



---

# Linking Visual and Textual Entity mentions with Background Knowledge

**Shahi Dost**

(Cycle-XXXIII)

shahi.dost@studenti.unipd.it

*School of Brain, Mind and Computer Science,  
Department of General Psychology,  
University of Padova, Italy*

*Data and Knowledge Management, research-unit  
Fondazione Bruno Kessler, Trento, Italy*

March 22, 2021

**Advisor:**

**Dr. Luciano Serafini**

Fondazione Bruno Kessler, Trento, Italy  
serafini@fbk.eu

**Prof. Alessandro Sperduti**

Department of Mathematics, University of Padova  
alessandro.sperduti@me.com

**Committee:**

**Prof. Natalia Díaz-Rodríguez**

Institut Polytechnique Paris, ENSTA (École Nationale Supérieure de Techniques Avancées), INRIA, France.

**Prof. Stefano Ferilli**

Department of Computer Science, University of Bari, Italy

**Dr. Abdullah Khan**

Department of Computer Sciece, University of Essex, UK





# Dedication

*To my parents*

# Abstract

*A picture is worth a thousand words, the adage reads. However, pictures cannot replace words in terms of their ability to efficiently convey clear (mostly) unambiguous and concise knowledge. Images and text, indeed reveal different and complementary information that, if combined will result in more information than the sum of that contained in a single media. The combination of visual and textual information can be obtained by linking the entities mentioned in the text with those shown in the pictures. To further integrate this with the agent's background knowledge, an additional step is necessary. That is, either finding the entities in the agent knowledge base that correspond to those mentioned in the text or shown in the picture or, extending the knowledge base with the newly discovered entities. We call this complex task Visual-Textual-Knowledge Entity Linking (VTKEL). In this thesis, after providing a precise definition of the VTKEL task, we present two datasets called VTKEL1k\* and VTKEL30k. These datasets consisting of images and corresponding captions, in which the image and textual mentions are both annotated with the corresponding entities typed according to the YAGO ontology. The datasets can be used for training and evaluating algorithms of the VTKEL task. Successively, we developed a baseline algorithm called VT-LINKER (Visual-Textual-Knowledge Entity Linker) for the solution of the VTKEL task. We evaluated the performances of VT-LINKER on both datasets. We also developed a supervised algorithm called VITKAN (Visual-Textual-Knowledge Alignment Network). We trained the VITKAN algorithm using features data of the VTKEL1k\* dataset. The experimental results of VITKAN on VTKEL1k\* and VTKEL30k datasets improved the accuracy with respect to the baseline.*

**Keywords:** AI, NLP, Computer Vision, Machine Learning, Knowledge Representation, Semantic Web, Entity recognition and linking.



# Acknowledgments

First, foremost, and the leading person, I want to thank is my Ph.D. advisor *Luciano Serafini*, who has shown me the light of cutting edge research in a true meaning. Constantly working with him, I had the opportunity to adopt hardworking, dedication, patience, quality work, and many skills expected from a researcher. I learn a lot from him in the fields of Artificial Intelligence, Semantic Web, Logics, Machine Learning, Big Data Analytics, and Knowledge Representations. Without his support, patience, knowledge, and guidance, it was impossible to complete this Ph.D. thesis.

The second most important person, I would like to thank is my co-advisor Prof. Alessandro Sperduti. I have improved my machine and deep learning skills under his supervision and teaching. He supported me during my Ph.D. duration directly or indirectly in the University administrative matters. I learn a lot from him during my Ph.D. courses and when we were discussing the challenging tasks of the Ph.D. project.

I would like to thank to my colleague and friend Prof. Marco Rospocher. He introduced me to the world of *Semantic-Web* in a very clarified and simple way. He taught me many tricks, best practices, and technicalities in the fields of Linked Data, Ontologies, Natural Language Processing, and Knowledge Representation. He advised me during my initial phase of Ph.D. and helped me in the development of the VTKEL dataset, and the VT-LinkER algorithm.

I would like to thank to Prof. Lamberto Ballan, who advised me on “*how we can combine the scientific community of Computer Vision and Natural Language Processing, with Knowledge Representation*”. I learn from him and improve my skills in the field of Computer Vision.

I would like to thank to Francesco Corcoglioniti, Ivan Donadello, and Alessandro Daniele, for their support, discussion, and guidance in fields of Computer Vision, Knowledge Representation, and Machine Learning.

Prof. Natalia Díaz-Rodríguez was my foreign-advisor during my study abroad at the Institut Polytechnique Paris, ENSTA (École Nationale Supérieure de Techniques Avancées), INRIA, France. I would like to thank for her support during my study abroad, for reviewing my Ph.D. thesis, and for her guidance in the fields of machine learning during the last phase of my Ph.D. thesis. I would like to thank Prof. Stefano Ferilli for reviewing my Ph.D. thesis with Prof. Natalia.

I acknowledge the funding provided by Fondazione Bruno Kessler (FBK), Trento that hosted me in the School of Brain, Mind, and Computer Science (BMCS), at the University of Padova and Data, Knowledge, and Management research unit at FBK, Trento for the whole duration of my Ph.D. studies. I would like to thank, Prof. Anna Spagnolli, Prof. Gianluca Campana, Prof. Giuseppe Sartori, and Ilaria Longo for their administrative support during my Ph.D. on the behalf of BMCS school.



I want to thanks to my colleagues and friends: Greta Adamo, Gaetano Calabrese, Loris Bozzato, Chiara Di Francescomarino, Mauro Dragoni, Chiara Ghidini, Radim Nedbal, Giulio Petrucci, Williams Rizzi, Sagar Malhotra, Abdullah Khan, Davide Rigoni, Tommaso Campari, Tahir Ahmad, Wasif Safeen, Hazrat Hussain, Hafeezullah Khan, Zaffar bhai, Kashif Ahmad, Sudipan Saha, Hussain Manikeya Bhai, Anwar Khan, Muhammad Haroon, Asif Khan, Tahir Ahmad (Ashna), Zaffar Bhai, Sana-Ullah, Mehtab Ali, Faheem Shahid, Athar Ali, Nicole Pendić, Setisemhal Getachew, Usmani Munazza, Dhouha Jemmeli, Shah Nawaz, for their advice and support in any form.

I would like to thanks my mentor Dr. Aftab Maroof and my friend Waqar Baig who motivated me to pursue Ph.D. and helped me during the starting phase.

I want to thanks to my friends (abroad): Hafeez Ur Rehman, Farooq Shah, Fazal Wahab, Asad Khan, Dawood Khan, Habib Dost, Shahid Iqbal, Atif Khattak, Faryal Saud, Habib-Ullah, Imtiaz Khan, and Muhammad Abdullah, who were even for away but supporting and motivating me from the long distance.

Last but not least, I warmly thanks to my family members: my dad (Alam Dost), my mom (Rahima Alam), my, sisters (Samena Yousuf, Shagufta Nisar, Ayesha Imtiaz, Sobia Jawad), my brothers (Faisal Dost & Waqas Dost), for their patience, prayers, and for supporting me during these years and, especially, for their love.



# Contents

<b>List of Figures</b>	<b>2</b>
<b>List of Tables</b>	<b>3</b>
<b>1 Introduction</b>	<b>5</b>
1.1 Contributions	8
1.2 Structure of the Thesis	9
1.3 Publication Note	10
1.4 Artifacts	10
<b>2 The Problem</b>	<b>13</b>
<b>3 State of the Art</b>	<b>18</b>
3.1 Vision and Language Integration	18
3.2 Multimodal Interpretation	20
3.3 Visual Grounding	21
<b>4 Background</b>	<b>25</b>
4.1 Flickr30k-entities dataset	26
4.2 YAGO	27
4.3 PIKES	28
4.4 Object Detection	28
4.4.1 YOLO	29
4.4.2 Mask-RCNN	29
4.4.3 Keras-RetinaNet	30
4.5 VGG16	30
<b>5 VTKEL Dataset</b>	<b>33</b>
5.1 Related work	34
5.2 A Data Model for multi-modal knowledge extraction	35
5.2.1 Resource Layer	35
5.2.2 Mention Layer	35
5.2.3 Entity Layer	36
5.3 VTKEL data model instantiation	37
5.4 Evaluations	39
5.5 Conclusion	40
<b>6 The VT-LinkEr Algorithm</b>	<b>42</b>
6.1 Related Work	42
6.2 Algorithm	43
6.2.1 Visual Mention Detection (VMD)	44

6.2.2	Visual Entity Typing (VET)	45
6.2.3	Textual Mention Detection (TMD)	45
6.2.4	Textual Entity Typing (TET)	46
6.2.5	Visual Textual Coreference (VTC)	46
6.3	Experimental Evaluations	47
6.3.1	Datasets	47
6.3.2	Evaluation	48
<b>7</b>	<b>The ViTKan Algorithm</b>	<b>52</b>
7.1	Introduction	52
7.2	Related Work	53
7.3	The ViTKAN Algorithm	54
7.3.1	Visual Module	55
7.3.2	Language Module	55
7.3.3	Neural-Network module and training of ViTKAN	56
7.4	Experimental Evaluations	57
7.4.1	Visual Textual Coreference	57
7.4.2	Textual Entity Grounding	58
7.5	Conclusions	59
<b>8</b>	<b>Conclusion</b>	<b>62</b>
	<b>Bibliography</b>	<b>65</b>

# List of Figures

2.1	The picture shows the output of the VTKEL task, which takes in input a picture with related text and an ontology. The output consists in a set of visual mentions (c.f. the bounding boxes in the image) and textual mentions (c.f. the highlighted words in the sentences), corresponding to the mentioned entities (in this case: a ball, a woman, the tennis court and a racket), and the extension (or alignment) of the ontology with entities of the correct (most specific) type.	15
4.1	Flickr30k-entities dataset single entry. Each image has five captions and the color-coding (for each visual co-reference chain) represents the visual entities in the image with their corresponding textual-mentions in the text.	26
4.2	A segment of <i>knowledge graph</i> from YAGO, which represent fact about <i>Max Planck</i> in the RDF form.	27
4.3	Passing caption “ <i>A young woman on a tennis court with a ball coming from behind her.</i> ” through the PIKES tool. In the mentioned graph, nodes represent the entity, event, or situation and arcs represent <i>relations</i> between them.	29
5.1	Data model overview (OWL ontology shown in UML notation).	36
5.2	A single document of the VTKEL dataset, which consists of an image, five captions (in the upper part), and co-reference chain with <i>ID = 255542</i> is represented in the RDF graph (in the lower part).	38
6.1	The (RDF) graph of the triples resulting from the VTKEL task. Each $vm_i$ corresponds to the bounding boxes mentioning some entity (visual mentions); each $ve_i$ represents an entity shown in some bounding box (visual entity); each $tm_i$ corresponds to the portion of text mentioning some entity (textual mentions); finally, each $te_i$ corresponds to some entity mentioned in the text. The other nodes of the knowledge-graphs are the concepts of the knowledge base, typing the entities.	44
6.2	An example of how data are stored in the VTKEL dataset. The RDF graph shows how the visual and textual mentions of the woman are mapped to a visual entity and 5 textual entities, all linked together by the <code>owl:sameAs</code> relation. The entities are linked to the most specific YAGO classes in this case person, woman, and player.	48
7.1	The ViTKAN algorithm architecture pipeline.	56
7.2	Some images of VTKEL dataset that consist of challenging objects.	59

# List of Tables

5.1	PIKES result over the caption " <i>A man is inside a truck looking out with his left arm in front of a door</i> ". . . . .	39
5.2	Examples of wrongly aligned YAGO Classes obtained by processing with PIKES, together with the manually corrected ones. The underline-word in textual mention column is processed wrongly. . . .	41
6.1	VT-LINKER and VITKAN evaluation results on VTKEL1k* and VTKEL30k using KRN object detector. . . . .	49
7.1	Results of state-of-the-art using Flickr30k-Entity, and VITKAN algorithm using VTKEL1k* and VTKEL30k datasets for TEG task. .	58



# Chapter 1

## Introduction

The universe of digital data is huge and a large amount of this data consists of multimedia data. A major portion of the multimedia data is available in the form of images and some text which described the contents of the image called *image caption*. On a daily basis, this digital data is growing exponentially and in the future after the implementation of 5G technology [1], these trends will be double. Thousands of news organizations are publishing their content on social networks, websites, repositories, hard drives, and other media. Approximately 95% of these news contents are in the form of some images and text, which describes the news (story) about war, pandemic, accident, technology, weather, sports, etc.,. This big amount of data (i.e. pictures with textual descriptions) conveys a huge quantity of information. Given a portion of news, which consists of a picture and some text, for human its take less time to understand, because most of the time human know the contents of images and text from their background knowledge. However, for the machines, it's a twofold challenging problem. In the first aspect, the machine should understand what is shown in the image and what is described in the textual part. While in the second aspect, the machine will interpret and understand the background knowledge of image and textual contents.

To process, this huge amount of multimedia data, we need automatic tools that analyze and extract the visual and textual portion of a document, with their background knowledge in an effective and efficient way. These tools, which based on artificial intelligence techniques will solve the tasks in the domain of vision and language more efficiently.

Understanding the contents of documents composed of images and text is an important task, which jointly described one particular topic. With the growing maturity and reliability of Natural Language Processing (NLP) and Computer Vision (CV) technologies have set the basis for deploying them in many products and real-world applications. However, the independent processing of the textual and visual part of a document is not sufficient to fully understand its content. We need a more integrated process. While focusing on the pictorial and textual parts of a document, it mostly consists of *entities*, which provide complementary information about them. For instance, in a news about a *car accident*, the text may mention the *brand* and *model* of the car involved in the accident as well as the name of the driver, while the picture may reveal the car brand and model as well, but also the car color and its status after the accident. Redundant information between text and images (c.f., the car brand, and model) enables matching the *visual* and *textual mentions* (i.e., portion of image or spans of text denoting some entities or facts) of the same entity (c.f., the car). Matching mentions, in turn, allows joining the complemen-



tary information (c.f., name of the driver, car color, and status after the accident) contributed independently by the two media. Furthermore, this information is usually interpreted by human agents also in the light of some background knowledge. This background knowledge, typically operationalized in terms of a *knowledge base* (=T-box + A-box), actually plays a double role: on the one side it is used as input for processing and understanding the content of the document; and, on the other side it is augmented with the additional knowledge resulting from the interpretation of the document, i.e., new facts contained in the document about entities either already present or to be added in the background knowledge base. We call this task *Visual-Textual-Knowledge Entity Linking* (VTKEL). More precisely, the VTKEL task aims at *detecting* and *linking* the maximum visual and textual portions of a document that refer to the same or individual entities of the document, a.k.a. *entity mentions*, with the corresponding entity (or a newly created one) in a knowledge base.

Given a document, composed of image and text, VTKEL is a complex task, which can be solved by using start-of-the-art tools in NLP, CV, and Knowledge Representation & Reasoning (KRR) communities. The VTKEL task can be analyzed and solved in different phases, where each phase described the interpretation of image, text, and ontological resource.

**First Phase:** This phase analyzed the contents of the image and text. The first task is to discovered (and analyzed) the objects (with their labels of classification) in the image. Moreover, it is possible to discover some attributes of the localized objects, such as, the shape, color, age, and gender (if the object is a person), etc. in this way, there is also the possibility to know what are the objects and where they exist in the image. Analysis of this portion is called *object detection* and *localization* of detected objects within bounding boxes (a rectangle around the object). In the second task, we have to analyze the textual part of the document for *entity recognition* and *localization* in the text. To solve this task, we have to process the given text with the help of a tool based on NLP techniques for *entity recognition* and *classification*.

**Second Phase:** The detected bounding box is called *visual mention* and the corresponding object, which is the instance of the class label, is called *visual entity*. The textual portion which consists of a noun (or more than one noun), is called *noun-phrase*, (e.g. “a man in a white shirt”, consists of two nouns, *man* and *shirt*) and their corresponding instances are called *textual-mentions*. After, visual and textual entity recognition and classification, the next task is to link them with the classes in the knowledge base called *entity linking* [2, 3].

**Third Phase:** In this phase, we have to make the alignment between visual and textual entity mentions. For this task, we have to exploit the class/sub-class hierarchy between the classes in the knowledge base. Let  $TE$  and  $VE$  be the set of textual and visual entities ( $te$  &  $ve$ ) that are mentioned in a visual-textual document, and that is present in the knowledge base with a given type. The coreference task has the objective of finding the coreference relation ( $CR$ ) [4]:  $CR \subseteq VE \times TE$  such that the following consistent properties hold:

1. For every  $ve \in VE$  there is at least one  $\langle ve, te \rangle \in CR$ ;
2. For every  $ve \in VE$  there is at most one  $\langle ve, te \rangle \in CR$ ;

3. If  $\langle ce, ve \rangle$  ( $ce$  is the coreference entity) and  $ve$  and  $te$  are of type  $C_v$  and  $C_t$  respectively then either  $C_v \sqsubseteq C_t$  or  $C_t \sqsubseteq C_e$  holds in the knowledge base.

The VTKEL task and its solution enable a set of important applications in the area of multimedia *indexing*, *retrieval*, and *vision-language* tasks. Here, we enlist some of the examples:

#### **Information Extraction from Multimedia Systems:**

A structured description of an image and text contents allows the retrieval of a given query (i.e. noun-phrase grounding or phrase-localization [5]) from visual and textual mentions. With the help of background knowledge, the system can process that query more intelligently. For example, in a *forensic system*, a textual structured query can retrieve all the people images (with their bounding boxes/visual mentions) having "*weapon in their hand*" from the documents automatically.

#### **Visual Question Answering (VQA):**

To help a blind and visually impaired user, a VQA system could provide information about an image on the web or social media. For example, "*Are there any human?*", "*how many players are in the image excluding the referee?*", "*Is it raining?*" etc.. can be solved by the VQA system.

#### **Visual Dialogue System:**

In the Visual Dialog (VD) system, an AI agent is responsible to answers a multi-round of questions about an image (visual-contents) from humans in natural (conversational) language. The VD system is addressing both the VQA and how to infer the co-reference between questions and the dialog history. For example, "*How many lamps are in the table?*", "*Are they on or off?*", "*What is the color of the left lamp?*" etc.,.

#### **Image Captioning:**

Image captioning is the process of passing an image to an AI agent to automatically generate the natural language description of that image. It connects the CV and NLP tools and approaches to solve the task of image captioning. In this task, the AI agent extracts the visual features, detect visual concepts (objects, regions, attribute, events, etc.), and in the final stage generates natural language sentences for the description of a given image.

#### **Robotics:**

A robot moving in an environment interact with different objects and scene, enforce him to perform a set of actions. By injecting the contextual plus structural (Ontological) knowledge, the robot can perform these actions in a more intelligent way. For example, "*a can of coke*" (i.e. cold drink tin) on the table can be grasped avoiding "*the cup of tea*" (warm drink).

#### **Complex Image and Textual Querying:**

The structured image and textual information with background knowledge can be converted to Resource Description Framework (RDF) graph [6]. This will enable us to perform structured queries on both images and text by utilizing Semantic Web languages, such as SPARQL [7]. The semantic base query can process a complex scenario, which consists of more than one query in an appropriate and efficient way. For example, retrieve those images showing "*a person*

*with German-shepherd during the hiking*". In this query, the challenge is not just to predict a person and a dog, but also the "hilly area" showing trees and mountains.

The aim of this Ph.D. thesis is to develop novel datasets and algorithms to solve the problem of the VTKEL. The solution of the VTKEL problem closed the loop between the scientific community of CV, NLP, and KRR. However, the VTKEL problem becomes challenging in some aspects:

**Datasets:** Recently the scientific community of NLP and CV devoted a reasonable effort in investigating the interaction and integration of text and image processing. However, there is not a single dataset that combines, NLP, and CV with KRR. The development of those datasets, which stored the annotations of visual, textual, and knowledge-base contents is one of the big challenges.

**Hybrid Domain:** The visual and textual entity-mentions can be represented in semantic, knowledge, and numeric features. The semantic features of visual and textual entities are (i) the labels describing the types of objects, for example, "person", "car" and "dog", (ii) the image region for objects (localization), and portion in the text for the noun-phrase. The knowledge-based features can be (i) class-hierarchy (i.e. subclass or superclass), (ii) entity gloss (i.e. description of entity mention), or (iii) synonyms of visual and textual entity mentions. The numeric features can be, (i) the bounding box coordinates of objects, (ii) the visual-features extracted with the help of CV techniques, (iii) the location (in characters) of the textual entity in the sentence, or (iv) embedding [8] of textual entity mentions.

