Semantic Interpretation of Image and Text

Shahi Dost[1,2], Luciano Serafini[1](Supervisor), and Alessandro Sperduti[2](Co-supervisor)

[1] Fondazione Bruno Kessler, Italy
[2] University of Padova, Italy

Abstract. Semantic Interpretation of Image and Text consists in recognizing the entities and relations shown in the image and mentioned in the text and to align them with some ontological resource that contains structured knowledge about these entities and relations. A proper semantical description of entities and relations aligned in images and texts would allow for more accurate retrieval of images in the search tasks.
In this research, we propose a framework for Semantic Interpretation of Image and Text, which will utilize both the low level and semantic features of image and text by using background knowledge extracted from online Knowledge base. A successful approach for the Semantic Interpretation of Image and Text should address the following challenges. First, we have to recognize entities shown in images and describe in text. Second to make links between entities, third to make links with the entities in the Knowledge base and fourth relations between entities. To solve these complex tasks, we will use state of the art methods for image object detection and textual entities recognition. Furthermore, for mapping textual, and visual entities with entities in the knowledge base, we will use supervised machine learning techniques that exploit background knowledge. In order to provide a method for training our algorithms, and to evaluate properly the results we intend to develop a dataset consisting of images, image captions, bounding box annotations, links between visual and textual entities, linked to the knowledge base and the semantic meaning of entities.

Keywords: Artificial Intelligence · Knowledge Representation · Knowledge Base · Computer Vision · Natural Language Processing · Machine Learning.

1 Introduction

When we see a picture surrounded by text on social media, website or in a book, the first important thing is the objects (entities) in the picture, entities in the text, the information of entities, and what relations these entities have with each other. The problem of identifying and semantically synchronizing entities and relations shown in an image and mentioned in text, with the alignment in those knowledge bases which already have knowledge about these entities and relations will enrich the information for the machine. This semantic interpretation of

an image and text is the results of very complex processing which involves: 1) recognition of entities in image and text, 2) linking of entities in image and text, 3) alignment of these entities with the knowledge base and, 4) relations between entities. In my PhD, I am proposing a framework, which will recognized entities in image and text, relations between entities, links between visual and textual entities, and the links of these entities with the knowledge base which stored structural knowledge about these entities.

The source of information is the image with text, and the goal is to make links between visual and textual entities and then connect these entities with the Knowledge base. Fig. 1 shows the scenario in details. Fig. 1a shows an image with text, which describes *(the girl is eating ice-cream in Piazza Duomo)*. The main objects that appear in the image are *girl, ice-cream, hands, shirt* and relation *eating* between *girl* and *ice-cream*. Entities that are mentioned in text are *girl, ice-cream* and *Piazza Duomo* with relation *eating* between *girl* and *ice-cream*. The first step is to detect the image bounding boxes correspond to each entity in the image and textual entity, and make alignment between image and textual entities, then detect relations between entities as shown in Fig. 1b.

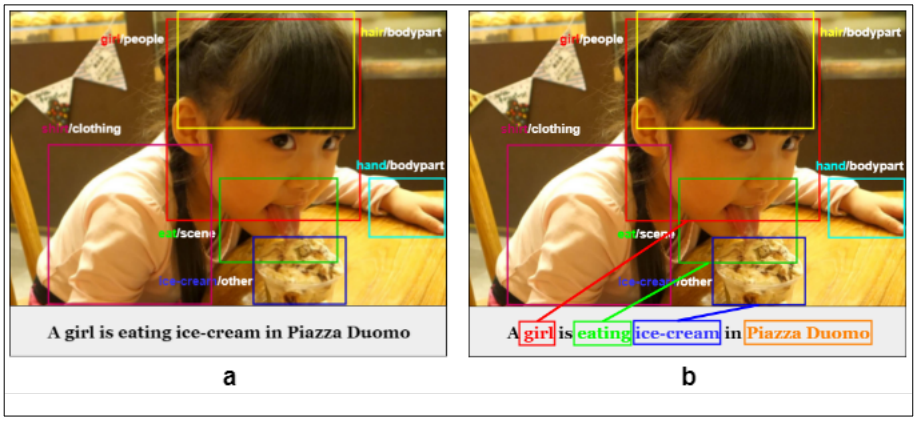

Fig. 1. (a) Image with visual entities and caption, (b) Alignment of visual (girl, ice-cream) and textual (girl, ice-cream, Piazza Duomo) entities, and relation (eating) between (girl) and (ice-cream)

After aligned visual and textual elements (i.e., entities and relations), the next step is to connect these objects shown in the image and describe in the text with some ontological resource that consists of structured knowledge about these entities and relations. The result is graphically represented in Fig. 2.

