SHORT TERM SCIENTIFIC MISSION (STSM)
SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: IC1307
STSM title: Learning spatial templates of actions and implicit spatial language
STSM start and end date: 27-06-2017 to 20-09-2017
Grantee name: Guillem Collell Talleda

PURPOSE OF THE STSM:

(max.200 words)

The research carried out during the stay at the Computer Vision Laboratory (CVL) in ETH focuses on building automated methods for acquiring common sense spatial knowledge. Endowing machines with common sense knowledge is one of the most important long-term goals of artificial intelligence research. Lack of common sense has been recurrently argued as one of the main reasons that prevents machines from exhibiting more human-like behavior when solving tasks. In particular, the research conducted during the short term mission is aimed at advancing our knowledge and existing methods for understanding spatial language. To this end, we design neural-network models that learn to predict common sense spatial knowledge that is left implicit in language. The models learn from multimodal, image-text paired data with annotations and we allow for both, a quantitative evaluation of their performance and a qualitative visualization of their predictions. The motivation for such interdisciplinary collaboration is to tackle the above challenges by combining the prior knowledge in natural language processing and representation learning of the visiting researcher with the extensive expertise in computer vision of the host laboratory.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

(max.500 words)

Overview of the research. During the short term mission in ETH, the visiting researcher has been working on the problem of designing automatic methods to acquire common sense spatial knowledge and in building predictive models for this type of knowledge. To enable learning such models, we have leveraged paired text-visual data from where we can extract both, the spatial relationships between objects (visual) and its corresponding structured (Subject, Relationship, Object) linguistic predicate (text).

Collaboration. Overall, the visiting researcher has greatly benefitted from the collaboration with the host laboratory in ETH. The main advantages obtained were: (i) expert advice in computer vision, and (ii) GPU computing power. In particular, regular discussions with postdocs and more spaced meetings with Prof. Luc Van Gool were specially fruitful. In addition, the availability of GPU computing power allowed to significantly speed up the research by enabling to try out many different models in a short time.

Relevance of the problem. To provide machines with common sense is one of the major long term goals of artificial intelligence research. Common sense knowledge regards knowledge that humans have

Funded by the Horizon 2020 Framework Programme
of the European Union

acquired through a lifetime of experiences. It is crucial in language understanding because a lot of content needed for correct understanding is not expressed explicitly but resides in the mind of communicator and audience. In addition, humans rely on their common sense knowledge when performing a variety of tasks including interpreting images, navigation and reasoning, to name a few. Representing and understanding spatial knowledge are imperative for any agent (human, animal or robot) that navigates in a physical world.

State-of-the-art. Current methods that learn and represent spatial knowledge focus on ***explicit*** spatial language. The representation of spatial knowledge is often modelled with ***spatial templates***, i.e., regions of acceptability of two objects that are related by means of an ***explicit*** spatial relationship (e.g., "on", "below", "above", "left", "right", etc.). In contrast to prior work that restricts spatial templates to explicit spatial prepositions (e.g., "glass *on* table"), here we extend this concept to ***implicit*** spatial language, i.e., those relationships (generally actions) for which the spatial arrangement of the objects is only implicitly implied (e.g., ``man riding horse") but not explicitly mentioned. In contrast to linguistic utterances containing explicit relationships, predicting spatial arrangements from implicit spatial language requires significant common sense spatial understanding. Prior approaches on explicit spatial language include, among others, the work of Yatskar et al. (2016) who propose a model to extract common sense facts from annotated images using co-occurrence statistics such as point-wise mutual information (PMI). These facts include six spatial relationships. Alternatively, Malinowski and Fritz (2014) propose a learning-based pooling approach to retrieve images given queries of the form (object_1, spatial_preposition, object_2). They learn the parameters of a spatial template for each explicit spatial preposition (e.g., "left" or "above") which computes a soft spatial fit of two objects under the relationship. E.g., an object on the left of the referent object obtains a high score for the "left" template and low for the "right" template.

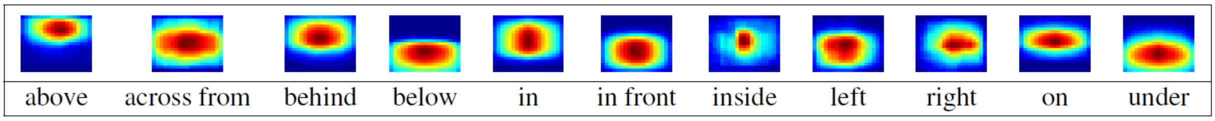

Figure 1. Spatial templates, adapted from Malinowski and Fritz (2015).

LITERATURE:

Malinowski, M., & Fritz, M. (2014). A pooling approach to modelling spatial relations for image retrieval and annotation. arXiv preprint arXiv:1411.5190.

Yatskar, M., Ordonez, V., & Farhadi, A. (2016, June). Stating the Obvious: Extracting Visual Common Sense Knowledge. In *HLT-NAACL* (pp. 193-198).