**Multimodal Domain:** Processing a document consists of images and text [5, 9] in different domains (e.g. news, social media, image-captioning, etc.) is a challenging task for machines. Moreover, the visual data and their features are different from the textual part. If we add the knowledge base (Ontological) to the visual and textual parts, the problem becomes more challenging, but in return provides the possibility to utilize huge structured background knowledge [2, 10, 11].

**Ambiguity:** The entities shown in images and described in the text and their relationships with each other make events (scenario) [12]. Sometimes it is difficult for a machine to fully understand that scenario. For example, in a picture, if a person is with a dog and in the background there is also a Llama (Lama). Most of the time the CV tool predicts Lama with horse and some time with sheep. Similarly, in the text "A woman holds a man's arm at a formal event", the machine sometimes predicts "arm" with weapon class instead of a human limb.

## 1.1 Contributions

The main contributions of this thesis are as follows:

**C1 - The VTKEL task:** We introduce a complex and novel task called VTKEL, which aims at linking the maximum visual and textual portions of a document that refer to the same individual entity a.k.a entity-mentions, with the corresponding entity (or a newly created one) in a knowledge base.

**C2 - The VTKEL datasets:** The second contribution of this thesis is the assembling of ground-truth datasets for the VTKEL task. We developed two datasets called VTKEL1k\*<sup>1</sup> and VTKEL30k<sup>2</sup>. These datasets can be used for the training and evaluations of algorithms to solve the VTKEL problem.

**C3 - The VT-LinKEr (baseline) algorithm:** The third contribution of this thesis is the development of an unsupervised algorithm called VT-LINKER (*Visual-Textual-Knowledge-Entity Linker*). The VT-LINKER algorithm solving the VTKEL task by combining state-of-the-art NLP, CV, and Ontological reasoning tools and techniques.

**C4 - The ViTKan algorithm:** The fourth contribution of this thesis is the development of a supervised algorithm called VITKAN. This algorithm solves the task of VTKEL with great accuracy.

## 1.2 Structure of the Thesis

The roadmap of the thesis is structured as follows:

**Chapter 2:** This chapter provides the definition of research problem in details.

**Chapter 3:** This chapter provides state-of-the-arts literatures and approaches used in this Ph.D. thesis.

**Chapter 4:** This chapter provides background of the state-of-the-art tools in the scientific communities of CV, NLP and KRR: including Flickr30k entity dataset<sup>[5]</sup>, Knowledge-base (YAGO)<sup>[13]</sup>, PIKES<sup>[14]</sup>, Keras-RetinaNet<sup>[15]</sup>, and VGG16<sup>[16]</sup>.

**Chapter 5:** This chapter provides the VTKEL datasets, their development, representation, and evaluations. This chapter represents *Contribution C1* and *C2*. We have published the results of this chapter in *the proceedings of the 35th annual ACM Symposium on Applied Computing(SAC2020)* (Semantic Web, and applications track)<sup>[17]</sup>.

**Chapter 6:** This chapter represents the introduction of VT-LINKER algorithm, their experiments, and evaluations in details. The *Contributions C3* are represented in this chapter. We have published the scientific results of this chapter in *the 24th European Conference on Artificial Intelligence (ECAI2020)*, *the 14th IEEE International Conference on Semantic Computing (IEEE-ICSC2020)*, and *the 25th International Conference on Natural Language & Information Systems (NLDB2020)*.

**Chapter 7:** In this chapter, we described VITKAN algorithm in details. This chapter represents *Contribution C4*. We have submitted the scientific results of this chapter with *Contribution C1, C2, C3* at the scientific Journal of *Data and Knowledge Engineering* (March 2021).

**Chapter 8:** This chapter provides the conclusion of Ph.D. thesis. This chapter also provides some possible futures research directions, strong points, and limitations of this thesis.

---

<sup>1</sup>This gold-standard dataset consists of 1000 documents.

<sup>2</sup>This big dataset consists of more than 30k documents.

### 1.3 Publication Note

The publications which comprise this Ph.D. thesis are listed below:

- **Shahi Dost**, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. "VTKEL: a resource for visual-textual-knowledge entity linking." *In the Proceedings of the 35th Annual ACM Symposium on Applied Computing, SAC 2020, Brno, Czech Republic, March 30 – April 3, 2020*, pages 2021-2028. ACM, 2020. [17]
- **Shahi Dost**, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. "VT-LINKER: Visual-Textual-Knowledge-Entity Linker." *In 24th European Conference on Artificial Intelligence (ECAI), Santiago de Compostela, Spain, August 29 – September 5, 2020*, pages 234-235. 2020. [18]
- **Shahi Dost**, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. "On Visual-Textual-Knowledge Entity Linking." *In IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, California, USA, February 3-5, 2020*, pages 190-193. IEEE, 2020. [19]
- **Shahi Dost**, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. "Jointly linking visual and textual entity mentions with background knowledge." *In 25th International Conference on Natural Language & Information Systems (NLDB), DFKI Saarbrücken, Germany, June 24 – 26 September 5, 2020*, pages 234-235. Springer Nature, 2020. [20]
- **Shahi Dost**, Luciano Serafini, and Alessandro Sperduti. "Semantic Interpretation of Image and Text". *In the Proceedings of the Doctoral Consortium (DC) co-located with the 17th Conference of the Italian Association for Artificial Intelligence (AI\*IA 2018), Trento, Italy, November 20-23, 2018., volume 2249 of CEUR Workshop Proceedings, pages 48-53* (CEUR-WS.org, 2018). [21]
- **Shahi Dost**, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. "ViTKAN: A supervised algorithm for visual-textual knowledge entity alignment and linking." *Journal of the Data & Knowledge Engineering*". [22].

### 1.4 Artifacts

The main artifacts supporting this thesis are listed below:

- A1** The VTKEL datasets for 300, 1000 and 30K+ documents are available online at: VTKEL<sup>1</sup>, VTKEL<sup>2</sup>, VTKEL<sup>3</sup> respectively.
- A2** The source code for the paper in [17] is available at: <https://github.com/shahidost/VTKEL>
- A3** The source code for the papers in [19], [20], [21] is available at: <https://github.com/shahidost/Baseline4VTKEL>

<sup>1</sup><https://bit.ly/2Bxu5mU>

<sup>2</sup><https://bit.ly/3etlbWq>

<sup>3</sup><https://bit.ly/2Bxu5mU>

**A4** A source code for the paper in [22] is available at: <https://github.com/shahidost/VTKEL-KENN>



## Chapter 2

# The Problem

In this thesis, we are interested to understand the contents of a document composed of an image and text. The visual part of the document consists of objects and the textual part described these objects and their relationship (visual contents) in natural language. We need to develop an artificial agent that involved cross-modal learning from image and text data and predicts the objects (i.e. visual entity mentions) shown in the image and parallelly recognized the textual entities mentions described in the text. After recognizing the visual and textual entity mentions, the agent will link them to its background knowledge by using the knowledge-bases (e.g. DBpedia [23], YAGO [13] etc.,).

*Visual-Textual-Knowledge Entity Linking* (VTKEL) is the task of taking in input a document composed of an image and text<sup>1</sup>. More precisely, a document  $d$  is a pair  $\langle d_t, d_i \rangle$ , where  $d_t$  is a text in natural language represented as a string of characters and  $d_i$  is an image, represented as a 3-channel ( $w \times h$ )-matrix. Notice that, for the sake of simplicity, we ignore all the structural information about the document, e.g. the relative position of the image w.r.t. the text, the explicit references to the figures, etc. If  $e$  is an *entity* of the domain of discourse in a document  $d$ , for example a specific *car* or a *person*, a *textual mention* of  $e$  in  $d$  is a portion of the text  $d_t$  that refers to the *entity*  $e$ . Such a mention can be identified by an interval  $\langle l, r \rangle$  with  $0 \leq l < r \leq \text{len}(d_t)$ , corresponding to the characters (in  $d_t$ ) of the mention. Analogously, a *visual mention* of an entity  $e$  is a region of the picture  $d_i$  that shows (a characterising part of) the entity  $e$ . E.g., the region of a picture that shows the (face of a) person is a *visual mention* of that person. If we restrict to rectangular regions (a.k.a. bounding boxes) a visual mention can be represented by a bounding box encoded by four integers  $\langle x, y, x + w, y + h \rangle$  with  $0 \leq x, x + w \leq \text{width}(d_i)$  and  $0 \leq y, y + h \leq \text{height}(d_i)$ , where  $\langle x, y \rangle$  represents the position of the pixel in the top left corner of the bounding box, and  $w, h$  represent the width and height of the bounding box (in pixels).

A logic-based *Knowledge-base* [2] is a logical theory that states attributes and relations about a set of entities, called the domain, using a logical language. In description logics, a *knowledge-base* is composed of a T-box and an A-box. The T-box contains a set of axioms of the form  $C \sqsubseteq D$  and  $R \sqsubseteq S$ , for some concept expressions  $C$  and  $D$  and relations  $R$  and  $S$  stating that  $C$  is a sub-class of  $D$  ( $R$  is a sub-relation of  $S$ ). The A-box contains assertions of the form  $C(e)$  (the entity

---

<sup>1</sup>For the sake of simplicity, we consider only documents that contain one single picture. The extension to multiple pictures, though intuitive, presents additional challenges that are out of the scope of this thesis

<sup>2</sup>[https://en.wikipedia.org/wiki/Knowledge\\_base](https://en.wikipedia.org/wiki/Knowledge_base)



$e$  is of type  $C$ ) and  $R(e, f)$  (the pair of entities  $\langle e, f \rangle$  are in relation  $R$ ) where  $e$  and  $f$  are entities of the *Knowledge-base* and  $C$  and  $R$  are concept and role expressions respectively. The *entities* of a knowledge-base are constant symbols that explicitly occur in some axiom of the T-box or assertion of the A-box. For instance, the T-box may contain the knowledge that every *car* has a *manufacturer* and that a *manufacturer* is a *company*. This knowledge can be formalized by the axioms  $\text{Car} \sqsubseteq \exists \text{hasManufacturer. Manufacturer}$  and  $\text{Manufacturer} \sqsubseteq \text{Company}$ , where *Car*, *Manufacturer*, and *Company* are concept names and *hasManufacturer* is a relation (or role). The A-box may contain the knowledge that a specific *car* (an entity), say  $\text{car}_{22}$ , is a *BMW* and that *BMW* is a *Manufacturer*. This is formalized by the assertional axioms  $\text{Car}(\text{car}_{22})$ ,  $\text{hasManufacturer}(\text{car}_{22}, \text{BMW})$ , and  $\text{Manufacturer}(\text{BMW})$ .

**Problem 1** (VTKEL). Given a document  $d$  composed of a text  $d_t$  and an image  $d_i$  and a Knowledge-base  $K$ , *VTKEL* is the problem of detecting all the entities mentioned in  $d_t$  and shown in  $d_i$ , and linking them to the corresponding named entities in  $K$ , if they are present, or linking them to new entities, extending the A-box of  $K$  with its type assertion(s), i.e. adding  $C(e^{new})$  for each new entity  $e^{new}$  of type  $C$  mentioned in  $d$ .

**Example 1.** To understand the *VTKEL* problem in details, consider the document shown in Figure 2.1, which is composed of one picture and two short sentences (image captions) in natural language. In Figure 2.1, one can find four visual mentions, shown in colored rectangles in the picture, and five<sup>3</sup> textual mentions, colored in the text. One could find many visual mentions in the picture (e.g., grass, white-line, t-shirt etc.) but suppose we are only interested in the mentions of certain types. Let us consider a *Knowledge-base* (e.g., YAGO [13]) that contains knowledge about the named entities  $e_{player}$ ,  $e_{woman}$ ,  $e_{court}$ ,  $e_{ball}$  and  $e_{racket}$  for "Player", "Woman", "Court", "Ball" and "Racket" respectively, with the corresponding types  $person(e_{player})$ ,  $person(e_{woman})$ ,  $location(e_{court})$ ,  $artifact(e_{ball})$ , and  $artifact(e_{racket})$ . Let us suppose that the *Knowledge-base* contains also the concepts *woman*, and *ball*, and we want to describe these mentioned concepts. The visual (blue box) and textual (blue text) mentions of a *woman* refer to the same entity, and they should be linked together and typed (using *rdf:Type*) according to the YAGO class *Woman110787470*<sup>4</sup>. The visual mention of *player* should be typed according to the YAGO class *Player110439851*<sup>5</sup>. The textual mentions *woman* and *player* (blue text) are refer to the same visual mention *person* (blue box) because *woman* and *player* are the *rdfs:subClassOf person* in the *Knowledge-base* YAGO. For the instance, if there are two *women* showing in the picture, we will assign *woman*<sup>1</sup> and *woman*<sup>2</sup> to differentiate them and type the visual mentions with YAGO class *Woman110787470*. The textual mention *court* (yellow text) and visual mention *location* (yellow box) are related entity and they are typed with YAGO class *PlayingField108570758*<sup>6</sup>. The remaining visual and textual mentions should be linked to entities with the corresponding YAGO type i.e., we should add the assertion *ball* and *racket* to *Ball102778669*<sup>7</sup> and *Racket104039381*<sup>8</sup> respectively.

<sup>3</sup>From the syntactic analysis of the noun phrase, for the instance we are only considering the head of the noun-phrase. For example, in "the group of people", the *head* of the noun-phrase is "group" while "people" is a *modifier*.

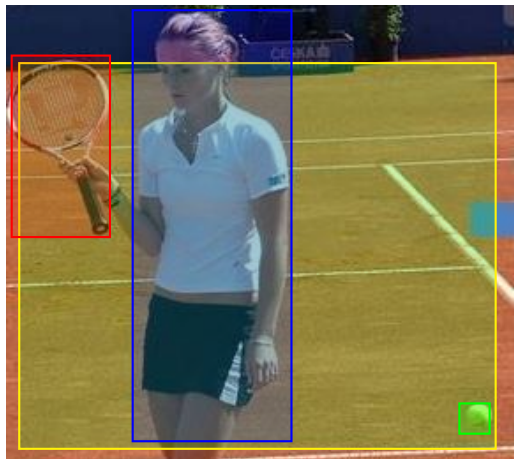
<sup>4</sup>The URI <http://dbpedia.org/class/yago/Woman110787470> stored all the background knowledge of class *Woman110787470* from YAGO ontology in structured (RDF) graph.

<sup>5</sup><http://dbpedia.org/class/yago/Player110439851>

<sup>6</sup><http://dbpedia.org/class/yago/PlayingField108570758>

<sup>7</sup><http://dbpedia.org/class/yago/Ball102778669>

<sup>8</sup><http://dbpedia.org/class/yago/Racket104039381>



1. A young woman on a tennis court with a ball coming from behind her.
2. A female tennis player casually swinging her tennis racket .

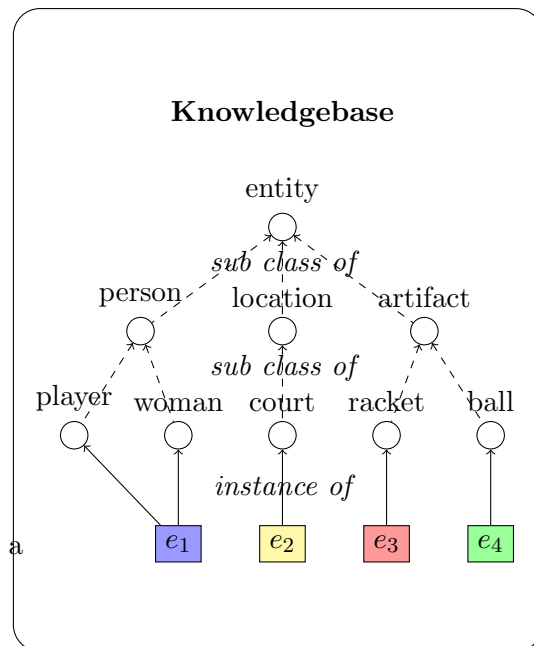


Figure 2.1: The picture shows the output of the VTKEL task, which takes in input a picture with related text and an ontology. The output consists in a set of visual mentions (c.f. the bounding boxes in the image) and textual mentions (c.f. the highlighted words in the sentences), corresponding to the mentioned entities (in this case: a ball, a woman, the tennis court and a racket), and the extension (or alignment) of the ontology with entities of the correct (most specific) type.

The *VTKEL* is a complex task that requires the solution of a set of well-studied elementary tasks in CV, NLP, and KRR. In particular, the following are the key subtasks of *VTKEL*:

1. The named entity recognition and classification (i.e. typing) in texts [24];
2. Visual Objects detection and classification in images [25];
3. Textual co-reference resolution<sup>9</sup> [26];
4. Linking textual entity to the *Knowledge-base*(Ontology) [2];
5. Linking visual entity to the *Knowledge-base*(Ontology) [27, 28];
6. Visual and textual co-reference resolution [29, 30, 31].

In the recent literature, a number of approaches focusing on one specific task, or a subset of the VTKEL tasks can be found. However, it is well-established in many areas of NLP and CV (see chapter 3) that there is a clear advantage in solving complex tasks in a collective/joint manner, rather than combining the results of task-specific tools used as a black-box. It is indeed clear that, the relation among the appearance of an entity in an image, its associated linguistic properties within

<sup>9</sup>The co-reference here means that when two or more expressions in a text refer to the same person or thing (i.e. they have the same referent), e.g. *David said he would become pilot*; the proper noun *David* and the pronoun *he* refers to the same person, namely *David*.

the text, and the semantic/axiomatic knowledge contained in the *Knowledge-base*, can jointly contribute to the solution of the complex task altogether.

We are particularly interested in this Ph.D. thesis, to develop a dataset (chapter 5) which is annotated with all the ground-truth data needed for the *VTKEL* problem. A baseline algorithm (chapter 6) to solve the problem of *VTKEL* by using state-of-the-art tools and techniques in the community of CV, NLP, and KRR. In the end, we want to solve the *VTKEL* problem by using supervised techniques (i.e. artificial-neural-network model) (chapter 7).



## Chapter 3

# State of the Art

The previous chapter defines the problem of *Visual-Textual-Knowledge Entity Linking* (VTKEL) in detail. In this chapter, we described the state-of-the-art methods and techniques used in this thesis and to accomplished the challenging tasks. We also described some of the approaches that are related to VTKEL problem, dataset, baseline algorithm (VT-LINKER), and the VITKAN model, separately in chapter [4](#), [5](#), [6](#), and [7](#) respectively.

### 3.1 Vision and Language Integration

There are several approaches and datasets that combined text and images (multi-modal), however, none of them have the three components that are vision, language, and knowledge necessary for the VTKEL task. A review of language and vision datasets are presented by *Francis et al.* [\[32\]](#). They categorized these datasets with respect to the tasks on image to language and vice-versa. *VisualGenome* [\[33\]](#) is an extremely large dataset that contains pictures in which objects are annotated with their types, attributes, and relationships. Annotations are mapped to *WordNet*<sup>[1](#)</sup> synsets. Objects can also be annotated with a short sentence that describes some qualitative property of the object. E.g., "*The girl is feeding the elephant*" or "*a handle of bananas*". However, there is no alignment between the objects mentioned in these phrases and the objects shown in the picture. E.g., there is no bounding box for the object "*bananas*" or "*elephant*". The *Visual Relationship Dataset* (VRD) [\[12\]](#) is a dataset of images annotated with bounding boxes around key objects. Furthermore, VRD contains annotations about relationships between objects in the form of triplets  $\langle \text{object.type}, \text{relation}, \text{subject.type} \rangle$  describing the scene. Examples of annotations are  $\langle \text{man}, \text{riding}, \text{bicycle} \rangle$  and  $\langle \text{car}, \text{on}, \text{road} \rangle$ . However, these annotations are not aligned to any knowledge base.

The *Microsoft COCO dataset* [\[34\]](#) contains pictures associated with five captions. They are annotated with object regions of any shape (not simple bounding boxes) and each region is assigned with an object type. This dataset does not contain any information about the relation between object regions, and the relation between regions and mentions in the captions. *Conceptual Captions* [\[35\]](#), [\[36\]](#) is a recently introduced dataset that has been developed for automatic image caption generation. It contains one order of magnitude more items than Microsoft COCO. It is a realistic dataset as images with captions have been automatically extracted and filtered from the web. However, there is no visual/textual mention annotation and

---

<sup>1</sup><https://wordnet.princeton.edu/>

visual textual entity linking. *VizWiz* [37] is a dataset generated by mobile users by talking pictures from their mobile, with image descriptions (captions), and recorded spoken questions about the picture. This dataset is very well known nowadays for developing models to assist people who are blind to overcome their daily (real) visual challenges. However, there is no background knowledge (Ontological) that described either pictures, captions, or voices contents of the dataset. Also, there is no visual, textual, and audio entity mentions annotations between each other.