2 State of the Art

In literature, there is a lot of work which involves text and online knowledge bases [1] [11] but the work which involves images, text and knowledge bases are very few [8] [9]. Tilak et al. [8] proposed a basic framework, which links visual entities to DBpedia and Freebase. In their work, they used Flickr8k dataset for training and testing to connect the image regions and text entities directly to DBpedia and Freebase. Weiland et al. [9] proposed a methodology to understand the gist (message) convey by image regions and caption with related contents extracted from Wikipedia. Many problems like *automatic caption generation* [7], *visual relationship detection* [4], *phrase grounding* [10] and *automatic question answering* [2] involve image and text processing.

In the first phase of our proposed framework, we will recognize and links the visual and textual entities to the online knowledge base and then find the relations between entities. We will use state of the art objects detection [6] and textual entities recognition [3] methods. For more specifically to map the entities in image, text and knowledge base, we will propose a multimodal approach based on CNN and LSTM to correctly encode the entities and relations shown in an image, mentioned in the text and extracted from the knowledge base.

3 Research Methodology and Approach

To develop my framework for the problem of semantic interpretation of image and text, we start from building a dataset consists of images, text, and knowledge base alignment. We start from Flickr30k [5] dataset, which consists of images, text (set of captions which describe the contents of images), bounding box annotations and linking of noun phrases between images and text. Next to make links between visual and textual entities to ontological resource for structural knowledge about entities and relations. We used PIKES[3] for structural knowledge graphs extracting from text. We passed 158,915 captions of the Flickr30k dataset through PIKES, which generate knowledge graphs of these captions. For links formation between visual and textual objects to an ontological resource, we used YAGO[4] Knowledge base extracted by PIKES. We are using YAGO ontology for structured and semantic knowledge discovery, which is a massive semantic knowledge base, build from WordNet[5], Wikipedia (DBpedia)[6] and GeoNames[7]. By aligning entities of text with YAGO, we can infer that these entities are also aligned with visual objects as shown in Fig. 2. After connecting visual objects with YAGO ontology, we can extract structured and semantic knowledge information of objects shown in the image and describe in the text. To solve these challenging tasks, we developed a dataset called *Visual Knowledge Stored*

[3] `https:http://pikes.fbk.eu/`
[4] `https://www.mpi-inf.mpg.de/yago-naga/yago/`
[5] `https://wordnet.princeton.edu/`
[6] `https://wiki.dbpedia.org/`
[7] `http://www.geonames.org/`

Flickr30k (VKS Flickr30k). VKS Flickr30k consists of images, image captions, bounding box annotations, links between visual and textual entities, visual and textual entities linked to the Knowledge base and semantic meaning of entities.

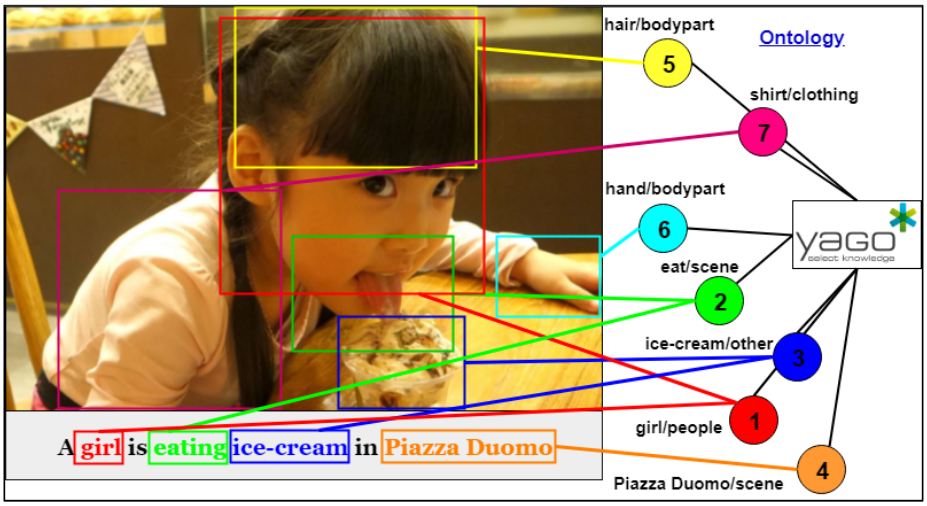

Fig. 2. Alignment of visual and textual objects (entities) and relation (eating) between (girl) and (ice-cream) with the Knowledge base (YAGO)

For visual and textual entities recognition, we will use state of the art object detection [6] and name entity recognition [3] algorithms. For more specifically to map the entities in image, text and knowledge base, we will propose a multimodal approach based on Convolutional neural network (CNN) and Long short-term memory (LSTM) to correctly encode the entities shown in the image, mentioned in the text and extracted from the knowledge base.