DESCRIPTION OF THE MAIN RESULTS OBTAINED

Our approach. To shed light on the problem and research questions described above, we have proposed both, a task and neural-network models to solve this task.

- **Task:** The proposed task consists in predicting the 2D relative spatial arrangement of two objects under a relationship given a structured text input of the form (Subject, Relationship, Object). More specifically, we predict the Object's location and size (**output**) given triplets (Subject, Relationship, Object) and the location and size of the Subject (**input**).

- **Model:** The proposed feed-forward neural network models map the input triplets (Subject, Relationship, Object) to the Object's location (relative to the Subject) and its size. These models employ an embedding layer to incorporate the representations of the words of the input triplet. These embeddings are initialized with pre-trained word embedding (Mikolov et al., 2013). A few hidden layers are also added in order to compose the embeddings of the triplets.

Datasets employed: We employed the Visual Genome dataset as our source of annotated data (Krishna et al. 2017). The Visual Genome consists of ~108K images containing ~1.5M human-annotated (Subject, Relationship, Object) instances with bounding boxes for Subject and Object.

Implementation details: Our models were implemented in Python 2.7 with the Keras framework running on Tensorflow backend. CUDA and CuDNN accelerated GPU support was used to speed up learning the models.

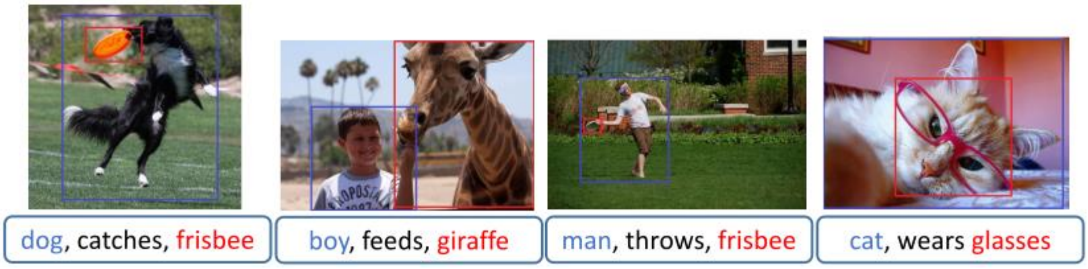

Figure 2. Sample of images with object boxes from the Visual Genome dataset.

Empirical results. The main results obtained in our experiments were:

- Models showed high predictive power for the Object's location in implicit spatial relationships. In particular, the performance is comparable to that obtained with explicit spatial prepositions. For example, above/below classification (i.e., predicting whether the Object is above or below the Subject) yielded an accuracy over 80%. Correlations between predicted and actual Object coordinates and size were as high as 0.9.
- Two model architectures have been proven successful in the task.
- The models are capable of predicting correctly unseen (Subject, Relationship, Object) combinations (e.g.,``man walking dog") by generalizing from the triplets present in the training data. Performance does not drop significantly in such scenario.
- Next, we go one step further by presenting the models with unseen objects (e.g., ``dog"). We find that transferring knowledge from word embeddings enables the models to output accurate spatial predictions.
- Qualitative visualizations of the model predictions further support the quantitative evaluation.

Conclusions. First, the good performance of our models reveals that spatial locations are to a large extent predictable from implicit spatial language. Second, the good predictive power of the models when predicting unseen combinations (Subject, Relationship, Object) and words shows that the models acquire solid common sense spatial knowledge allowing for such generalization.

Research outcomes. The work and results discussed above has been submitted to the AAAI 2018 conference and is currently under review.

LITERATURE:

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision 123(1):32–73.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. NIPS (pp. 3111-3119).

FUTURE COLLABORATIONS (if applicable)

We foresee that the completed work (submitted to AAAI 2018) will lead to follow up studies and more research outputs in the future. So far, we engaged in a series of discussion sessions with two postdoc researchers from CVL aimed at defining future research goals. In particular, we plan to extend the aforementioned models that acquire and predict common sense spatial knowledge in the direction of learning such knowledge in a weakly supervised manner—instead of a fully supervised setting. We additionally plan to employ the spatial priors that our models are able to output in an image retrieval task. A number of studies have shown that spatial information (e.g., obtained in images or in text) can improve performance in several natural language and computer vision tasks. For example, Elliott and Keller (2013) have shown that accounting for the spatial structure in images can improve the tasks of image captioning (i.e., describing images with text) and image retrieval. Shiang et al. (2017) have shown that by leveraging spatial knowledge (which they obtain by mining texts and labeled images in the form of object co-occurrences and relative positions between objects), they are able to improve the performance over the state-of-the-art in object recognition. Hence, we foresee that our models have potential to boost performance in the task of image retrieval by properly integrating the predictions of our models (taken as spatial priors) into the retrieval system.

LITERATURE:

Elliott, D., & Keller, F. (2013). *Image Description using Visual Dependency Representations.* EMNLP (Vol. 13, pp. 1292-1302).

Shiang, S. Rosenthal, S., Gershman, A., Carbonell, J., Oh. J., (2017) Vision-Language Fusion for Object Recognition. AAAI