Grew et al. [38] recently proposed a dataset called *GQA* for real-word visual reasoning and compositional questions answering to leverage key limitation in the Visual-Question-Answering (VQA) datasets. It consists of 113K images and 22M questions of assorted types with answers. In their dataset, images, questions, and corresponding answers are all represented by semantic matching. Each image is annotated with a dense *scene-graph*, which represent objects, attributes, and relations it contains. Each question is associated with a functional-program, which lists the series of reasoning steps needed to be performed to reach into the answer. Each answer is enlarged with both visual and textual justification, denoting (pointing) to the corresponding region in the image. In this dataset, the image regions are annotated with the portion of the text, however, they missed the annotation of textual entity mentions (of  $d_t$ ) with the corresponding visual objects (of  $d_i$ ). Ferrari et al. [39] propose a *Localized Narrative*, a new form of multimodal image annotations technique connecting vision and language by using voice and mouse-trace. During the annotations process, they ask annotators to describe an image with voice, while simultaneously hovering their mouse over the image region they are describing. The voice and the mouse pointer are aligned, which in return localized every word of text with image region. They annotated 849k images with Localized Narratives exploiting from COCO, Flickr30k, ADE20K, and 671k images of Google-Open-Images datasets. The drawback of this dataset is that no background knowledge of vision or language contents is associated, which is necessary for the VTKEL task. *Google Open Images* [40] dataset consists of 9.2 Million images with unified annotations for *image classification*, *object detection* and, *visual relationship detection*. This dataset offers large scale across several dimensions: (i) 30.1M image-level labels for 19.8k concepts, (ii) 15.4M bounding boxes for 500 object classes, and (iii) 375k visual relationship annotations involving 57 classes. For object detection, in particular, the authors provide  $15\times$  more bounding boxes than the next largest datasets (15.4M boxes on 1.9M images). The images often show complex scenes with several objects (8 annotated objects per image on average). They have annotated visual relationships between them, which support visual relationship detection, an emerging task that requires structured reasoning. This dataset is state-of-the-art and recently widely used for competitions in the area of object detection, classification, and visual-relation detection. However, there are no captions that described the contents of images with textual portions.

From the analysis and literature, it becomes clear that there is not any dataset, which combines images, texts, and the background knowledge of them. This justifies the development of VTKEL1k\* and VTKEL30k datasets, which not only combined images and corresponding text but also their background knowledge in the form of Ontological (structured) data.

## 3.2 Multimodal Interpretation

There is a huge research history of investigating the intersection and integration of vision and language called multimodal interpretation. The NLP and CV scientific communities are trying to solve various tasks such as textual grounding [5, 41, 42], visual question answering [43], and visual reasoning [44, 45] and various models have developed to solve them.

For an exhaustive survey of the approaches in the area of entity information extraction and linking, we refer the reader to [46]. In particular: [31] exploits natural language descriptions of a picture in order to understand the content of the scene itself. The proposed approach solves the image-to-text coreference problem. It successively exploits the visual information and visual-textual coreference previously found to solve coreference in text. In their approach, they did not mention the use of semantic or ontological knowledge of natural language descriptions and associated pictures contents. The work described in [47, 48] tackles the problem of ranking the concepts from the knowledge base that best represents the core message expressed in an image. This work involves the three elements: Image, Text, and Knowledge, but it does not provide information about the entities mentioned in the text and shown in the image. The approach in [49] adapts Markov Random Fields to represent the dependencies between what is shown in the frames of videos about the wild-life animal and the subtitles. The main objective is to detect the animal shown in a frame, and the mentions of animal in the subtitle. The set of entities are the animal names available in WordNet [50]. Object detection is not performed: the approach assumes that only one animal is shown in a frame, and the vision part consists of image classification. Furthermore, no background knowledge about animals is used. [27] proposes a basic framework for visual entity linking to DBpedia [23] and Freebase [51]. The approach involves also textual processing since the link of bounding boxes to DBpedia and Freebase entities is found passing through an automatically generated textual description of the image. The approach uses the Flickr8k dataset, which is a subset of the Flickr30k-Entities dataset. A combination of textual coreference resolution and linking of image and textual mentions is described in [52] with the objective of solving the problem of assigning names to people appearing in TV-show.

The approach in [53] presented the real logic and their implementation with *Logic Tensor Network* (LTN). The LTN is capable of learning from numerical data and logical reasoning via integrating with the help of first-order logic syntax for Semantic Image Interpretation. They used LTN for the task of classifying visual objects and their parts (e.g. “left hand of a person”) in images, using state-of-the-art object detectors by exploiting part-of ontology. They did not address particularly the task of recognizing visual (in  $d_i$ ) and textual (in  $d_t$ ) entity mentions, established alignment, and later linking them with the class instances of YAGO knowledge-base. The task of *Referring expression comprehension* (REF) is identifying a particular object in a scene by a natural language expression. The approaches for solving the problem of REF need to jointly process both the textual resource with visual domains [54, 55, 56]. In their approaches, they rely only on visual and textual data, and the semantic information coming from their contents. The work in [57] proposed a supervised algorithm called *Cops-Ref* (COmPoSitional Referring expression comprehension), which utilized their developed dataset for solving referring expression task. The Cosp-Ref takes in input a natural language referring expression, and a set of similar images, to make predictions on the REF task. They are not using

the background knowledge of input images and corresponding natural language referring expressions. The approach in [58] proposed an integrated framework that connects classification algorithms for the recognition of simple visual objects while using ontologies to recognize the complex objects by means of reasoning. They are using ontological knowledge in the visual part, but there is no textual descriptions or annotations which connect the image portions (objects) with textual data. In [59], the authors presented an unsupervised clustering method for automatic video contents annotation with Ontologies. In their approach, they present pictorially enriched ontologies and discuss a solution for their implementation for the soccer videos. The annotations are performed associating occurrences of events, or entities, to higher-level concepts by checking their proximity to visual concepts that are hierarchically linked to higher-level semantics. They based on visual data only, and there is no textual data or background knowledge linked to the description of soccer videos. The overview and analysis presented by [60] used the background knowledge and ontologies to provide a rich image understanding and image annotation in order to efficiently solve the tasks of retrieval. In the systematic review, they discussed a number of techniques and approaches which solved image annotation and interpretation to narrow the “semantic gap”. They also highlight the importance of reasoning and contextual knowledge in the image understanding process, emphasizes the limitations of current approaches, and provide solutions that can overcome these limitations. They have very limited textual data and annotations between textual and image entities. In [61] design a systematic approach using ontologies for *visual activity recognition* tasks from the video. In their approach, they draw on general ontology design principles and adapt them to the specific domain of ontology of human activities for bank and airport tarmac surveillance domains. Gomez et al. [62] developed a framework by using Ontology-based context representation and reasoning on *object tracking* and *scene interpretation* in videos. They have constructed a symbolic model of the scene by integrating *tracking data* and *contextual ontological information*. The scene model represented a formal ontology and supports the execution of reasoning procedures in order to: (i) obtain a high-level interpretation of the scenario; and (ii) provide feedback to the low-level tracking procedure to improve their performance. [63] uses a low-level action recognition on which an ontological approach works at the activity recognition level. In both object tracking and activity, recognition approaches, they have missed textual descriptions, which necessary for the VTKEL task, by connecting visual entity mentions in  $d_i$  with textual entity mentions in  $d_t$ .

The main limitations of the above approaches are either (i) using visual contents with ontological knowledge but missing textual descriptions, or (ii) using both visual and textual contents but missing the background knowledge. In the problem of VTKEL, we are using the three modalities i.e. *vision*, *text*, and *background knowledge* in one pipeline. This justified the creation of VTKEL datasets, and algorithms for solving VTKEL.

### 3.3 Visual Grounding

Given an image and textual description of the image, the problem of *phrase grounding* tries to localize visual objects in the image with the corresponding phrases described in the captions. The main challenge in the phrase grounding problem is the correlation between visual and textual modalities. Karpathy et al. [30] align noun-phrases and image regions, using (i) *convolutional neural network* (CNN) over



images, (ii) bidirectional *recurrent neural network* (RNN) over sentences, and (iii) a structured objective that aligns the two modalities. One of the popular baselines for image-text embedding is *Canonical Correlation Analysis* (CCA), which finds linear projections that maximize the correlation between projected vectors from the two image-regions and text domains described in [64]. Wang et al. [65] employ structured matching of phrases and regions which develop the semantic relations between phrases to agree with the visual relations between image-regions. They formulate structured matching as a discrete optimization problem into a linear program and use neural networks for embedding visual regions and phrases into vectors.

Plummer et al. [5] augment the CCA model to leverage extensive linguistic cues in the phrases. Rohrbach et al. [66] propose grounding by reconstruction, an approach using an attention mechanism for phrase grounding by ranking proposal in an unsupervised scenario. During training their approach encodes the phrase using a recurrent network language model and then learns to attend for the relevant image region in order to reconstruct the input phrase. Hu et al. [67] propose a *Spatial Context Recurrent ConvNet* model which based on a 2-layers LSTM to rank visual proposals using embedded query and visual features. Dogan et al. [68] proposed a sequential and contextual process, which encode region proposals and all phrase into two stacks of LSTM cells, along with so-far grounded phrase-region pairs. These LSTM stacks collectively capture context for grounding of the next phrase. The resulting architecture supports many-to-many matching by allowing an image region to be matched to multiple phrases and vice versa. ViLBERT (Vision-and-Language BERT) [69] learn representation jointly from both visual and textual domains using two-stream co-attentional transformer layers independently. In contrast to ViLBERT, VisualBERT [70] consists of a stack of transformer layers, which indirectly align elements of an input text and regions in an associated input image with self-attention. The VisualBERT further demonstrates elements of language to image in syntactic relationships, for example, associations between verbs and image regions corresponding to their arguments.

Yang et al. [71] propose a linguistic structured guided propagation network for one-stage phrase grounding. In their model, they explore the linguistic structure of the sentence and perform relational propagation among noun-phrases under the guidance of the linguistics relation between them. Specifically, they first constructed a linguistic graph parsed from the sentence and then capture visual and textual (multimodal) feature maps for all noun-phrases nodes independently. Jing. et al. [72] formulate the problem of phrase grounding as a graph matching problem to find the nodes of visual and textual entities and to represent them in structured layouts of the image and sentence respectively. In their approach, they build a cross-modal graph convolutional network to learn cohesive node representations, which distinguish both node information and structured information to reduce the inconsistency of visual and textual graphs. Yu et al. [73] propose a *Cross-Model Omni Interaction network* (COI-Net) composed of (i) a neighboring interaction module, (ii) a global interaction module, (iii) a cross-modal interaction module, and (iv) a multilevel alignment module. They formulate the complex spatial and semantic relationship between image regions and phrases using these multi-level multi-modal interactions. To further enhance the interaction between two modalities, they use a co-attention module with the cross-modal context for all image regions and phrases. [74] presents a neural-symbolic approach on which instead of background knowledge to enhance the explanation, a Knowledge-Base(KB) populated from the DNN's training dataset is created. A way to measure the alignment between textual and object entities is

using semantic fidelity, a metric for which it is possible to optimize for [75]. An approach that uses KBs to align compositional image classifiers with *How does state of the art with transformers performs with image captioning*, as described by [76].

These existing methods lack the ability to model the background knowledge of visual and textual modalities coming from the knowledge bases (Ontologies) by linking the visual and textual entities mentions. From the above literature, it becomes clear that there is not a single comprehensive approach, which used visual and textual modalities with the background knowledge for the task of phrase (noun, entity) grounding.



## Chapter 4

# Background

To solve the problem of VTKEL, we start by developing *VTKEL* dataset. We used the Flickr30k-Entities dataset [5] is a starting point. The Flickr30k-Entities dataset provides the annotation of coreference chains, i.e., linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in automatic image description and grounded language understanding. They enable us to define a new benchmark for the localization of textual entity mentions in an image. To link textual entities to an ontological resource, namely YAGO [13], we use PIKES [10]. PIKES is a state-of-the-art tool to extract knowledge from textual resources in the form of Knowledge graphs.

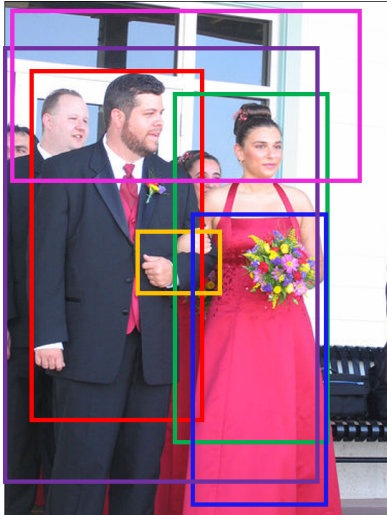
We developed a baseline algorithm called VT-LINKER, which used state-of-the-art tools and techniques in the fields of CV, NLP, and KRR. We used YOLO [77] in the starting, and then MaskRCNN [78] object detector for the implementation of VT-LINKER. These algorithms are trained on COCO [34] dataset. The COCO dataset consists of 328k images, every image has five captions (natural language sentences). The visual objects are labeled with 80 classes (object types) and amount to a total of 2.5 million instances (bounding-boxes) in the images. We linked manually the 80 classes of the COCO dataset to the corresponding instances (classes) in the knowledge base YAGO [13]. To checked the quality of VT-LINKER (baseline) using YOLO and MaskRCNN, we used VTKEL\* and VTKEL datasets. After the detailed experiments using YOLO and MaskRCNN as backbone object detectors, the accuracy of VTKEL problem was not promising. Due to these limitations, we select a better candidate for the object detection task called *Keras-RetinaNet* (KRN) [1].

One of the important tasks in the problem of *VTKEL* is the alignment of visual entities in the image with the textual entities in the text. We developed a supervised algorithm using neural network architecture for the alignment of visual and textual entities' tasks. To train the supervised algorithm, we used the features of visual entities (bounding box), features of textual entities (textual mention), and the background knowledge of visual and textual entities. The details of the algorithm are described in chapter 7.

The following sections provide in detail the background of tools and techniques used during the development of the datasets and algorithms.

---

<sup>1</sup><https://github.com/ZFTurbo/Keras-RetinaNet-for-Open-Images-Challenge-2018>



- C0: **A bearded man**, and **a girl** in a **red dress** are getting married.
- C1: **The group of people** are assembling for **a wedding**.
- C2: **A man** and **woman** dressed for **a wedding function**.
- C3: **A woman** holds **a man's arm** at **a formal event**.
- C4: **A wedding party** walks out of **a building**.

Figure 4.1: Flickr30k-entities dataset single entry. Each image has five captions and the color-coding (for each visual co-reference chain) represents the visual entities in the image with their corresponding textual-mentions in the text.

## 4.1 Flickr30k-entities dataset

In the development of VTKEL datasets, as a starting point, we used the Flickr30k-Entities [5] dataset. Flickr30k-entities<sup>2</sup> is a comprehensive dataset of image-region to textual-phrase correspondences for image description. This dataset consists of 31k images and every image consist of five captions (total 158k captions). Every document<sup>3</sup> is annotated with bounding-boxes (visual mentions) in the image part and noun-phrase (textual mentions) in the captions part. In total, there are 244k coreference chains<sup>4</sup> which are linking the bounding-box objects with the noun-phrases of the same entity in the captions. Figure 4.1 illustrates a single document entry of Flickr30k-entities dataset annotations in details. The textual mentions are categorize into *people*, *body-parts*, *animals*, *clothing*, *instruments*, *vehicles*, *scene*, *other*, and *non-visual* types.

The Flickr30k-Entities dataset has become a standard benchmark for sentence-based image description tasks such as image captioning. The annotations of this dataset are essential for grounded language understanding of visual data and they have allowed the recent progress in the *text-to-image reference resolution* (i.e. noun-phrase localization in an image) and *bidirectional image-sentence retrieval* tasks. The availability of such ground-truth annotations is also a key resource for experimenting in other high-level tasks, involving both visual and textual data, such as *Visual Question Answering* (VQA), multimedia retrieval, and indexing.

<sup>2</sup><https://github.com/BryanPlummer/flickr30kentities>

<sup>3</sup>Here we are considered one document, which consists of an image and five captions.

<sup>4</sup>Coreference chains are the ids, which annotated visual-objects (e.g. the bounding-box of a person in the image) with the textual occurrence of the same entity in five captions. For example, the visual-entity “person” may appear (1) “A man” in caption no.1, (2) “A young boy” in caption no.2, (3) “A player” in caption no.3, and no.5 and (4) “A person” in caption no. 4.

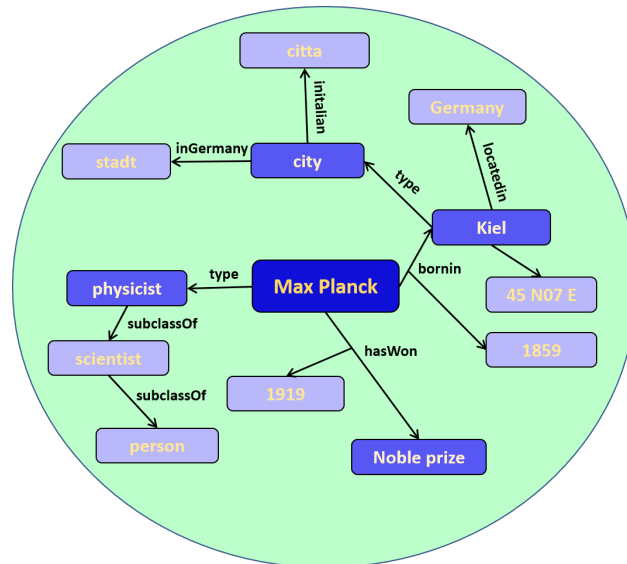


Figure 4.2: A segment of *knowledge graph* from YAGO, which represent fact about *Max Planck* in the RDF form.

## 4.2 YAGO

YAGO<sup>5</sup> (Yet Another Great Ontology) is a large-scale semantic knowledge base automatically derived from several data sources, including Wikipedia (e.g., categories, redirects, infoboxes, etc.), WordNet (e.g., synsets, hyponymy, etc.), and GeoNames. Particularly, in YAGO an entity (e.g., person, organization, city, etc.) is associated with its corresponding page in Wikipedia, and facts about the entity are extracted from the page *infobox*. Figure 4.2 represents a segment of knowledge graph (fact) from YAGO for a well-known scientist *Max Planck*. In this knowledge graph, the nodes represent entities and arc attached relationships (properties) between nodes. The entities of the knowledge graph are typed according to the classes and organized in a class/sub-class hierarchy obtained from the categories of the *WordNet synset taxonomy*.

The current version (v3) of YAGO contains more than 350K classes and 17M entities, with over 150M facts about them. YAGO is special in several aspects:

1. YAGO combines the taxonomy of wordnet with the richness of the Wikipedia category system, which assigns the entities to more than 350K classes.
2. The accuracy of YAGO has been manually evaluated with confirmed accuracy of 95%. Every relation is annotated with its confidence value.
3. YAGO is anchored in time and space, which are linked to a temporal dimension.
4. YAGO extracts and combines entities and facts from 10 different languages.
5. YAGO also has thematic domains such as “music” and “science” from Wordnet domains.

<sup>5</sup><https://yago-knowledge.org/>

### 4.3 PIKES

PIKES<sup>[6]</sup> is state-of-the-art frame-based framework for extracting knowledge from the natural language text resource. PIKES extracts entities and complex relations between entities by identifying semantic frames in a text, i.e., events and situations describing relations between entities (i.e., frame participants). PIKES works in two phases. In the first linguistic feature extraction phase, an RDF graph of mentions is obtained by running and combining the outputs of several state-of-the-art NLP tools, including Stanford CoreNLP<sup>[7]</sup> (tokenization, lemmatization, part-of-speech tagging, temporal expression recognition and normalization, named entity recognition and classification, coreference resolution, parsing), DBpedia Spotlight<sup>[8]</sup> (entity linking), UKB<sup>[9]</sup> (word sense disambiguation), Semafor<sup>[10]</sup> and Mate-tools<sup>[11]</sup> (semantic role labeling).