In the first phase of the proposed framework, VKS Flickr30k dataset will be used for training and testing algorithms to map textual and visual entities with entities in the knowledge base. In the second phase, we will extend the work for relations (events) detection and identification between entities. To evaluate the results of our proposed framework, we will use accuracy, precision, recall (sensitivity) and F1-score. We will compare our proposed framework with the existing state of the art approaches [8] [9].

4 Preliminary Results

We develop VKS Flickr30k dataset, which stored enrich details of image and textual entities and linked entities to a structural knowledge base. VKS Flickr30k dataset consists of 31,783 images with five captions per image and 244,035 coreference chains (connect single bounding box entity to multiple instances in the

image captions) to link 275,775 bounding box to 513,644 textual objects (entities). Visual and textual entities are connected to YAGO knowledge base by YAGO entity ids. Flickr30k entities are divided into eight distinct classes of people, clothing, bodyparts, animals, vehicles, instruments, scene and other.

During the development of VKS Flickr30k dataset, we used Flickr30k dataset for images, text, bounding box annotations and links between visual and textual phrases. To connect these noun phrases with the Knowledge base, we passed text through PIKES for semantic knowledge graph extraction. We connect YAGO entity Ids with textual entities extracted by PIKES. Fig. 3 shows step by step processes of VKS Flickr30k development.

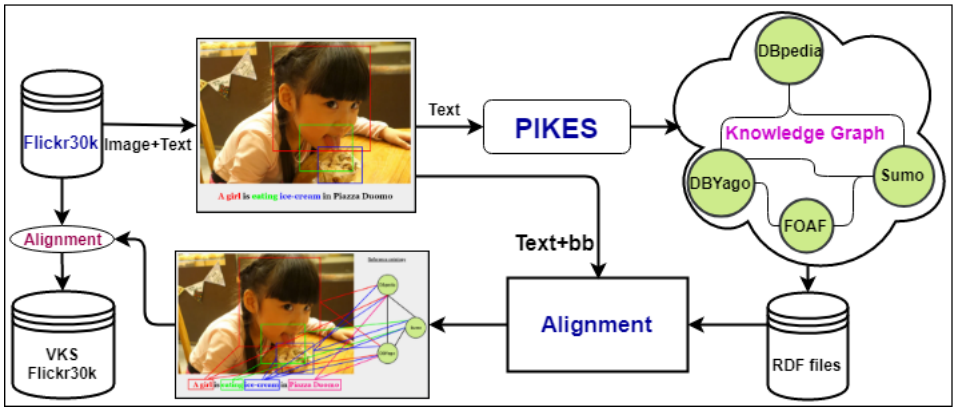

Fig. 3. VKS Flickr30k development stages

5 Conclusions

In my PhD, we are proposing a framework for semantic interpretation of image and text, which will link entities and relations shown in the image and mentioned in the text and to align them with reference Knowledge base that contains structural knowledge about these entities and relations. For training our algorithms to solve these complex tasks and accurately evaluate the results, we develop a dataset called VKS Flickr30 for semantic interpretation of image and text. In the future, we will use state of the art methods for image objects detection and textual entities recognition. Also, we will use supervised machine learning techniques for mapping visual, and textual entities with entities in the knowledge base. In the second phase, we will extend the work for relations (events) detection and identification between entities.

References

1. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with wikipedia. Artificial intelligence **194**, 130–150 (2013)
2. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: End-to-end module networks for visual question answering. CoRR, abs/1704.05526 **3** (2017)
3. Marrero, M., Sanchez-Cuadrado, S., Lara, J.M., Andreadakis, G.: Evaluation of named entity extraction systems. Advances in Computational Linguistics, Research in Computing Science **41**, 47–58 (2009)
4. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive image-language cues. In: Proc. ICCV (2017)
5. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
7. Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.J.: Deep reinforcement learning-based image captioning with embedding reward. arXiv preprint arXiv:1704.03899 (2017)
8. Tilak, N., Gandhi, S., Oates, T.: Visual entity linking. In: Neural Networks (IJCNN), 2017 International Joint Conference on. pp. 665–672. IEEE (2017)
9. Weiland, L., Hulpuş, I., Ponzetto, S.P., Effelsberg, W., Dietz, L.: Knowledge-rich image gist understanding beyond literal meaning. Data & Knowledge Engineering (2018)
10. Yeh, R.A., Do, M.N., Schwing, A.G.: Unsupervised textual grounding: Linking words to image concepts. In: Proc. CVPR. vol. 8 (2018)
11. Zugarini, A., Morvan, J., Melacci, S., Knerr, S., Gori, M.: Combining deep learning and symbolic processing for extracting knowledge from raw text. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition. pp. 90–101. Springer (2018)