In the second knowledge distillation phase, the mention graph is transformed into an RDF knowledge graph through the evaluation of mapping rules, using the RDF-pro<sup>[12]</sup> [14] tool for RDF processing. In the DBpedia-YAGO and WordNet-YAGO mappings, entities resulting in the final RDF knowledge graph are typed according to the classes in YAGO. In the mention graph, each node uniquely identifies an entity of the world, event, or situation, and arcs represent relations between them (e.g., the participation and role of an entity in an event). Figure 4.3 represents the mention graph (two phases) of caption passing through the PIKES tool in detail. In the caption, *woman*, *court*, *ball* and, *tennis* are the entity nodes and linked with YAGO class *Woman110787470*<sup>[13]</sup>, *Court108329453*, *Ball102778669*, and *Tennis100482298* respectively. However *come* and *young* are the event/situation nodes of the knowledge graph. In Figure 4.3, each arc of the knowledge graph represents relations between the nodes. For the instance, *age* and *arriving* are the relations between (*young*, *woman*) and (*ball*, *come*) respectively.

### 4.4 Object Detection

Visual-object detection is a well studied task in the community of CV, which involves identifying the presence, location, and type of one or more than one object in a given photograph (image) or videos (frames). It is not a simple task, and requires the development of methods to solve sub-tasks of object recognition (e.g. where the visual-objects are), object localization (e.g. what are the extent of the objects), and classification (e.g. what are the type of objects). To solve the problem of VTKEL, one of the important tasks is the visual entities (objects) detection and linking to a knowledge base (YAGO Ontology in our case). We used state-of-the-art object detectors YOLO<sup>[77]</sup>, Mask-RCNN<sup>[78]</sup> and KRN<sup>[14]</sup> for visual-entities detection and classification tasks in the VTKEL problem. The details of these object-detectors

<sup>6</sup><http://pikes.fbk.eu/>

<sup>7</sup><https://nlp.stanford.edu/software/corenlp.shtml>

<sup>8</sup><https://www.dbpedia-spotlight.org/>

<sup>9</sup><http://ixa2.si.ehu.es/ukb/>

<sup>10</sup><http://www.cs.cmu.edu/~ark/SEMAFOR/>

<sup>11</sup><https://code.google.com/archive/p/mate-tools/>

<sup>12</sup><http://rdfpro.fbk.eu/>

<sup>13</sup>The URI <http://dbpedia.org/class/yago/Woman110787470> stored the background knowledge information of class *Woman110787470* from YAGO ontology in the form of structured data (i.e. RDF triples).

<sup>14</sup><https://github.com/ZFTurbo/Keras-RetinaNet-for-Open-Images-Challenge-2018>

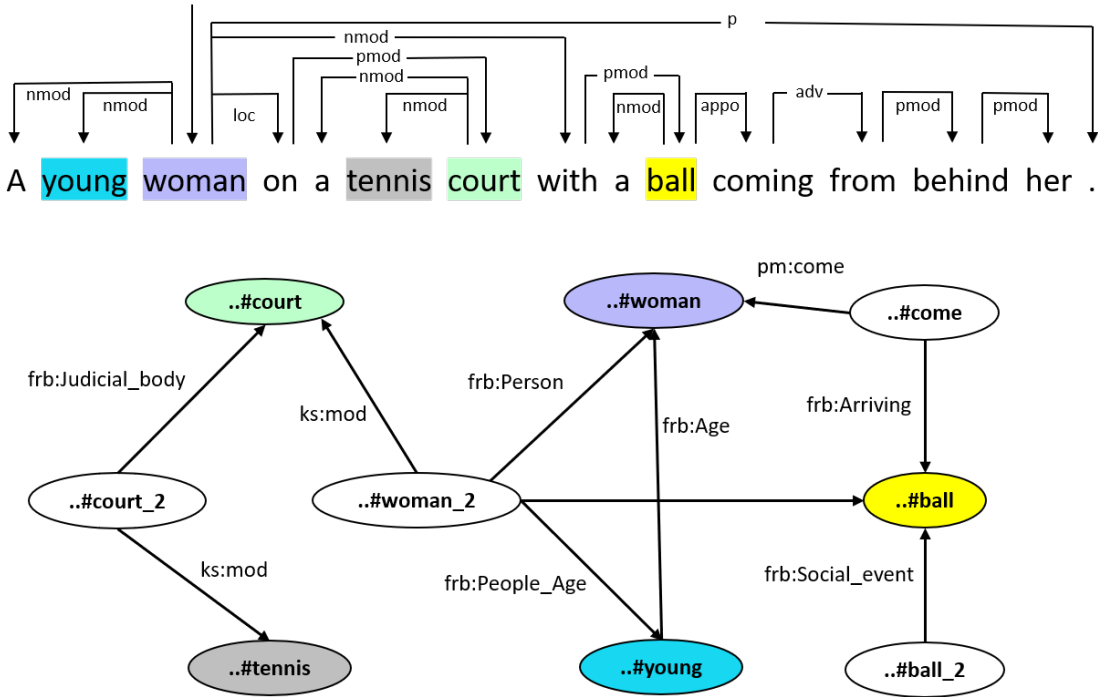


Figure 4.3: Passing caption “A young woman on a tennis court with a ball coming from behind her.” through the PIKES tool. In the mentioned graph, nodes represent the entity, event, or situation and arcs represent *relations* between them.

are listed below.

#### 4.4.1 YOLO

The “You Only Look Once” or YOLO<sup>15</sup>, is the family of end-to-end deep learning models for fast object detection and classification in real-time. The YOLO object detector involves a single deep convolutional neural network, which splits the input image into a grid of cells and each cell directly predicts a bounding box and object classification. In the result, a large number of candidate bounding boxes are consolidated into a final prediction by a post-processing step. There are three main versions of YOLO objects detector that is YOLOv1, YOLOv2, and YOLOv3. We use a version of YOLO that detects and classifies objects according to 80 pre-defined categories<sup>16</sup>, such as person, car, dog, etc. YOLO reasons globally about the image unlike sliding window or proposal-based techniques, while making predictions on the basis of seeing the entire image and encodes contextual information about classes as well as their appearance. We used the third version of YOLO, with an image resolution of 416x416 (i.e. YOLOv3-416).

#### 4.4.2 Mask-RCNN

The Mask-RCNN is one of the most recent variations of the family models and supports both object detection, classification, and segmentation. We utilized the object *detection*, and *classification* portion of Mask-RCNN in this thesis. The Mask-RCNN achieves state-of-the-art results on CV benchmark datasets. We used the pre-

<sup>15</sup><https://pjreddie.com/darknet/yolo/>

<sup>16</sup>c.f. COCO dataset <https://bit.ly/2KuioA0>



trained model of mask-RCNN on the COCO dataset, which detects and classifies objects according to the 80 categories (classes).

In Mask-RCNN, the Region-based Convolution-Neural-Network (R-CNN) approach is used for bounding-box object detection to obtain a number of candidate object regions. The R-CNN approach also evaluates convolution networks independently on each region of interest (RoI). The MaskR-CNN is conceptually simple, efficient, and works on three branches. For each candidate object, a class label and a bounding-box offset are generated using the two branches, while a third branch outputs the object mask. The key element of Mask-RCNN is pixel-to-pixel alignment, which is the main missing piece in Fast R-CNN [79] and Faster R-CNN [80].

#### 4.4.3 Keras-RetinaNet

At the start, we used YOLO and Mask-RCNN object detectors for the development of a baseline algorithm called *Visual-Textual-Knowledge-Entity Linker* (VT-LINKER). These object detectors performed efficiently on the categories of *people*, *vehicles* and *animals*. However, they failed to make predictions on visual entities in the categories of *clothing*, *human body part*, and *other* from the VTKEL dataset. On the basis of these shortcomings, the quality of our baseline algorithm was not good. The second problem was the limited classes (80) of the COCO dataset on which these object detectors are trained. To solve these problems, we select a better object detector which can make prediction on more than 80 classes and perform very well in all the categories of the VTKEL dataset.

The *KRN* [15] is a single-stage detector but at the cost of being slower. It is composed of a unified network with a backbone network and two task-specific sub-networks. The backbone is responsible for computing a convolutional features map over an entire input image and is an off-the-self convolutional network. The first subnet performs convolutional object classification on the backbone's output; the second subnet performs convolutional bounding box regression. The two subnetworks feature a simple design that we propose specifically for one-stage, dense detection. *KRN* uses a feature pyramid network to efficiently detect objects at multiple scales and introduces a new loss, the Focal loss function, to alleviate the problem of the extreme foreground-background class imbalance.

We used *KRN* object detector, trained on the *Google-Open images* (GOI) [17] dataset. This dataset has approximately 17 million images, which are annotated with 121.95 million bounding boxes over 500 categories (classes). We linked the 500 classes of *GOI* dataset to their corresponding class instance in the knowledge base YAGO manually. This object detector has covered the majority of the classes of the VTKEL dataset. We achieved very good results during the evaluations of VT-LINKER algorithm. It also covers the categories of *human-body part*, *clothing*, *instruments*, and *other* in an efficient way which was missed by YOLO and Mask-RCNN in the first version of the baseline.

## 4.5 VGG16

VGG16 [16] is a convolution neural network (CNN) architecture, which was used for the competition of *Large Scale Visual Recognition Challenge* (ILSVR) using Imagenet [81] dataset. It is working on convolution layers of 3x3 filter with a stride

<sup>17</sup><https://opensource.google/projects/open-images-dataset>

1 and always used the same padding and maxpool layer of 2x2 filter of stride 2. It considers one of the excellent vision model architectures to date. It follows the arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end, it has 2 FC (fully connected layers) each with 4096 nodes and followed by a softmax classifier for the output. The 16 in VGG16 refers to 16 layers that have weights.

We used VGG16 architecture to find the visual features from the bounding boxes and used these features to train our supervised neural network to solve the problem of *visual* and *textual entity alignment* task. We passed the ground-truth bounding boxes of the VTKEL dataset through VGG16 architecture for visual features. We detach the last FC layers and received a 4096 features matrix to be used as visual features.

In this chapter, we described the Flickr30k-Entities dataset, tools, object detectors, and approaches used in the development of our *VTKEL* dataset, baseline algorithm, and supervised way of solving the task of Visual-textual entity alignment. The importance of this chapter is to have in-depth knowledge of these approaches to understand better the rest of the chapters.



## Chapter 5

# VTKEL Dataset

In the previous chapter, we described in detail the background of state-of-the-art dataset (Flickr30k-Entities), tools, and techniques used in this thesis. In this chapter, we are introducing state-of-the-art dataset, its development, and the evaluation procedure for checking the quality. We also described the representation schema in the form of an RDF graph for storing the information of the dataset.

The scientific community of Natural Language Processing (NLP) and Computer Vision (CV) have devoted a reasonable effort in investigating text and image processing. In this thesis, we also added the community of Knowledge Representation (KR) by adding the background knowledge of visual and textual resources, extracted from a knowledge-base (Ontology). One can find a bunch of works and datasets which consist of images, text (caption), and the intersection of visual and textual modalities. However, there is not a single work that combines the three modalities necessary for our dataset.

To build *Visual-Textual-Knowledge Entity Linking* (VTKEL) dataset. We start from the *Flickr30k-Entities dataset* [5], which provides documents composed of a picture and five captions, describing the contents of an image. The Flickr30k-Entities dataset, also provides the annotation of coreference chains, i.e., linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in *automatic image description* and *grounded language understanding*. They enable us to define a new benchmark for the localization of textual entity mentions in an image.

VTKEL was automatically derived by (i) applying a knowledge graph extraction tool (PIKES [10]) to the textual captions of the Flickr30k-Entities dataset, and (ii) leveraging the picture-caption coreference annotations contained in the original Flickr30k-Entities dataset. As a result, visual and textual mentions of each picture and captions are annotated and aligned to entities typed with classes from YAGO [13], a well-known Semantic Web (SW) ontology. Such a dataset is essential for providing training and evaluation material for automatic algorithms tackling the VTKEL task.

To check the quality of our automatically developed VTKEL [4] dataset, we randomly sample 1000 documents and developed a subset of VTKEL dataset called VTKEL1k\*[2]. For every caption, we manually checked the correctness of the YAGO

---

<sup>1</sup>The VTKEL30k dataset can be download from the link: [https://figshare.com/articles/VTKL\\_dataset\\_file/7882781](https://figshare.com/articles/VTKL_dataset_file/7882781)

<sup>2</sup>The VTEKL1k\*(1000 documents) dataset can be download from the link: [https://figshare.com/articles/VTKEL\\_dataset/10318985](https://figshare.com/articles/VTKEL_dataset/10318985)

class associated with each textual-mentions found by PIKES. The details are described in the next sections.

## 5.1 Related work

There are several datasets available that combine text and images, however, none of them have all the three components (i.e. NLP, CV, and KR) necessary for the VTKEL task. A review of language and vision datasets are presented by Francis et al. [32]. They categorized these datasets with respect to the tasks on image to language and vice-versa. *VisualGenome* [33] is an extremely large dataset that contains pictures in which objects are annotated with their types, attributes, and relationships. Annotations are mapped to *WordNet*<sup>3</sup> synsets. Objects can also be annotated with a short sentence that describes some qualitative property of the object. E.g., "The girl is feeding the elephant" or "a handle of bananas". However, there is no alignment between the objects mentioned in these phrases and the objects shown in the picture. E.g., there is no bounding box for the object "bananas" or "elephant". The *Visual Relationship Dataset* (VRD) [12] is a dataset of images annotated with bounding boxes around key objects. Furthermore, VRD contains annotations about relationships between objects in the form of triplets  $\langle \text{object\_type}, \text{relation}, \text{subject\_type} \rangle$  describing the scene. Examples of annotations are  $\langle \text{man}, \text{riding}, \text{bicycle} \rangle$  and  $\langle \text{car}, \text{on}, \text{road} \rangle$ . However, these annotations are not aligned to any knowledge base. The *Microsoft COCO dataset* [34] contains pictures associated with five captions. They are annotated with objects regions of any shape (not simple bounding boxes) and each region is assigned with an object type. This dataset does not contain any information about the relation between object regions, and the relation between regions and mentions in the captions. *Conceptual Captions* [35] is a recently introduced dataset that has been developed for automatic image caption generation. It contains one order of magnitude more items than Microsoft COCO. It is a realistic dataset as images with captions have been automatically extracted and filtered from the web. However, there is no visual/textual mention annotation and visual textual entity linking. *VizWiz* [37] is a dataset generated by mobile users by talking pictures from their mobile, with image descriptions (captions), and recorded spoken questions about the picture. This dataset is very well known nowadays for developing models to assist people who are blind to overcome their daily visual challenges. However, there is no background knowledge (Ontological) that described either pictures, captions, or voices contents of the dataset. Grew et al. [38] recently proposed a dataset called *GQA* for real-word visual reasoning and compositional questions answering dataset to leverage key shortcoming in the Visual-Question-Answering (VQA) datasets. It consists of 113K images and 22M questions of assorted types with answers. In their dataset, the images, questions and corresponding answers are all represented by semantic matching. Each image is annotated with a dense Scene Graph, which representing objects, attributes and relations it contains. Each question is associated with a functional-program, which lists the series of reasoning steps needed to be performed to reach into the answer. Each answer is enlarged with both visual and textual justification, denoting (pointing) to the corresponding region in the image. The dataset has annotated region of image with the portion of text, which missing the annotation of nouns in the text with the corresponding object image. *Google Open Images* [40] dataset consists of

---

<sup>3</sup><https://wordnet.princeton.edu/>

9.2 Million images with unified annotations for *image classification*, *object detection* and, *visual relationship detection*. This dataset offers large scale across several dimensions: (i) 30.1M image-level labels for 19.8k concepts, (ii) 15.4M bounding boxes for 600 object classes, and (iii) 375k visual relationship annotations involving 57 classes. For object detection in particular, the authors provide 15× more bounding boxes than the next largest datasets (15.4M boxes on 1.9M images). The images often show complex scenes with several objects (8 annotated objects per image on average). They annotated visual relationships between them, which support visual relationship detection, an emerging task that requires structured reasoning. This dataset is state-of-the-art and recently widely used for competitions in the area of object detection, classification and visual-relation detection. However, there is no background or ontological knowledge associated with this dataset.

From the above analysis and literature, it becomes clear that there is not a single dataset, which combines images (vision), texts (language), and the Ontological knowledge of them. This justifies the development of a ground truth dataset, which not only combined images and corresponding text but also their background knowledge in the form of Ontological (structured) data.

## 5.2 A Data Model for multi-modal knowledge extraction

The VTKEL dataset contains multiple images, each associated with five textual captions. Visual and textual entity mentions in images and captions are annotated with entities types/class of the YAGO ontology, and with coreferring mentions annotated with the same entity. All the resource is represented in RDF (Resource Description Framework)<sup>4</sup> using the representation schema.

In order to represent the mention and entity content, as well as its links to the pieces of text or image where it derives from, we encode all the information in an RDF model organized in three distinct yet interlinked representation layers — *Resource*, *Mention*, and *Entity* — extending the data model proposed in [82] (where it serves as data model for a framework —the KnowledgeStore— supporting the interlinking of unstructured and structured content), and later refined in [14]. The extension mainly regards the representation of visual content, and in particular visual textual mentions. Figure 5.1 provides an overview of the main model elements relevant to the VTKEL. An instantiation of the data model is shown in Figure 5.2.

### 5.2.1 Resource Layer

This is the textual or image content from which knowledge was extracted. It consists of text or image resources identified by URIs. Each resource may be characterized by metadata (e.g., the `dct:title` or the document creation time `dct:created`) that are expressed with standard vocabularies (e.g., Dublin Core<sup>5</sup>) and may be exploited during processing. Resources may be complex objects (e.g., documents) composed of (via `dct:isPartOf`) other resources (e.g., images, textual documents).

### 5.2.2 Mention Layer

This layer consists of mentions of entities (class `ks:EntityMention`). As shown in Figure 5.1, we identify two (disjoint) main types of mentions: `ks:Textual-`

<sup>4</sup><https://www.w3.org/RDF/>

<sup>5</sup><http://dublincore.org/>

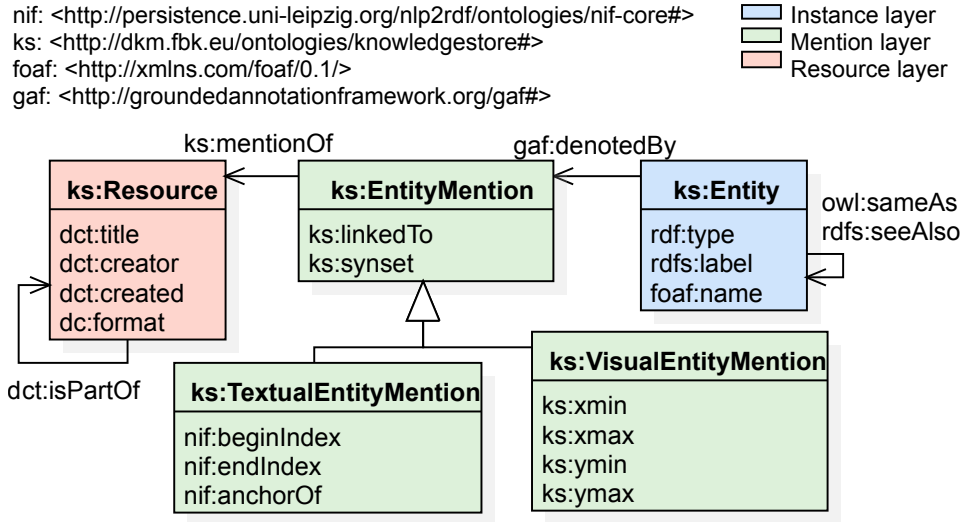


Figure 5.1: Data model overview (OWL ontology shown in UML notation).

`EntityMention` and `ks:VisualEntityMention`. The first corresponds to mentions of entities in textual resources. Textual mentions are characterized by attributes, whose values are extracted from NLP tools. In particular, the NLP Interchange Format (NIF) [83] is used for anchoring mentions in the resource text, using properties `nif:beginIndex`, `nif:endIndex` and `nif:anchorOf` (the mention textual extent) and minting mention URIs according to the RFC 5147 URI scheme (c.f., textual mention `vtkel:resource65567C0#char=0,13` in Figure 5.2). `ks:VisualEntityMention` corresponds to mentions of entities in image resources. We associate a visual mention to the bounding box around the entity identified by CV tools. In particular, `ks:xmin`, `ks:xmax`, `ks:ymin`, and `ks:ymax` identify the coordinates of the top-left and bottom-right corner of the bounding box in the image. Similarly to the RFC 5147 URI scheme for textual mentions, also URIs of visual entity mentions encode these coordinates (c.f. visual mention `vtkel:resource65567I#xywh=24,63,208,500` in Figure 5.2, where `ks:xmin=24`, `ks:xmax=208`, `ks:ymin=63`, and `ks:ymax=500`).

### 5.2.3 Entity Layer

The entity layer describes the things of interest contained in a textual or visual resource, abstracting from the actual ways they are expressed in the text or they appear in an image. Its main objects are instances of entities (class `ks:Entity` in Figure 5.1) such as *persons*, *body parts*, *objects*, and so on. Entities result from clustering `ks:EntityMentions` denoting the same referent, as determined using (visual-textual) co-reference resolution techniques that exploit the information available at the mention level (e.g., from multiple mentions of a person in text or the corresponding bounding boxes in images, a single instance uniquely identifying the person is obtained, possibly by smushing `owl:sameAs` entity links). That is, the entity layer compacts the knowledge coming from the mention layer, where content is spread and redundantly replicated over several mentions. For entity types, we use in particular the classes defined in YAGO [13], as well as the types used in the Flickr30k-Entities dataset.

Mention and Resource layers are related by property `ks:mentionOf` that links

a `ks:hadMention` to the `ks:Resource` it belongs to. Entity and Mention layers are related by property `gaf:denotedBy`<sup>6</sup> that links a `ks:Entity` to the `ks:Entity-Mention` denoting it (e.g., `vtkel:resource/65567C0/#man` is linked to the mention `vtkel:resource/65567C0/#char=10,13` in Figure 5.2).

### 5.3 VTKEL data model instantiation

The VTKEL dataset contains commented images annotated with all the ground truth of the simple tasks composing the visual-textual-knowledge entity linking task. To support the integration of this resource in the semantic web and its access via standard SPARQL<sup>7</sup> language, we propose to represent this information in RDF using the representation schema described in the previous section. In details, every document  $d = \langle d_i, d_t \rangle$  of the dataset is annotated with the following information:

- *Visual mentions* represented as the bounding boxes around the objects detected in  $d_i$ ;
- *Links* between *visual mentions* and the corresponding entity in the ontology, represented by the relation `ks:denotedBy`;
- *Textual mentions* represented as the text spans around the text referring to an entity in  $d_t$ ;
- *Links* between *textual mentions* and the corresponding entity in the ontology, represented by the relation `ks:denotedBy`;
- *Ontological types* of each entity, represented with the `rdf:type` relation;
- The classes from the Flickr30k-Entities dataset (e.g. people, clothing, instruments, bodypart etc..) are linked with the corresponding classes of YAGO by using `owl:equivalentClass` relation.
- *Coreference* information about visual-textual mentions and textual-textual mentions, represented as `owl:sameAs` link between entities.

For every complex document, we associate a unique URI, e.g., `vtkel:resource/65567`. The URI of the corresponding image is `vtkel:resource/65567I`, and the URIs of the associated five captions are `vtkel:resource/65567C<i>` for  $i \in \{0, \dots, 4\}$ . The URIs of image and captions are related to the URI of the complex document via `dct:isPartOf` property. The segment of the graph shown in blue in Fig. 5.2, is obtained by encoding in RDF the annotation of Flickr30k-Entities dataset, the segment shown in brown is obtained by automatically processing each caption with PIKES and then extracting the information about entities and entity mentions.

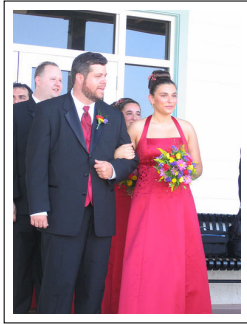
The `owl:sameAs` link among the entities of Flickr30k-Entities dataset and those introduced by PIKES are automatically computed with the following heuristic.

Since PIKES annotates with a finer granularity than the Flickr30k-Entities dataset, it can happen that a Flickr30k-Entities dataset mention contains more than one PIKES mention. For instance, consider caption  $C_1$  of Figure 5.2, the textual mention “*The group of people*” that refers to the

<sup>6</sup><http://groundedannotationframework.org/>

<sup>7</sup><https://www.w3.org/TR/rdf-sparql-query/>





- $C_0$ : A bearded man, and a girl in a red dress are getting married.
- $C_1$ : The group of people are assembling for a wedding.
- $C_2$ : A man and woman dressed for a wedding function.
- $C_3$ : A woman holds a man's arm at a formal event.
- $C_4$ : A wedding party walks out of a building.

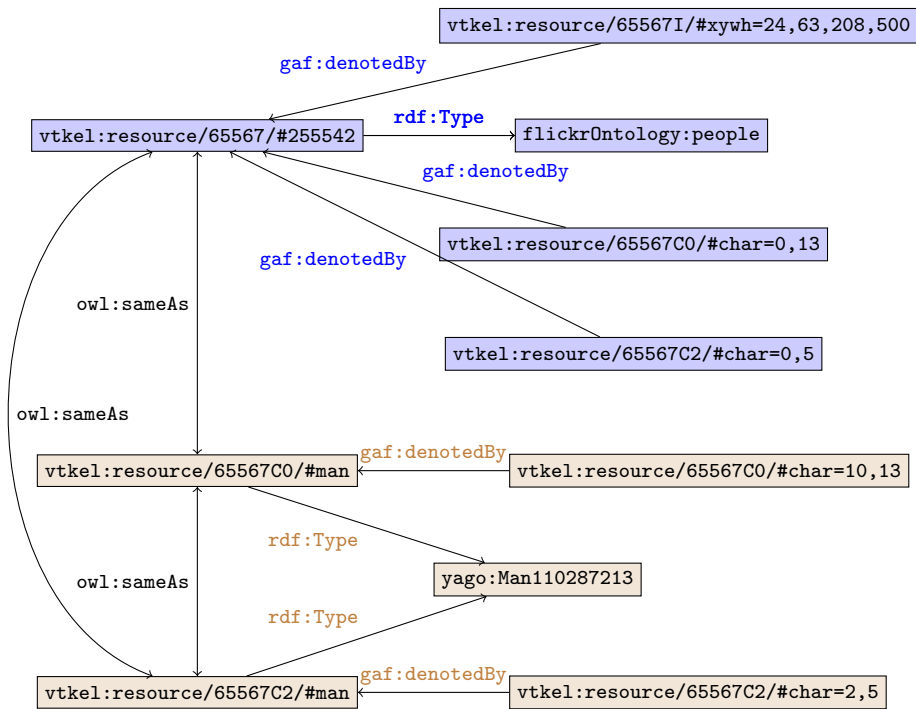


Figure 5.2: A single document of the *VTKEL* dataset, which consists of an image, five captions (in the upper part), and co-reference chain with  $ID = 255542$  is represented in the RDF graph (in the lower part).

Table 5.1: PIKES result over the caption "A man is inside a truck looking out with his left arm in front of a door".

Mention	YAGO class	WordNet gloss
arm	Weapon104565375	any instrument ... used in fighting ...
door	Doorway103224032	the entrance ... you enter ... a room
front	Front108573472	the side that is forward or prominent
he	Homo102472293	any living or extinct member ...
truck	Truck104490091	an automotive vehicle ...

Flickr30k entity `vtkel:resource/65567/#255544` contains the two textual mentions "group" and "people" identified by PIKES and linked to two entities `vtkel:resource/65567C1/#group` and `vtkel:resource/65567C1/#people`. From the syntactic analysis of the noun phrase, we obtain that the *head* of the noun phrase is "group" while "people" is a *modifier*. So we decided to align only the *head* of the noun phrase to the Flickr30k entity, that is to add the relation `owl:sameAs` between `vtkel:resource/65567C1/#group` and `vtkel:resource/65567/#255544`. We also link the second entity (i.e. `vtkel:resource/65567C1/#people`) to `vtkel:resource/65567/#255544` by relation `ks:hasPart`, which described that the second entities is the part of noun-phrase (i.e. `vtkel:resource/65567/#255544`). The same heuristic is applied to the Flickr30k-Entities dataset annotations "wedding function", "man's arm", "wedding party" of the captions shown in Figure 5.2.

The Flickr30k-Entities dataset has some limitations in the annotations of co-reference chains. There are cases in which an object (e.g., a cup of coffee) is shown in the image (i.e. in the background), but this object is not described in the corresponding caption, and thus no alignment can be drawn. Similarly, there are cases in which an entity is described in a caption but not shown in the image. These shortcomings should be taken into consideration when evaluating the performances of systems solving the VTKEL task. The second limitation of the Flickr30k-Entity dataset is by exploiting mostly common nouns in the description (textual data) of images by missing the proper-nouns in the images (visual data).

The VTKEL dataset is stored in a unique, comprehensive Linked Data resource, available at [https://figshare.com/articles/VTKL\\_dataset\\_file/7882781](https://figshare.com/articles/VTKL_dataset_file/7882781), in which pictures<sup>8</sup>, associated captions, mentions, entities, and links to entities are all encoded as RDF triples.

## 5.4 Evaluations

The VTKEL30k dataset consists of millions of RDF triples, and manually evaluating these triples is a challenging task, which requires a lot of human efforts. We followed the same methodology used for the evaluations of the YAGO knowledge base by randomly sampling 1000 entries from the VTKEL30k dataset. The VTKEL1k\* consists of 20,356 textual entity mentions distributed over 5000 captions, with an average of amount 4 entities per caption. For every caption processed by PIKES, we manually checked the correctness of the YAGO class associated with each textual mention found by PIKES. To assess the correctness of the YAGO class, we looked

<sup>8</sup>All the pictures of Flickr30k-Entities dataset can be downloaded from <http://hockenmaier.cs.illinois.edu/DenotationGraph/>

at the textual description of the class and the gloss of the corresponding WordNet synset (if any) from which the class was derived. For example, the resultant running of PIKES over the caption “A man is inside a truck looking out with his left arm in front of a door”, is shown in the Table 5.1<sup>9</sup>. All the detected YAGO classes for the mentions in the given sentence are correct except the one for “arm”. Concerning mention “arm”, the correct annotation would have been `dbyago:Arm105563770`, with WordNet gloss “a human limb”.

For 1000 documents, PIKES recognized 19440 entity mentions correctly and make alignments with the YAGO Ontology. Among all the alignment, we found a total of 916 incorrectly linked mentions. Further examples of the wrong mention with YAGO class (i.e. linking) produced by PIKES are shown in Table 5.2.

These errors are mainly due to the incorrect *word sense disambiguation*: e.g., in some cases e.g. “bus” was linked to the concept of *computer bus*, instead of that of *coach*, and “arm” to *weapon* instead of *body-part*. The construction of VTKEL1k\* dataset allows us also to estimate the *error-rate* of the larger VTKEL30k dataset. In particular, we found no missing link (i.e., recall is 100%) and 916 incorrectly linked mentions, which amounts to *Precision* = 0.955, *Recall* = 0.893, and *F1* = 0.923. We believe that an *error-rate* of 5% is physiological also in manually developed datasets, and therefore we believe that the VTKEL30 dataset can be reasonably considered a ground truth.

While performing the evaluation on the reduced subset, we also manually corrected all the wrong alignments found, by replacing the wrong YAGO classes with the correct ones. As a “by-product” of the evaluation, we obtained manually validated dataset containing gold visual-textual-entity alignments for 1000 complex documents, separately released as part of the VTKEL30k dataset.

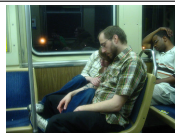




## 5.5 Conclusion

In this chapter, we introduced two state-of-the-art datasets called VTKEL1k\* and VTKEL30k, their development and evaluation in detail. These datasets consist of images, textual descriptions of the images, annotation of bounding-boxes with the corresponding noun-phrases in (image) captions, and the textual entity mentions are linked with YAGO ontology. We used PIKES a state-of-the-art tool for extracting background knowledge from the textual resources. We evaluated the quality of the dataset by sampling VTKEL1k\* dataset, which consists of 1000 documents. This evaluation also allows us to estimate the error rate of the big dataset called VTKEL30k. We believe that the VTKEL dataset can be reasonably considered a ground truth for the evaluation of algorithm(s) which solve the problem of *Visual-Textual-Knowledge-Entity linking*. In the next chapter, we introduce an unsupervised algorithm called VT-LINKER, which solves the problem of the Visual-Textual-Knowledge-Entity linking by using both VTKEL1k\* and VTKEL datasets.

---

<sup>9</sup>This can be checked by submitting the sentence to the online demo of PIKES at <https://pikes.fbk.eu/>, and selecting the tab “instances”.

Table 5.2: Examples of wrongly aligned YAGO Classes obtained by processing with PIKES, together with the manually corrected ones. The underline-word in textual mention column is processed wrongly.

Image	Textual mention	Incorrect Yago Class	Correct YAGO class
	Couple is sleeping in a <u>bus</u> .	Busbar102924713: subclass of conductor, used in electronic chip	Bus102924116: a vehicle carrying many passengers.
	A man with a tattoo on his right arm is playing a guitar on stage at a <u>bar</u> .	Bar102788689: a rigid piece of metal or wood	Bar102789487: a room or establishment where alcoholic drinks are served over a counter
	A man is dressed as a movie character holding a gun in the <u>lobby</u> of a movie theater .	Lobby108375526: political unit	Anteroom102715513: a large entrance or reception room or area
	In this <u>picture</u> , a child is playing with a large blue ball .	Movie106613686: a form of entertainment that enacts a story by sound and video	Picture103931044: a visual representation of objects
	Three people are pushing a heavy <u>machine</u> .	Car102958343: an automobile car	Machine103699975: electrical devices that transmit energy

## Chapter 6

# The VT-LinKEr Algorithm

In the previous chapter, we described a state-of-the-art dataset called VTKEL (*Visual-Textual-Knowledge-Entity Linking*). In this chapter, we are introducing an unsupervised algorithm called VT-LINKER (*Visual Textual Knowledge Entity Linker*) in detail.

The scientific community of Computer Vision (CV) and Natural Language Processing (NLP) are collectively trying to efficiently solve the reasoning capabilities of visual informed systems. They are mostly focusing on the problem of recognizing what objects are present in image and described in text (i.e. phrase-grounding) [5, 41], vision-and-language tasks such as captioning [84, 85], visual question answering [43], and visual dialogue systems [11]. However, their approaches mainly depend on a limited range of details coming from visual and textual resources. The knowledge part (i.e. background knowledge) of visual and textual resources, which play a key role in the development of common-sense and reasoning [45] phase are missing in their approaches.

The VT-LINKER algorithm solving the VTKEL task by combining the state-of-the-art tools and techniques in the fields of NLP, CV and Ontological reasoning. Given a document composed of text and images, VT-LINKER applies an object detector to the image part, resulting in a set of bounding boxes labelled with classes of the ontology. Each bounding box is called visual mention and the corresponding object, which is an instance of the class label, is called visual entity. In parallel, VT-LINKER processes the text with a tool for entity recognition, which labels the noun phrases with classes of the ontology. The recognized noun phrases are called textual mentions and the corresponding instances of the ontological class are textual entities. Finally, the VT-LINKER attempts to link visual and textual mentions which correspond to the same entity. This final task is done by exploiting ontological knowledge about class/sub-class hierarchy, and similarity information available in the textual mentions.

The structure of this chapter is: in section 6.1 we represented the state-of-the-art approaches that combining vision ( $d_i$ ) modality with natural language text ( $d_t$ ). In section 6.2, we described the VT-LINKER in details, and in section 6.3, we performed the evaluations experiments.

### 6.1 Related Work

There is a long research history of investigating the intersection and integration of vision and language. The NLP and CV scientific communities are trying to solve various tasks such as textual grounding [5, 41], visual question answering [43], visual

reasoning [44, 45], and various models have been developed to solve them.

For an exhaustive survey of the approaches in the area of entity information extraction and linking, we refer the reader to [46]. In particular: [31] exploits natural language descriptions of a picture in order to understand the content of the scene itself. The proposed approach solves the image-to-text coreference problem. It successively exploits the visual information and visual-textual coreference previously found to solve coreference in text. The work described in [47, 48] tackles the problem of ranking the concepts from the knowledge base that best represents the core message expressed in an image. This work involves the three elements: Image, Text, and Knowledge, but it does not provide information about the entities mentioned in the text and shown in the image. The approach in [49] adapts Markov Random Fields to represent the dependencies between what is shown in the frames of videos about the wild-life animal and the subtitles. The main objective is to detect the animal shown in a frame, and the mentions of animal in the subtitle. The set of entities are the animal names available in WordNet [50]. Object detection is not performed: the approach assumes that only one animal is shown in a frame, and the vision part consists of image classification. Furthermore, no background knowledge about animals is used. [27] proposes a basic framework for visual entity linking to DBpedia and Freebase. The approach involves also textual processing since the link of bounding boxes to DBpedia and Freebase entities is found passing through an automatically generated textual description of the image. The approach uses the Flickr8k dataset, which is a subset of the Flickr30k-Entities dataset. A combination of textual coreference resolution and linking of image and textual mentions is described in [52] with the objective of solving the problem of assigning names to people appearing in TV-show.

From the literature, there is not a single comprehensive approach corresponding to solve the problem of VTKEL task collectively. To the best of our knowledge, VT-LINKER is the first algorithm that extracts visual, and textual entity mentions from images and texts and jointly linking them to their entities mentions in the knowledge base.

## 6.2 Algorithm

VTKEL is a complex task: for every input document, composed of some text, an image, and knowledge base, it produces a set of assertions (RDF triples to be added to the A-box of the knowledge base), each belonging to one of the following five types:

- VMD *Visual mention detection triples*:  $\langle e, \text{isDenotedBy}, vm \rangle$  the entity  $e$  is denoted by the visual mention  $vm$  ( $vm$  is a bounding box);
- VET *Visual entity typing triples*:  $\langle e, \text{hasType}, c \rangle$ , the entity  $e$ , corresponding to a visual mention, is an instance of the knowledge base concept  $c$ ;
- TMD *Textual mention detection triples*:  $\langle e, \text{isDenotedBy}, tm \rangle$  the entity  $e$  is denoted by the textual mention  $tm$  ( $tm$  is a portion of text);
- TET *Textual entity typing triples*:  $\langle e, \text{hasType}, c \rangle$ , the entity  $e$ , corresponding to a textual mention, is an instance of the knowledge base concept  $c$ ;
- VTC *Visual Textual Coreference triples*:  $\langle e, \text{sameAs}, e' \rangle$ , the two entities  $e$  and  $e'$  denotes the same real world entity.

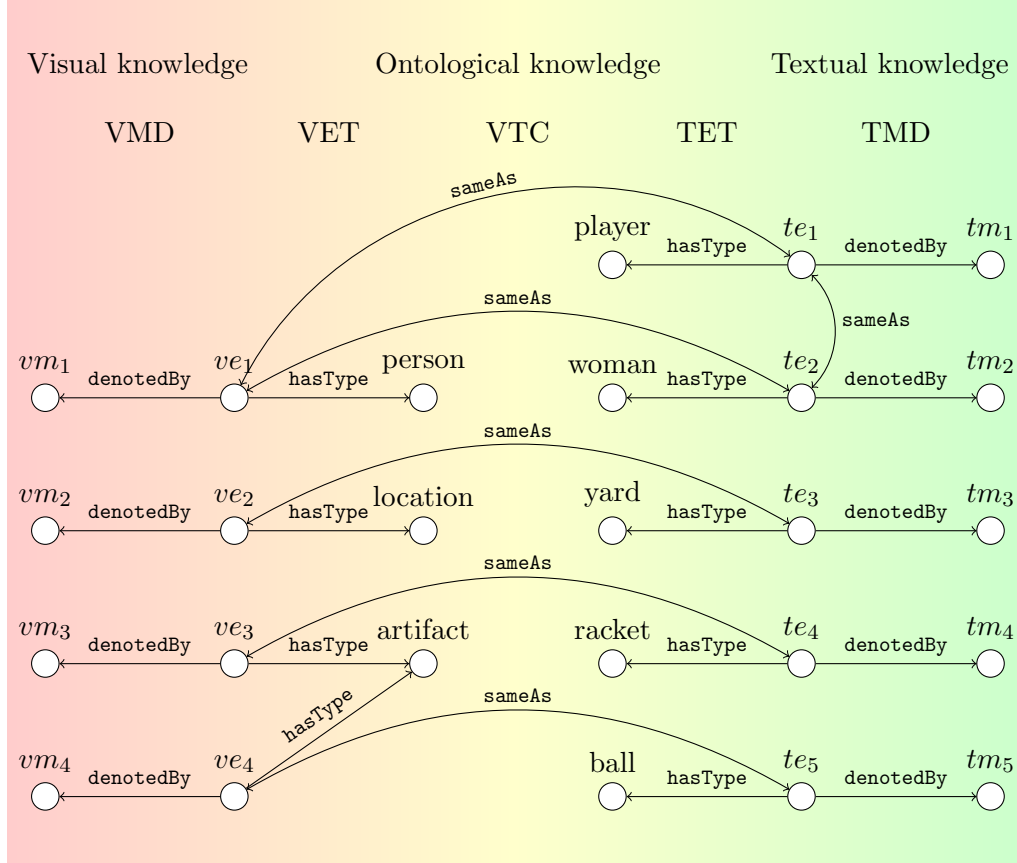


Figure 6.1: The (RDF) graph of the triples resulting from the VTKEL task. Each  $vm_i$  corresponds to the bounding boxes mentioning some entity (visual mentions); each  $ve_i$  represents an entity shown in some bounding box (visual entity); each  $tm_i$  corresponds to the portion of text mentioning some entity (textual mentions); finally, each  $te_i$  corresponds to some entity mentioned in the text. The other nodes of the knowledge-graphs are the concepts of the knowledge base, typing the entities.

The output of the example shown in Figure 2.1 is shown in Figure 6.1.

The VT-LINKER algorithm is composed of two sequential phases: The first phase, *the entity detection phase*, focuses on visual and textual entity detection and typing (VMD-VET and TMD-TET); the second phase, *the matching phase*, attempts to match the discovered entities (VTC). The entity detection phase is based on the output of state-of-the-art tools in NLP and CV. The matching phase is realized by a basic form of semantic matching that exploits the knowledge available in the T-box (i.e. class/sub-class hierarchy). In the following, we illustrate the different steps for each phase.

### 6.2.1 Visual Mention Detection (VMD)

To implement the VMD task, we used Keras-RetinaNet (KRN) <sup>[1]</sup>. We start from YOLO and Mask-RCNN (pre-trained object detectors models), which are trained on COCO dataset <sup>[34]</sup>. The accuracy of these object detectors was not promising during the experiments on the VMD task due to the limited classes (80) of COCO datasets. To improve the accuracy of the VMD task, we used *KRN* object detector,

<sup>1</sup><https://github.com/ZFTurbo/Keras-RetinaNet-for-Open-Images-Challenge-2018>

which is trained on the *Google-Open Images* [40] (GOI) dataset. The details of KRN and GOI dataset are described in chapter 4. The *KRN* has covered the majority of visual-categories of the VTKEL dataset.

We process images through the object detectors, which returns a set of bounding box proposals each of which is associated with a corresponding class and confidence score in  $[0,1]$ . For YOLO and Mask-RCNN, we used the model pre-trained on the 80 classes of COCO dataset [34], while for KRN we used the pre-trained model on 500 classes of GOI dataset. Among the bounding box candidates, we retain only those having confidence equal or greater than a specified threshold (in the experiments we set it to 0.5). In general, one could use some more sophisticated selection criteria that take into account also the co-occurrence with the other bounding box candidates (e.g., glass and bottle are more probable than glass and elephant) and the output of the textual mention detection and ontological knowledge. For the picture of Figure 2.1, the Object-detector returns three bounding box candidates with score higher than 0.5, labeled with *person*, *ball* and *racket*, but no bounding box has been found for the tennis court (due to the lack of appropriate classes for locations in the object-detector class set).

### 6.2.2 Visual Entity Typing (VET)

The objective of this sub-task is to find the correct most specific class in the knowledge base that can be associated to each visual entity associated to the visual mention detected in the VMD step. Notice that the COCO and GOI classes do not correspond one-to-one with the YAGO classes, this implies that we need to map the class returned by YOLO, Mask-RCNN and KRN into YAGO. A naïve way to implement this task is to map the label contained in the output of the object detector to its corresponding ontology class. Also, here more sophisticated methods can be implemented that take into account also the weight of the labels or additional visual/numerical features. In the VT-LINKER algorithm, we adopt the straightforward approach of manually mapping the 80 COCO and 500 GOI classes to the corresponding (most specific) classes of the YAGO ontology.<sup>2</sup> Examples of mappings from COCO to YAGO are: *person*  $\rightarrow$  *yago:Person100007846*, *ball*  $\rightarrow$  *yago:Ball102778669*, and *hotdog*  $\rightarrow$  *yago:Frank107676602*, etc.,.

### 6.2.3 Textual Mention Detection (TMD)

To detect textual mentions of entities we process the text with the PIKES suite, which provides services for both textual mention detection and textual entity typing to the YAGO ontology. These two tasks are tightly integrated in PIKES, however, for conceptual clarity, here we present them separately. Let us focus on entity mention detection. Given a text in input PIKES applies different state-of-the-art NLP techniques to discover entity mentions depending on their “nature”:

- *named entity mentions* (e.g., Barak Obama, Trento, IBM) refers to entities for which there is an individual in the knowledge base. They are recognized and linked (performing a task called Entity Linking) to the corresponding entity in YAGO (the knowledge base is not extended).

<sup>2</sup>The whole mapping can be downloaded from [https://figshare.com/articles/YOLO\\_to\\_YAGO\\_classes\\_mapping/8889848](https://figshare.com/articles/YOLO_to_YAGO_classes_mapping/8889848)



- *common nouns* (e.g., racket, ball, player, and woman) implicitly identify entities, by referring to their type (e.g., the mention of “racket” does not refer to the general notion of racket, but to a specific object, of type racket). Common nouns are discovered via word sense disambiguation (WSD). For every common noun, WSD returns the WordNet synset corresponding to the correct sense in which the noun is used. For instance, the correct sense of “racket” is the one indicating a sport equipment, and not a loud and disturbing noise. A new entity is created and added to the knowledge base for common nouns occurring in the text. Some further processing is performed to properly handle compound noun phrases (e.g., “a female tennis player”). PIKES also performs a syntactic analysis of the text: in particular, words in a noun phrase can be tagged either with *head* or with *modifier*, depending on their syntactic role in the noun phrase (e.g., in “a female tennis player” the noun “player” is the head and “female” and “tennis” are modifiers). In the current version of the VT-LINKER algorithm, a new entity is added to the knowledge base only for the head noun, and not for its modifiers.

For example, for the first sentence of the caption in Figure 2.1, PIKES detects three textual mentions: *woman*, *court* and *ball*.

#### 6.2.4 Textual Entity Typing (TET)

This task is also implemented using PIKES primitives. Typing for named entities is not necessary since these entities are in the YAGO knowledge base, and thus already typed according to the YAGO ontology. For the common nouns, we exploit the mapping from WordNet to YAGO also available in PIKES to obtain the (more specific) YAGO class associated to the WordNet synset of the mention, and the corresponding type assertion will be added to the knowledge base. For example, for the first sentence of the caption in Figure 2.1, PIKES types the entities corresponding to the textual mentions *woman*, *court* and *ball*, with the YAGO classes `yago:Woman110787470`, `yago:Court108329453`, and `yago:Ball1102778669`, respectively.

#### 6.2.5 Visual Textual Coreference (VTC)

This is the last sub-task that has to be accomplished by VT-LINKER. For this task, we exploit the class/sub-class hierarchy between the classes in the knowledge base. Let  $TE$  and  $VE$  be the set of textual and visual entities that are mentioned in a visual-textual document, and that are present in the knowledge base with a given type. The coreference sub-task has the objective of finding the coreference relation  $CR \subseteq VE \times TE$  such that the following consistent properties hold:

1. For every  $ve \in VE$  there is at least one  $\langle ve, te \rangle \in CR$ ;
2. For every  $ve \in VE$  there is at most one  $\langle ve, te \rangle \in CR$ ;
3. If  $\langle ce, ve \rangle$  ( $ce$  is the coreference entity) and  $ve$  and  $te$  are of type  $C_v$  and  $C_t$  respectively then either  $C_v \sqsubseteq C_t$  or  $C_t \sqsubseteq C_e$  holds in the knowledge base.

In simple situations, the above criteria uniquely define the coreference relations. In our case, the coreference chain can be defined as the total number of mapping pairs (clauses) between one or more than one visual entity (bounding box) with

the corresponding textual entity mentions in five captions. The entity mentions of  $VE$  and  $TE$  can be considered predicate formulas, and  $CR$  will be true if at least one pair of coreference-chain become true. This is the case for instance of the example presented in Fig 2.1. However, in many cases the relation  $CR \subseteq VE \times TE$  is not uniquely defined by the above criteria. Nevertheless, the problem can be straightforwardly encoded as a *MaxSat* problem<sup>3</sup>. In case of CRs with equal total weight, a random choice is taken although additional heuristics could be implemented either by using some supervised learning method or by handcrafting the weight of a pair  $\langle ve, te \rangle$  by exploiting some additional features of the mentions of  $ve$  and  $te$ .

## 6.3 Experimental Evaluations

To evaluate the performance of VT-LINKER, we used two ground truth datasets. The first consists of more than 31k documents called VTKEL30 and has been derived from Flickr30k-Entites. This dataset is generated automatically by typing the visual and textual entities with classes from the YAGO ontology. The second dataset called VTKEL1k\*, has been obtained by randomly selecting 1000 pictures (and the corresponding captions) from the VTKEL30k dataset, and manually validating and revising the proposed alignments to YAGO. In the following, we provide some details on the datasets, and then we describe the evaluations conducted.

### 6.3.1 Datasets

The VTKEL30k<sup>4</sup> dataset has been obtained by extending the Flickr30k-Entities dataset by linking textual and visual mentions to entities assigned with the most specific YAGO class. Looking at Figure 2.1, we started from the left part of the figure (the picture and captions, with annotated visual and textual mentions, and alignment between corresponding mentions), available in Flickr30k-Entities, and we extended it with the right part, by populating a knowledge base with corresponding entities typed according to the YAGO ontology. The VTKEL30k dataset has been automatically produced by processing the captions of Flickr30k-Entities with PIKES for entity recognition and linking to YAGO. Specifically, for each textual mention (aligned to a visual mention) in Flickr30k-Entities, detected also by PIKES, a corresponding entity is created (or aligned to, if already existing) and typed according to the appropriate YAGO ontology.

The VTKEL1k\*<sup>5</sup> has been obtained by randomly sampling 1000 entries from the VTKEL dataset (corresponding to 20,356 textual mentions, and 8,673 visual mentions). Every entry of VTKEL\* has been manually checked for the correctness and completeness of the YAGO classes associated to the mentioned entities. Wrong and missing links are manually adjusted. Errors are mainly due to the incorrect word sense disambiguation: e.g., in some cases “bus” was linked to the concept of computer bus, instead of that of coach, and “arm” to weapon instead of bodypart. The construction of VTKEL1k\* dataset allows us also to estimate the error rate

<sup>3</sup>MaxSat (maximum satisfiability problem), is the problem of finding the maximum number of clauses of a given Boolean formula in the conjunctive normal form (CNF), and this can be made true by assigning the true values of the variables of the formula.

<sup>4</sup>[https://figshare.com/articles/VTKL\\_dataset\\_file/7882781](https://figshare.com/articles/VTKL_dataset_file/7882781)

<sup>5</sup>The VTKEL\* dataset can be downloaded from [https://figshare.com/articles/VTKEL\\_dataset/8896487](https://figshare.com/articles/VTKEL_dataset/8896487)

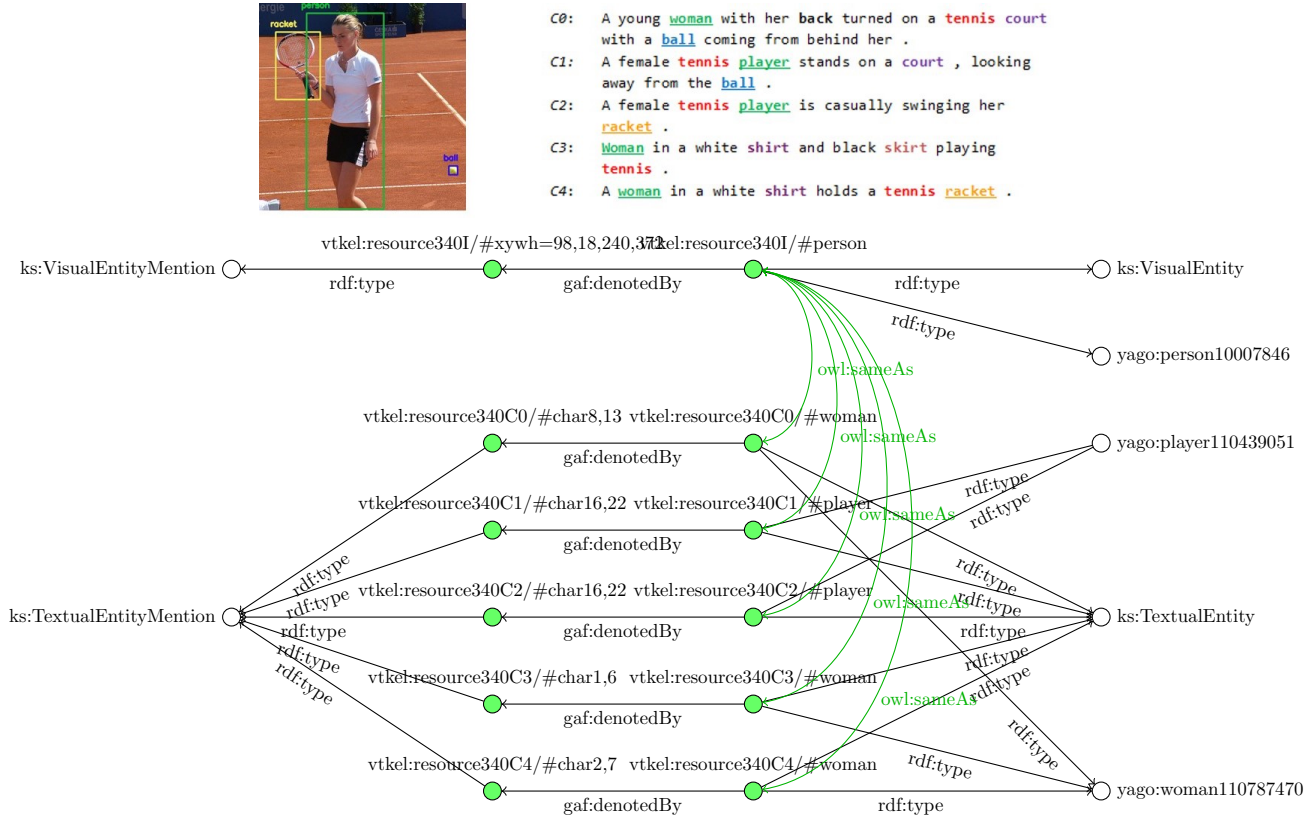


Figure 6.2: An example of how data are stored in the VTKEL dataset. The RDF graph shows how the visual and textual mentions of the woman are mapped to a visual entity and 5 textual entities, all linked together by the `owl:sameAs` relation. The entities are linked to the most specific YAGO classes in this case person, woman, and player.

of the larger dataset (VTKEL30k). In particular, we found no missing link (i.e., recall is 100%) and 916 incorrectly linked mentions, which amounts to an accuracy of 95%. We believe that the VTKEL30k dataset can be reasonably considered a ground truth.

To maximize reusability and connection with the Semantic Web, we represent the datasets in RDF. This representation will also support semantic visual query answering via standard SPARQL language [7]. To organize the dataset, we adopt the model proposed in [14], extending it for representing visual mentions. The model is organized in three distinct yet interlinked representation layers: *Resource*, *Mention*, and *Entity* layer.

An example of how the information is stored in the dataset is provided in Figure 6.2.

### 6.3.2 Evaluation

We evaluated the performances of VT-LINKER on both VTKEL1k\* and VTKEL30k datasets. We separately assessed the performance on the three sub-tasks described in Section 6.2. We use the standard metrics, namely precision ( $P$ ), recall ( $R$ ), and F-score ( $F_1$ ). The figures obtained from the evaluation are reported in Table 6.1.

Table 6.1: VT-LINKER and ViTKAN evaluation results on VTKEL1k\* and VTKEL30k using KRN object detector.

task	VTKEL1k* dataset			VTKEL30k dataset		
	$P$	$R$	$F_1$	$P$	$R$	$F_1$
VMD + VET	0.786	0.867	0.825	0.769	0.873	0.818
TMD + TET	0.954	0.884	0.918	0.953	0.898	0.925
VTC (VT-LINKER)	<b>0.314</b>	<b>0.381</b>	<b>0.344</b>	<b>0.303</b>	<b>0.403</b>	<b>0.346</b>
VTC (ViTKAN)	<b>0.716</b>	<b>0.771</b>	<b>0.742</b>	<b>0.701</b>	<b>0.782</b>	<b>0.739</b>

### Visual entities detection and typing (VMD)+(VET)

To evaluate the visual detection part, we use standard method adopted for evaluating object detection. A visual mention  $b_p$  of type  $t_p$  produced by VT-LINKER on an image is considered to be correct if the ground truth annotation of the image contains a bounding box  $b_g$  of type  $t_g$  such that the intersection over union ratio ( $\frac{area(b_p \cap b_g)}{area(b_p) \cup area(b_g)}$ ) is greater or equal to  $\frac{1}{2}$  and if the predicted type  $t_p$  is equal or a sub-class of  $t_g$  in YAGO.

For the 1000 documents VTKEL1k\* dataset, VT-LINKER predicted 13,365 visual objects (visual-entities) by using *KRN* object detectors, with respect to the 10,074 ground-truth bounding-boxes. The VT-LINKER correctly predicted 8633 of them. By using the same procedure for 30K+ documents VTKEL30k dataset, VT-LINKER predicted 355,464 visual-entities with respect to the 275,770 ground-truth annotated visual entity (i.e. bounding boxes). The VT-LINKER correctly predicted 232,474 of them. The results from VMD and VMT (i.e. VMD+VMT) are listed in Table [6.1](#) in details.

In the majority of cases, the VT-LINKER framework by using YOLO and Mask-RCNN ignored *human – bodyparts* and *clothing* during the prediction of visual mentions. To improve the quality of VT-LINKER, we used *KRN* object detector. The *KRN* object detector not only predicted *human – bodyparts* and *clothing*, but also improve the results of other classes, missed by YOLO and Mask-RCNN.

### Textual entities detection and typing (TMD)+(TET)

To evaluate the performance of this sub-task, we apply a criterion analogous to the visual entity detection and typing sub-task. A textual mention  $w_p$  of an entity of YAGO class  $t_p$  predicted by VT-LINKER on a caption, is considered to be correct if the ground truth annotation on the caption contains a mention  $w_g$  of an entity of type  $t_g$  such that  $w_p$  is equal or a sub-string of  $w_g$  and the type  $t_p$  is equal or a sub-type of  $t_g$  according to the YAGO class hierarchy. From the 5000 captions of VTKEL1k\* dataset, VT-LINKER wrongly recognized and linked 935 out of total 20,374 textual entities, which amount to  $Precision = 0.954$  and  $Recall = 0.884$ . Similarly, for 158,605 captions of VTKEL30k dataset, VT-LINKER correctly recognized and linked 576,769 out of total 612,281 textual entities. Most of the errors during entity recognition and linking are due to the *word sense disambiguation*. The details evaluations of TMD+TET is listed in Table [6.1](#).

**Visual textual coreference (VTC)**

We evaluate the capability of VT-LINKER on aligning visual and textual entities. A coreference (alignment) pair  $\langle ve_p, te_p \rangle$  produced by VT-LINKER is correct, if the ground truth contains the triple  $ve_g \text{ owl:sameAs } te_v$  such that the visual mentions (bounding boxes) of  $ve_p$  and  $ve_g$  matches (under the IOU ratio) and the textual mention of  $te_p$  matches the textual mention of  $te_g$  (i.e.,  $te_p$  is equal or a substring of  $te_g$ ). To be noticed that, we are not considering the types of the entities this time. Type compatibility is indeed guaranteed by the fact that coreference pairs are added only if their types are compatible (i.e., they are either equal or in sub-class relation in YAGO). From the 1000 entities VTKEL1k\* dataset, the VT-LINKER algorithm predicted 27,247 pairs (i.e. visual entities mapped with textual entities) in total. From predictions, 8289 pairs were correctly aligned with respect to 21,732 ground-truth pairs ( $P = 0.314, R = 0.381$ ). Similarly, for VTKEL30k dataset, the VT-LINKER correctly aligned 266,312, from 878,766 predicted pairs with respect to 663,457 ground-truth ( $P = 0.303, R = 0.403$ ). In some pairs from the categories of *human-body parts* and *scenes* (e.g. *building, playgrounds, etc.*), the VT-LINKER made wrong predictions due to the complexity and challenges facing in these categories.



## Chapter 7

# The ViTKan Algorithm

In the previous chapter, we described our baseline VT-LINKER algorithm in detail. In this chapter, we are describing a supervised algorithm called ViTKAN (*Visual-Textual-Knowledge Alignment Network*), their development, architecture, training, and evaluations in detail.

### 7.1 Introduction

We learn in a better way if an image is represented with its description, the reason is the association of textual-words with the image-regions. For instance, it becomes easy for us to acquire facts from the newspaper (which described an event with both text and image), making a diagnosis from the MRI scans with the reports, watching a post on social media (consists of images and text), or enjoying a movie with subtitles. This problem of associations between textual-entities and image-regions is called noun, or entity, or phrase grounding (localization). An additional step of linking the background knowledge of those visual and textual entities by exploiting a knowledge-base (Ontology) can extract huge structural facts about them. The text-region association and their structured background knowledge of both modalities can be utilized in high-level tasks, such as *visual question answering* [43, 86, 87], *image captioning* [30, 88, 89], and *images from text, or text from images retrieval* [90, 91, 92].

Existing natural language models can be used to provide *textual entity recognition* and *linking* tasks [93, 10] from the noun-phrases of image captions. In parallel, an object detector [77, 94, 78] can detect and represent the object regions in an image. However, learning the mapping between these two independent modalities is a challenging problem, which requires first to parse language queries and then relating knowledge of these queries to ground (i.e. localized) objects in the visual domain. To address the problem of mapping, the current state-of-the-art methods [95, 96, 97, 9, 98] rely on a proposal generation system to produce a set of bounding boxes as grounding candidates. In this approach, there are two main challenges (1) how to learn the correlation between language (query) and visual (proposals) modalities, (2) and how to localize the objects shown in the image and described in the text (i.e. multimodal associations). State-of-the-art methods have solved the first problem by learning a subspace to measure the similarities between proposals and queries. After learning the subspace they treat the second problem as a retrieval task, where proposals are ranked with respect to their input query.

These approaches based on the information of visual and textual modalities, by using features data of annotated images and text from Flickr30k-Entity dataset

[5, 9] during training. A popular baseline for image-text embedding is the *Canonical Correlation Analysis* (CCA), which finds linear projections that maximize the correlation between projected vectors from the two image-regions and text domains [64]. In some approaches, they used also the structured information of language in the form of language-scene graphs [72]. However, none of these approaches have used *background knowledge* of visual and textual entities in the form of structured knowledge coming from a knowledge-base.

In this chapter, we introduced a supervised algorithm called ViTKAN. During training, the ViTKAN takes in input an image, natural text, and the Ontological knowledge of visual entities described in the image, and textual entities mention in captions. The ViTKAN algorithm solved the problem of mapping (i.e. alignment, association, grounding) between visual entities (in  $d_i$ ) and textual entities (in  $d_t$ ) efficiently. We trained the ViTKAN on VTKEL1k\* dataset [17]. We performed the evaluations to check the quality of the ViTKAN algorithm on VTKEL1k\* and VTKEL30k datasets, and also compared the results with state-of-the-art methods. Figure [7.1], shows the architecture of ViTKAN in details.

## 7.2 Related Work

Given an image and textual description of the image, the problem of *phrase grounding* tries to localize visual objects in the image with the corresponding phrases described in the captions. The main challenge in the phrase grounding problem is the correlation between visual and textual modalities. *Karpathy et al.* [30] align noun-phrases and image regions, using (i) *convolutional neural network* (CNN) over images, (ii) *bidirectional recurrent neural network* (RNN) over sentences, and (iii) a structured objective that aligns the two modalities. One of the popular baselines for image-text embedding is *Canonical Correlation Analysis* (CCA), which finds linear projections that maximize the correlation between projected vectors from the two image-regions and text domains described in [64]. *Wang et al.* [65] employ structured matching of phrases and regions which develop the semantic relations between phrases to agree with the visual relations between image-regions. They formulate structured matching as a discrete optimization problem into a linear program and use neural networks for embedding visual regions and phrases into vectors.

*Plummer et al.* [5] augment the CCA model to leverage extensive linguistic cues in the phrases. *Rohrbach et al.* [66] propose grounding by reconstruction, an approach using an attention mechanism for phrase grounding by ranking proposal in an unsupervised scenario. During training their approach encodes the phrase using a recurrent network language model and then learns to attend for the relevant image region in order to reconstruct the input phrase. *Hu et al.* [67] propose a *Spatial Context Recurrent ConvNet* model which based on a 2-layers LSTM to rank visual proposals using embedded query and visual features. *Dogan et al.* [68] proposed a sequential and contextual process, which encode region proposals and all phrase into two stacks of LSTM cells, along with so-far grounded phrase-region pairs. These LSTM stacks collectively capture context for grounding of the next phrase. The resulting architecture supports many-to-many matching by allowing an image region to be matched to multiple phrases and vice versa. ViLBERT (Vision-and-Language BERT) [69] learn representation jointly from both visual and textual domains using two-stream co-attentional transformer layers independently. In contrast to ViLBERT, VisualBERT [70] consists of a stack of transformer layers, which indirectly align elements of an input text and regions in an associated input image



with self-attention. The VisualBERT further demonstrates elements of language to image in syntactic relationships, for example, associations between verbs and image regions corresponding to their arguments.

Yang *et al.* [71] propose a linguistic structured guided propagation network for one-stage phrase grounding. In their model, they explore the linguistic structure of the sentence and perform relational propagation among noun-phrases under the guidance of the linguistics relation between them. Specifically, they first constructed a linguistic graph parsed from the sentence and then capture visual and textual (multimodal) feature maps for all noun-phrases nodes independently. Jing. *et al.* [72] formulate the problem of phrase grounding as a graph matching problem to find the nodes of visual and textual entities and to represent them in structured layouts of the image and sentence respectively. In their approach, they build a cross-modal graph convolutional network to learn cohesive node representations, which distinguish both node information and structured information to reduce the inconsistency of visual and textual graphs. Yu *et al.* [73] propose a *Cross-Model Omni Interaction network* (COI-Net) composed of (i) a neighboring interaction module, (ii) a global interaction module, (iii) a cross-modal interaction module, and (iv) a multilevel alignment module. They formulate the complex spatial and semantic relationship between image regions and phrases using these multi-level multi-modal interactions. To further enhance the interaction between two modalities, they use a co-attention module with the cross-modal context for all image regions and phrases.

These existing methods lack the ability to model the background knowledge of visual and textual modalities coming from the knowledge bases (Ontologies) by linking the visual and textual entities mentions. From the above literature, it becomes clear that there is not a single comprehensive approach, which used visual and textual modalities with the background knowledge for the task of phrase (noun, entity) grounding.

### 7.3 The ViTKan Algorithm

In this section, we describe in detail the VITKAN algorithm, which based on visual, textual, and Ontological pipelines, that linked Vision, Language, and Knowledge-Representation modalities. We take the advantages of using background knowledge of visual and textual pipelines coming from the knowledge-base (YAGO) and handling it explicitly. Later, we used the subsymbolic (embedding) approaches for extracting features from these pipelines. The visual part of VITKAN algorithm takes in input an image and passed through an object detector for bounding-boxes prediction and later passed these bounding boxes through VGG16 [16] classifier for visual features (details in [4]). In the textual part, we passed the natural language text ( $d_t$ ) of image through an *entity detection* and *linking* tool called PIKES [10, 82] and later extract the embedding of textual entity mentions using Word2Vec [8]. After visual and textual features extraction, the next step is to use the Ontological knowledge of visual and textual entity mentions in the form of features vectors (encoders) exploiting the sub and super-class hierarchy (i.e. taxonomy) of YAGO [13] a well-known web-semantic knowledge-base. In section [7.3.1] we explained the visual module, section [7.3.2] described the language module, section [7.3.3] described the architecture of VITKAN, and section [7.4] described in details the evaluations of VITKAN algorithm using VTKEL1k\* and VTKEL30k datasets.

### 7.3.1 Visual Module

In the ViTKAN algorithm, the most significant part is the detection of visual entity mentions (objects) shown in the images. We used a pre-trained object detector framework called *Keras-RetinaNet* (KRN) [\[4\]](#), which utilize the ResNet-101 [\[99\]](#) (i.e. Residual Network with 101 layers deep) as a backbone network. The KRN object detector is trained on the *Google Open-Images* (GOI) [\[40\]](#) dataset. The GOI dataset object categories do not correspond one-to-one with the YAGO classes, which entails to map the class returned by KRN into the YAGO type. A naïve way to implement this task is to map the label contained in the output of the object detector to their corresponding Ontology class. Also, here more sophisticated methods can be implemented that can take into account also the weight of the labels or additional visual/numerical features. We adopted the straightforward approach of manually mapping the 500 GOI classes to the corresponding (most specific) classes of the YAGO ontology.

We process every image through KRN object detector for, (i) bounding boxes labels, (ii) bounding boxes values in the form of  $\{x, y, x + w, y + h\}$ , (iii) YAGO type (class) of labels, and (iv) class probability vectors from KRN. The detected bounding boxes are processed through a convolution-neural-network classifier called *VGG16*, trained on ImageNet [\[81\]](#) dataset to extract visual features. For every bounding box, we extracted their visual-features in a vector of size 4600, which consist of (i) bounding-box values (4), (ii) visual-features from the last max-pooling layer of VGG16 (4096), (iii) and classes probability vector from KRN (500) for training. During the training, we consider only those bounding boxes predicted by KRN, which has *intersection-over-union*  $\geq 0.5$  with ground-truth bounding boxes. We used VTKEL1k\* dataset [\[17\]](#) during training of the ViTKAN algorithm.

### 7.3.2 Language Module

After visual, the next important module of the ViTKAN algorithm is the extraction of textual features data from image captions, which can be used for the training phase. We used a textual knowledge extraction tool called PIKES [\[10\]](#), which produce the knowledge graphs from the textual resource. We process captions through PIKES, which recognized textual entity mentions and linked them to YAGO knowledge-base. Linking these entity mentions to YAGO ontology can be used to extract huge structured background knowledge in the form of *entity gloss* (i.e. description of entity), *taxonomy* (sub and super-classes of entity), *Wikipedia pages*, and other information in the form of RDF graphs using Linked-Open-Data [\[100\]](#). In ViTKAN, we exploit the taxonomy information of each entity mention.

For every textual entity mentions, we process their sub and super-classes hierarchy using the taxonomy of YAGO Ontology and find the mapping with the corresponding class(es) of visual-entity mention. We represent this mapping into two one-hot encoders called (i) sub-class encoder, and (ii) super-class encoder. The size of each encoder is 500, because GOI dataset has 500 total classes.

For example, if we have an image and a natural language text describing the image, i.e. “*A man with his Rottweiler*”. After passing the image through the KRN object detector, we received two objects (i.e. “*person*” and “*dog*”). In parallel, PIKES recognized two entity mentions *man* and *rottweiler*, and linked them with two URIs of the knowledge-base <http://dbpedia.org/class/yago/Man110287213>

<sup>1</sup><https://github.com/ZFTurbo/Keras-RetinaNet-for-Open-Images-Challenge-2018>

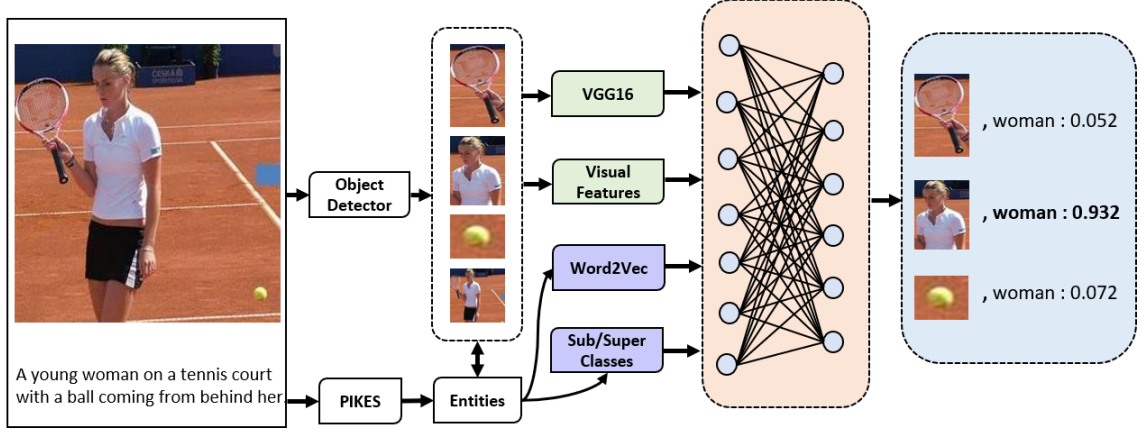


Figure 7.1: The ViTKAN algorithm architecture pipeline.

and <http://dbpedia.org/page/Rottweiler> respectively. We process the sub and super-class hierarchy with respect to YAGO taxonomy and found that (i) *person* is the super-class of *man*, and (ii) *rottweiler* is the sub-class of *dog*. Later, we process the recognized textual-entity mentions through *Word2Vec* and extract the textual features in the form of (300 values) word-embedding vector. For each textual-entity mention, we extract 1300 features values. We stacked the visual and textual entities mentions features into a vector of size  $4600 + 1300$ , to be used as a input during training of the ViTKAN algorithm.

### 7.3.3 Neural-Network module and training of ViTKan

In the previous sections, we described the techniques for extracting visual and textual features data. After features data, the next step is to develop a *neural-network* part of the ViTKAN algorithm and utilized these features data for training. We developed a neural-network architecture, which consists of an input layer of size  $5900 \times 2048$ , two fully-connected (FC1) layers of size  $2048 \times 512$  and (FC2)  $512 \times 1$  respectively. From input-layer to FC1 has *ReLU*, FC1 to FC2 layers has *LeakyReLU*, and FC2 to output-layer has *sigmoid* activation functions respectively. We used *Adam* [101] (adaptive moment estimation) optimization function for updating the network weights on the training data. The most important part in ViTKAN module is to solve the problem of mapping (i.e. association, or alignment) between visual entities shown in the image with textual entity mentions described in captions. We used the *Binary Cross-Entropy* (i.e. sigmoid) loss function at the output layer to predict the alignment (True) and penalized non-alignment (False) pairs between visual and textual entities. During training, we used the learning rate value:  $\alpha = 0.001$ . The ViTKAN algorithm is trained for 10 epochs with a batch size of 100. Figure 7.1 shows the components and architecture of the ViTKAN algorithm in details.

We trained the neural-network model using VTKEL1k\* dataset. The VTKEL1k\* dataset consists of images (1000), captions (5000), annotations for bounding-boxes (8673), and textual mentions (20,356). Moreover, each bounding box is annotated and mapped (associations) with one or more than one textual mention and each textual-mention is linked with YAGO Ontology.

We used features data of the VTKEL1k\* dataset during training and testing of the NN by applying *10-fold cross-validation*. The details of this dataset are described in chapter 6. Each document  $d$  of the VTKEL1k\* consist of image  $d_i$

and five captions  $d_t$ . We process  $d_i$  through KRN and predict set of visual-objects  $O_N = \{o_1, o_2, \dots, o_n\}$ . In parallel  $d_t$  is passed through PIKES for textual-entity recognition and linking tasks i.e.  $T_{CN}\{t_{11}, t_{12}, \dots, t_{1n}\}$ , where  $C$  denote caption-number, and  $N$  is used for the total number of textual entity mentions in  $C$ . For each  $d$ , we extract features data of  $O_N$  and  $T_{CN}$  as described in the previous sections. During the processing of VTKEL1k\* features data, we found that the mappings of visual and textual entity mentions are imbalanced in the visual classes. From the imbalance, we mean that a massive portion of the VTKEL1k\* dataset belongs to the *people* class. To solve this problem, we added 66.66% of false mapping to the imbalance portion, i.e., for every true pair, we added two false pairs. We trained end-to-end the NN, utilizing the balanced data.

## 7.4 Experimental Evaluations

To check the quality of ViTKAN algorithm, we used VTKEL1k\* and VTKEL30k datasets for evaluation experiments. The evaluation results of VMD+VET and TMD+TET (described in chapter: [6](#)) tasks are the same due to the KRN and PIKES tools used in both VT-LINKER and ViTKAN. We check the performance of ViTKAN primarily on two tasks called *Visual Textual Coreference* (VTC), and *Textual Entity Grounding* (TEG). VTC is a major task in the VT-LINKER algorithm, described in details in chapter [6](#), while the TEG (or phrase grounding) task is jointly addressed by the community of CV and NLP for *textual-entity localization* problem [5](#), [97](#), [9](#), [102](#). In the following, we describe these evaluations in detail.

### 7.4.1 Visual Textual Coreference

The ViTKAN algorithm solves the problem of VTKEL with better performance than the baseline (VT-LINKER) algorithm in the VTC task. Each pair ( $p_i$ ) predicted by the ViTKAN consists of a visual-object ( $b_p$ ) and a textual-entity mention ( $t_p$ ). We checked  $b_p$  against the ground-truth bounding-box ( $b_g$ ) and  $t_p$  against the ground-truth textual mention ( $t_g$ ): a pair is considered to be correct if the *intersection-over-union* (IOU) ratio of  $b_p$  over  $b_g$  (i.e.  $\frac{area(b_p \cap b_g)}{area(b_p) \cup area(b_g)}$ ) is greater or equal than 0.5, and  $t_p$  is equal or a substring of  $t_g$ . The improved results of ViTKAN on VTC task using VTKEL1k\* and VTKEL30k datasets are presented in Table [6.1](#).

In the visual module, we exploit the KRN object detector, which plays a major role in predicting visual objects from images. The KRN performed very well on *people, clothing, human-body-parts, animals, vehicles, and instrument* categories, while missing most of the objects from visual-scenes categories (e.g. buildings, graces, playing grounds, etc..). In the community of CV, correctly detecting and typing objects from the *scenes* category is still a challenging problem, and any improvement in this respect would potentially improve the ViTKAN performance. In some images, which consist of very few and small objects (even a human can see the objects with great effort), the KRN object detector can predict these objects. The visual objects in Figure [7.2](#) pictures are mostly missed or wrongly predicted by state-of-the-art object detectors (e.g. *YOLO and SSD*), while KRN predicts correctly these objects.

Table 7.1: Results of state-of-the-art using Flickr30k-Entity, and ViTKAN algorithm using VTKEL1k\* and VTKEL30k datasets for TEG task.

Methods	Dataset	R@1	R@5	R@10	R@50	R@100
BRNN [30]	Flickr30k-Entity [5]	22.24	48.22	61.40	-	-
NMLM [103]		23.12	50.70	62.81	-	-
m-RNN [104]		35.42	63.80	73.72	-	-
GMM+FV [105]		33.30	62.12	74.73	-	-
KAC+KBP [106]		38.71	-	-	-	-
Yeh et al. [107]		36.93	-	-	-	-
MATN [108]		33.10	-	-	-	-
CCA [5]		41.77	64.52	70.77	-	80.30
StructMatch [65]		42.08	-	-	-	-
DSPE [109]		43.89	64.46	69.66	-	-
GroundR [97]		47.81	-	-	-	-
Embedding-N [110]		50.67	70.21	75.51	-	-
Similarity-N [110]		51.05	70.30	75.04	-	-
$OT_T$ [102]		35.98	70.33	78.97	-	-
$OT_S$ [102]		41.12	70.42	77.48	-	-
<b>ViTKan</b> (ours)		VTKEL1k*	<b>46.81</b>	<b>70.46</b>	<b>77.01</b>	<b>84.02</b>
<b>ViTKan</b> (ours)	VTKEL30k	<b>25.69</b>	<b>51.03</b>	<b>62.48</b>	<b>80.68</b>	<b>83.82</b>

### 7.4.2 Textual Entity Grounding

The problem of TEG tries to localize the visual objects in an image with the corresponding textual entity described in captions. State-of-the-art [5, 97, 9, 102] approaches are using *recall* on TEG<sup>2</sup> task by utilizing *Flickr30k-Entity dataset* [5]. We follow the same evaluation procedure with respect to state-of-the-art methods by predicting the top 100 bounding boxes (proposals) of images using KRN object detector. During inference, the ViTKAN algorithm uses features data of visual and textual entity mentions (described in sections: 7.3.1, 7.3.2).

We consider the mapping between textual-portion and image-region as the *grounding problem*, which entails *textual entity mention* as a *query* and the 100 bounding box proposals from input image as the database to search over for the alignments. We followed the evaluation settings of state-of-the-art approaches and measure *Recall@K*, where  $\{K=1, 5, 10, 50, \text{ and } 100\}$ . The prediction performance of the ViTKAN algorithm on every pair consists of a textual entity mention and 100 bounding box proposals. We ranked the proposal (i.e. classifier probabilities in descending order) and report *Recall@100*, which gives us the upper bound on grounding performance. The results for *Recall@K* obtained by evaluating ViTKAN on the TEG task using VTKEL1k\* and VTKEL30k datasets are presented in Table 7.1, together with a reported performance of state-of-the-art approaches.

On the TEG task, the performance on VTKEL30k is substantially lower than on VTKEL1k\*, especially for low  $K$  values in *Recall@K*. For training ViTKAN we used the gold standard VTKEL1k\* dataset, which consists of 1000 documents, randomly selected from the VTKEL30k dataset. Given the substantially smaller

<sup>2</sup>Note: The state-of-the-arts solve the problem as *noun-phrase grounding*. In our case, the textual-entity mentions recognized by PIKES are considered to be the noun-phrases of the corresponding caption(s).

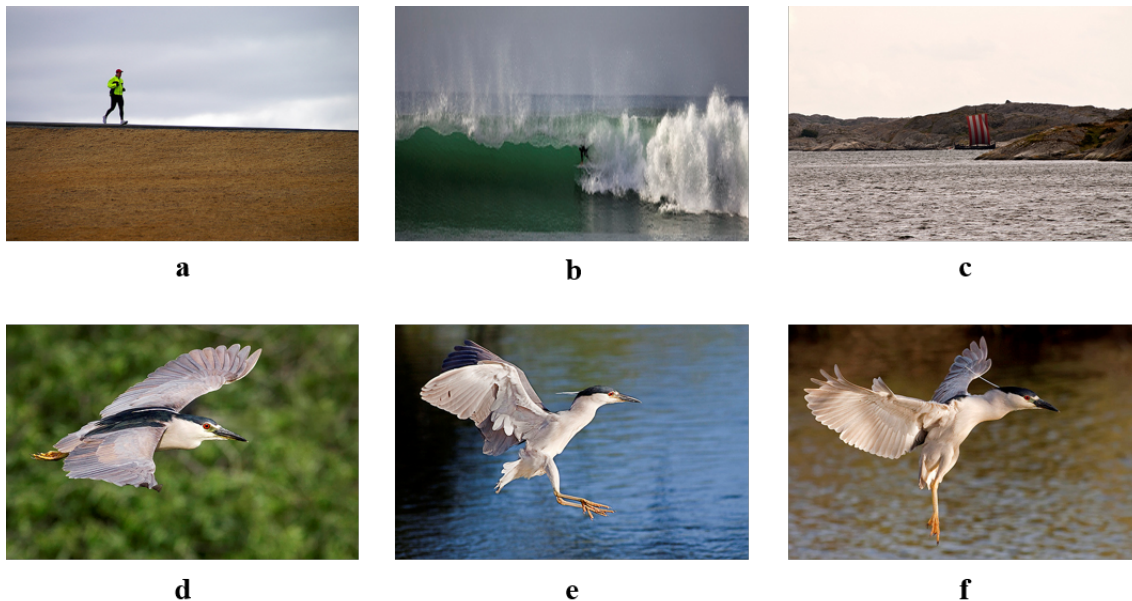


Figure 7.2: Some images of VTKEL dataset that consist of challenging objects.

size of VTKEL1k\* compared to VTKEL30k, it is likely that the feature data from the entity mentions in VTKEL1k\* do not sufficiently cover the variability of the entity mentions on VTKEL30k. A solution to achieve higher recall scores also on VTKEL30k would be to use a larger subsample from VTKEL30k, possibly cross-validating the performance on the whole VTKEL30k dataset. However, training the approach on such a large dataset is quite demanding computationally, and would require highly-spec'd computing hardware that was not available at the time of training VITKAN.

## 7.5 Conclusions

In this chapter, we introduced a supervised algorithm called VITKAN, their development, architecture, training, and evaluations in detail. The VITKAN algorithm solves the problem of visual-textual entity matching, based on a supervised method by using a more complete set of visual classes (from the GOI dataset), a sophisticated object detector (KRN), and a knowledge extracting tool from the textual resource (PIKES). The visual part of VITKAN algorithm takes in input an image and passed through KRN object detector for bounding-boxes prediction and process bounding boxes through a classifier for visual features extraction. In the textual part, we passed the natural language text (captions) of the image through an *entity recognition* and *linking* tool for extracting the embedding and background knowledge of textual entity mentions. We used the Ontological knowledge of visual and textual entity mentions in the form of features vectors (encoders) exploiting the sub and super-class hierarchy (i.e. taxonomy) of YAGO [13] knowledge-base. We also described the VTKEL1k\* dataset, used during the training, and performed evaluations to check the quality of VITKAN on VTKEL1k\* and VTKEL30k datasets. The VITKAN algorithm, efficiently solves the visual-textual alignment task of VT-LINKER algorithm. We also performed evaluation experiments on the textual-entity grounding task. The novelty of VITKAN algorithm is the exploitation of structured background knowledge coming from the knowledge-base YAGO, which was ignored

by the traditional methods.

In the future, we are interested to improve the quality of VITKAN algorithm, by training it on the VTKEL30k dataset. We also want to apply the method to a dataset that includes more pictures and textual resources other than captions (e.g., short news with pictures).





## Chapter 8

# Conclusion

In this thesis, we understand the contents of a document which composed of both *image* and *text*. The image part of the document consists of visual objects and the textual part described these objects in natural language text. To process and interpret these contents, we need to develop an *artificial agent* that involved cross-modal learning from image and text data and predict objects (i.e. *visual entity mentions*) shown in the image and parallelly recognized the *textual entity mentions* described in the text. After recognizing the visual and textual entity mentions, the agent will link them to its background knowledge by using the *knowledge-base* (Ontology). We called this problem *Visual-Textual-Knowledge-Entity Linking* (VTKEL). After defining the *VTKEL* problem, the next challenging part was to use the state-of-the-art tools and techniques in Computer vision, NLP, and Knowledge representation to solve the problem of VTKEL. We describe in detail these tools and techniques in chapter 4. We developed two datasets called VTKEL1k\* and VTKEL30k, which check the quality of algorithms solving the VTKEL problem. These datasets can also be used in the training and evaluations of algorithms. The innovation of VTKEL problem and datasets are the integration and intersection of vision, language, and knowledge modalities. In chapter 5, we described in detail the VTKEL1k\* and VTKEL30k datasets. The accuracy of VTKEL1k\* amounts to 95%, and we believe that an error rate of 5% also exists in the manually developed datasets.

In chapter 6, we described a *baseline algorithm* called, which solve the problem of VTKEL. In this chapter, we described the development of VT-LINKER, by exploiting state-of-the-art tools and techniques. Given a document ( $d$ ) composed of text ( $d_t$ ) and image ( $d_i$ ), the VT-LINKER applies an object detector to  $d_i$  part, resulting in a set of bounding boxes labeled with classes of the YAGO Ontology 1. In parallel, VT-LINKER processes  $d_t$  part with a tool for entity recognition, which labels the noun phrases with classes of the YAGO Ontology. Finally, the VT-LINKER attempts to link visual and textual mentions which correspond to the same entity. This final task is done by exploiting Ontological knowledge about *class/subclass hierarchy* (taxonomy), and similarity information available in the textual mentions. We performed the evaluation experiments on VTKEL1k\* and VTKEL30k datasets, to check the quality of VT-LINKER. The highest F1 scores of the VTC task using VTKEL1k\*, and VTKEL30k datasets are 0.303 and 0.346 respectively.

In chapter 7, we described a supervised algorithm called VITKAN (Visual-Textual-Knowledge-Alignment-Network). We presented in detail the development, architecture, training, and evaluations of VITKAN on (i) Visual-Textual-Coreference (VTC), and (ii) Textual-Entity-Grounding (TEG) tasks. The VITKAN,

---

<sup>1</sup><https://yago-knowledge.org/>

takes in input  $d_i$  and applies an object detector, resulting in a set of bounding boxes labeled with classes of the YAGO Ontology. In  $d_t$  part, the VITKAN takes captions and process for textual entity recognition and linking them with YAGO ontology for background knowledge extraction. We trained the VITKAN by utilizing features data of the VTKEL1k\* dataset. For the evaluation of VITKAN algorithm, we used VTKEL1k\* and VTKEL30k datasets. The highest F1 scores of VTC task using VTKEL1k\* and VTKEL30k datasets are 0.742 and 0.739 respectively. We also checked the quality of VITKAN on the TEG task and compared the results with state-of-the-art approaches.

The outcomes of the thesis are the introduction of a novel VTKEL task, the development of two datasets, and solving the VTKEL problem by introducing VT-LINKER and VITKAN algorithms for the communities of CV, NLP, and Knowledge representation. One of the significant innovations of this thesis is the use of background-knowledge coming from a knowledge base (YAGO) with visual and textual data, and later explicitly handling it for subsymbolic approaches. The problem of VTKEL and their solutions with the help of VT-LINKER and VITKAN algorithms have opened new research directions in the areas of multimedia indexing, retrieval, and vision-language. Using our techniques in these areas can improve the accuracy of image captioning, visual-dialogue system, and visual-question answering tasks.

By using the Flickr30k-Entity dataset as a starting point for VTKEL datasets, we were limited to their annotations. As a result, we are missing some visual entity mentions shown in the image and described by text and vice-versa. Another limitation of this thesis is the use of KRN object detector, which is trained on the GOI dataset. By using KRN, we are limited to 500 visual classes of GOI dataset.

## Future work

In the future, we are interested to capture the relationships between pairs of visual-objects [12] (e.g. "man feeding dog", "man on bicycle") shown in the image and described in the text, and later linked these relationships (i.e. verb, predicate) of entities with the knowledge-base (e.g. SUMO Ontology [111]). Linking visual objects, textual entities, and their relationships with the knowledge-base can be used to efficiently understand the contents of the image and text. We are also interested to apply the problem of VTKEL to a dataset that includes more pictures and textual descriptions different from captions (e.g., short news with pictures).

In addition, we want to use the pipelines of VTKEL task on a short *clip of video*, and the natural language text, which described the contents of video (i.e. *Grounded Video Description* [112]). This problem consists of two subtasks, (i) grounding of noun-phrases in video-captions to the corresponding bounding-boxes in one of the frames of video, and (ii) linking the textual entities mention of noun-phrases to the corresponding class in the knowledge-base. This approach can be used to efficiently solve the tasks of video description [113, 114], video paragraph description [115, 116], and video indexing & retrieval [117].



# Bibliography

- [1] Asvin Gohil, Hardik Modi, and Shobhit K Patel. 5g technology of mobile communication: A survey. In *2013 international conference on intelligent systems and signal processing (ISSP)*, pages 288–292. IEEE, 2013.
- [2] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2014.
- [3] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014.
- [4] David Crystal. *A dictionary of linguistics and phonetics*, volume 30. John Wiley & Sons, 2011.
- [5] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
- [6] Dan Brickley, Ramanathan V Guha, and Brian McBride. Rdf schema 1.1. *W3C recommendation*, 25:2004–2014, 2014.
- [7] Bob DuCharme. *Learning SPARQL: querying and updating with SPARQL 1.1*. ” O’Reilly Media, Inc.”, 2013.
- [8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [9] Bryan Allen Plummer, Kevin Shih, Yichen Li, Ke Xu, Svetlana Lazebnik, Stan Sclaroff, and Kate Saenko. Revisiting image-language networks for open-ended phrase detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [10] Francesco Corcoglioniti, Marco Rospoche, and Alessio Palmero Aprosio. Extracting knowledge from text with pikes. In *14th International Semantic Web Conference (ISWC 2015)*, volume 1486, 2015.
- [11] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 801–809, 2018.
- [12] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

- [13] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217, 2008.
- [14] Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio. Frame-based ontology population with pikes. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3261–3275, 2016.
- [15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [17] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. Vtkel: a resource for visual-textual-knowledge entity linking. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 2021–2028, 2020.
- [18] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. Vt-linker: Visual-textual-knowledge entity linker. *The 24th European Conference on Artificial Intelligence (ECAI)*, 249:42–43, 2020.
- [19] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. On visual-textual-knowledge entity linking. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 190–193. IEEE, 2020.
- [20] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. Jointly linking visual and textual entity mentions with background knowledge. In *International Conference on Applications of Natural Language to Information Systems*, pages 264–276. Springer, 2020.
- [21] Shahi Dost, Luciano Serafini, and Alessandro Sperduti. Semantic interpretation of image and text. *The 17th International Conference of the Italian Association for Artificial Intelligence*, 2249:37–40, 2018.
- [22] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. Aligning and linking entity mentions in image, text, and knowledge base. *Journal of Data and Knowledge Engineering*, 2021.
- [23] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, 7(3):154–165, 2009.
- [24] Archana Goyal, Vishal Gupta, and Manish Kumar. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43, 2018.
- [25] Junwei Han, Dingwen Zhang, Gong Cheng, Nian Liu, and Dong Xu. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine*, 35(1):84–100, 2018.

- [26] Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020.
- [27] Neha Tilak, Sunil Gandhi, and Tim Oates. Visual entity linking. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 665–672. IEEE, 2017.
- [28] Tokinori Suzuki, Daisuke Ikeda, Petra Galuščáková, and Douglas Oard. Towards automatic cataloging of image and textual collections with wikipedia. In *International Conference on Asian Digital Libraries*, pages 167–180. Springer, 2019.
- [29] De-An Huang, Joseph J Lim, Li Fei-Fei, and Juan Carlos Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2183–2192, 2017.
- [30] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [31] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565, 2014.
- [32] Francis Ferraro, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell, et al. A survey of current datasets for vision and language research. *arXiv preprint arXiv:1506.06833*, 2015.
- [33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- [34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [36] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *arXiv preprint arXiv:2102.08981*, 2021.
- [37] Nilavra Bhattacharya and Danna Gurari. Vizwiz dataset browser: A tool for visualizing machine learning datasets. *arXiv preprint arXiv:1912.09336*, 2019.

- [38] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.
- [39] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664. Springer, 2020.
- [40] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.
- [41] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2601–2610, 2019.
- [42] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019.
- [43] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [44] Alane Suhr and Yoav Artzi. Nlvr2 visual bias analysis. *arXiv preprint arXiv:1909.10411*, 2019.
- [45] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.
- [46] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. Information extraction meets the semantic web: a survey. *Semantic Web*, (Preprint):1–81, 2020.
- [47] Lydia Weiland, Ioana Hulpus, Simone Paolo Ponzetto, and Laura Dietz. Using object detection, nlp, and knowledge bases to understand the message of images. In *International Conference on Multimedia Modeling*, pages 405–418. Springer, 2017.
- [48] Lydia Weiland, Ioana Hulpuş, Simone Paolo Ponzetto, Wolfgang Effelsberg, and Laura Dietz. Knowledge-rich image gist understanding beyond literal meaning. *Data & Knowledge Engineering*, 117:114–132, 2018.
- [49] Aparna Nurani Venkitasubramanian, Tinne Tuytelaars, and Marie-Francine Moens. Entity linking across vision and language. *Multimedia Tools and Applications*, 76(21):22599–22622, 2017.

- [50] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [51] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.
- [52] Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. Linking people in videos with “their” names using coreference resolution. In *European conference on computer vision*, pages 95–110. Springer, 2014.
- [53] Luciano Serafini, Ivan Donadello, and Artur d’Avila Garcez. Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation. In *Proceedings of the Symposium on Applied Computing*, pages 125–130, 2017.
- [54] Sibeiyang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4145–4154, 2019.
- [55] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4683–4693, 2019.
- [56] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.
- [57] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095, 2020.
- [58] Daniele Porello, Marco Cristani, and Roberta Ferrario. Integrating ontologies and computer vision for classification of objects in images. In *Proceedings of the Workshop on Neural-Cognitive Integration in German Conference on Artificial Intelligence*, pages 1–15, 2013.
- [59] Marco Bertini, Alberto Del Bimbo, and Carlo Torniai. Automatic video annotation using ontologies extended with visual information. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 395–398, 2005.
- [60] Hichem Bannour and Céline Hudelot. Towards ontologies for image interpretation and annotation. In *2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 211–216. IEEE, 2011.
- [61] Umut Akdemir, Pavan Turaga, and Rama Chellappa. An ontology based approach for activity recognition from video. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 709–712, 2008.



- [62] Juan Gómez-Romero, Miguel A Patricio, Jesús García, and José M Molina. Ontology-based context representation and reasoning for object tracking and scene interpretation in video. *Expert Systems with Applications*, 38(6):7494–7510, 2011.
- [63] Natalia Díaz Rodríguez, Manuel P Cuéllar, Johan Lilius, and Miguel Delgado Calvo-Flores. A fuzzy ontology for semantic modelling and recognition of human behaviour. *Knowledge-Based Systems*, 66:46–60, 2014.
- [64] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
- [65] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *European Conference on Computer Vision*, pages 696–711. Springer, 2016.
- [66] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [67] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016.
- [68] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019.
- [69] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.
- [70] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [71] Sibe Yang, Guanbin Li, and Yizhou Yu. Propagating over phrase relations for one-stage visual grounding. In *European Conference on Computer Vision*, pages 589–605. Springer, 2020.
- [72] Chenchen Jing, Yuwei Wu, Mingtao Pei, Yao Hu, Yunde Jia, and Qi Wu. Visual-semantic graph matching for visual grounding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4041–4050, 2020.
- [73] Tianyu Yu, Tianrui Hui, Zhihao Yu, Yue Liao, Sansi Yu, Faxi Zhang, and Si Liu. Cross-modal omni interaction modeling for phrase grounding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1725–1734, 2020.

- [74] Adrien Bennetot, Jean-Luc Laurent, Raja Chatila, and Natalia Díaz-Rodríguez. Towards explainable neural-symbolic visual reasoning. *arXiv preprint arXiv:1909.09065*, 2019.
- [75] Pranav Agarwal, Alejandro Betancourt, Vana Panagiotou, and Natalia Díaz-Rodríguez. Egoshots, an ego-vision life-logging dataset and semantic fidelity metric to evaluate diversity in image captioning models. *arXiv preprint arXiv:2003.11743*, 2020.
- [76] Ronghang Hu and Amanpreet Singh. Transformer is all you need: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021.
- [77] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [78] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [79] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [80] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [81] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [82] Francesco Corcoglioniti, Marco Rospocher, Roldano Cattoni, Bernardo Magnini, and Luciano Serafini. The knowledgestore: a storage framework for interlinking unstructured and structured knowledge. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 11(2):1–35, 2015.
- [83] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating nlp using linked data. In *International semantic web conference*, pages 98–113. Springer, 2013.
- [84] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.
- [85] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.
- [86] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.

- [87] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [88] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.
- [89] Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, and Tieniu Tan. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition*, 98:107075, 2020.
- [90] Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. Amc: Attention guided multi-modal correlation learning for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2644–2652, 2017.
- [91] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016.
- [92] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3536–3545, 2020.
- [93] Pedro Henrique Martins, Zita Marinho, and André FT Martins. Joint learning of named entity recognition and entity linking. *arXiv preprint arXiv:1907.08243*, 2019.
- [94] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [95] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.
- [96] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016.
- [97] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.
- [98] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. *arXiv preprint arXiv:2006.09920*, 2020.
- [99] Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. *arXiv preprint arXiv:1510.07493*, 2015.

- [100] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.
- [101] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [102] Siyang Yuan, Ke Bai, Liqun Chen, Yizhe Zhang, Chenyang Tao, Chunyuan Li, Guoyin Wang, Ricardo Henao, and Lawrence Carin. Weakly supervised cross-domain alignment with optimal transport. *arXiv preprint arXiv:2008.06597*, 2020.
- [103] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.
- [104] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.
- [105] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014.
- [106] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050, 2018.
- [107] Raymond A Yeh, Minh N Do, and Alexander G Schwing. Unsupervised textual grounding: Linking words to image concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6125–6134, 2018.
- [108] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5705, 2018.
- [109] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.
- [110] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.
- [111] Adam Pease, Ian Niles, and John Li. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28, pages 7–10, 2002.
- [112] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6578–6587, 2019.

- [113] Jie Xu, Haoliang Wei, Linke Li, Qiuru Fu, and Jinhong Guo. Video description model based on temporal-spatial and channel multi-attention mechanisms. *Applied Sciences*, 10(12):4312, 2020.
- [114] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. Multimodal video description. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1092–1096, 2016.
- [115] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016.
- [116] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*, 2020.
- [117] Luca Rossetto, Ralph Gasser, Jakub Lokoc, Werner Bailer, Klaus Schoeffmann, Bernd Muenzer, Tomas Soucek, Phuong Anh Nguyen, Paolo Bolettieri, Andreas Leibetseder, et al. Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019. *IEEE Transactions on Multimedia*, 2020.