Sujet de thèse: Deep Neural Networks and Multimodal Semantic Role Labeling for Question Answering

Context

In recent years, deep learning (Bengio 2013, LeCun 2015) has established state of the art techniques for many tasks in domains involving semantic data processing like vision, speech decoding, natural language processing or even language translation (Krizhevsky 2012, Graves 2013, Mikolov 2013, Socher 2014, Cho 2014). About 2 years ago (2014), this research community started to consider more complex, higher level tasks involving the use of background knowledge and of analogical reasoning. Question Answering (Q/A) is such a task involving the ability to perform sophisticated inferences. Q/A has been first studied in the NLP community for answering factual questions on text modalities. In artificial intelligence it has the broader acceptation of developing systems able to answer questions using facts extracted from data. This requires the ability to automatically: 1) extract knowledge from semantic data, 2) learn a relevant representation of this knowledge, 3) develop learning and inference mechanisms for high level tasks using this representation. General purpose Q/A is a long term objective. For now, research has focused on less ambitious targets, but it is undergoing a fast development at the junction of different scientific communities. For example, researchers at Facebook have make Q/A one of their priority and have recently developed a benchmark based on artificially generated datasets, requiring the ability to understand sets of facts and assertions and to perform reasoning so as to answer questions (Weston 2015). For real world applications, knowledge is to be extracted from semantic data, most often text or images. In recent years, deep learning has been used to learn joint representations of text and images, for tagging image with words, for generating sentences describing what is in an image (Vinyals 2014, Kiros 2014), and more recently for Visual Query Answering (Antol 2015), a new Q/A task consisting in answering questions regarding the content of an image.

Thesis project

The thesis participates to this challenge: developing systems able to extract facts and relations from data that improve our understanding of semantic. More precisely, it targets the task of semantic role labeling (SRL) for spatial and temporal information. The goal of this research is to couple textual modality together with perceptual visual input for building structured multi-modal semantic representations for the recognition of objects and actions and their spatial and temporal relations. Humans perceive the physical world according to space and time. They move in the physical world by encoding objects and their relations and expressing them using language. Contextual world knowledge and common sense are essential to understand linguistic expressions and relations referring to objects, actions and events. This knowledge is hard to extract from text alone because it is prone to complex formulations and to ambiguities. While text provides a versatile semantic description of situations and events, perceptual visual information grounds this information in the physical world. Spatial and temporal role labeling for Q/A are recent topics in NLP. They have been part of the SemEval[1] challenge in 2015 using only the textual modality. Our goal here is to go beyond the limitations of current systems by using enriched multi-modal content.

[1] http://alt.qcri.org/semeval2015

Concerning the methodology, we will focus on Neural Network based representation learning. NNs have already shown their potential for learning joint multimodal representations for different tasks. They allow us combining multiple complex processing modules into end to end architectures for performing complex information processing tasks involving static (images) and sequence (text) data.

The thesis objective is then to develop neural network models for grounding our understanding of human language in the physical world using complementary visual information with a focus on representing objects, actions and two fundamental types of relations between real world objects and actions: spatial and temporal relations. The thesis will be organized around three main tasks.

- Labeling semantic spatial and temporal semantic roles on textual data

The first step is to analyze the potential of recent deep learning sequence models for extracting spatial and temporal information (objects, actions, events and their relations) from text sentences. Although representation learning has been used in different context, SRL has never been studied from this perspective. We will consider different families of recurrent gated NN (Graves 2013) models and develop new NNs for these specific tasks. Relation extraction may require several steps of reasoning/ inference. This is a grand challenge in Artificial Intelligence. Recently, neural network architectures, called memory models have been proposed specifically with this type of task as targets (Kumar 2016, Sukhbaatar 2015). They are trained to learn sequence item representations together with some attention mechanism able to combine or select the representations relevant for a given task. The thesis will explore their use for learning to perform complex inferences for the extraction of spatio-temporal relations.

- Integrating visual information for spatial role labeling

Up to now, in multi-modal information processing, textual information has been used to help computer vision models in tasks such as image captioning, and description generation, video description, retrieval. We take here a paradigm shift and rely on computer vision as an aid to enrich textual description. For this purpose, we will develop NN models for learning joint representations for objects, actions and their relations using images and text. Recently, very large databases have been developed that provide precise labeling of object and relations in images together with their textual description (e.g. Krishna et al 2016). We will used enhanced version of these data to extract joint multi-modal representation of textual and image entities pertaining to spatial relations between objects. We focus here on spatial information which concern still images, a more reasonable objective for the thesis than dealing with time and videos.

- Integrating multi-modal representation into spatial language processing systems

The integration of multi-modal representation into language processing systems is a completely new research path. We will study how to integrate the multi-modal representations mentioned above in order to enhance spatial labeling models. We will start from the systems developed for language alone and use the enhanced representations together with NN using attention mechanisms for spatial role labeling.

International cooperation

This work comes in the context of a collaboration with the Language Intelligence and Information Retrieval group at KU Leuven in Belgium. Pr S. Moens, head of the group is an internationally renowned figure in NLP. She is chair of the European Chapter of the Association for Computational Linguistics (ACL). The PhD will have the opportunity for long term visits to the KUL lab.

References

(Antol 2015) Antol S., Agrawal A., Lu J., Mitchell M., Batra D., Zitnick C.L., Parikh D.,VQA: Visual Question Answering, International Conference on Computer Vision (ICCV),2015

(Bengio 2013) Y. Bengio, A. Courville, P. Vincent, "Representation Learning: A Review and New Perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, Aug., 2013

(Kiros 2014) Kiros R., Salakhutdinov R. and Zemel R. S. Unifying visual-semantic embeddings with multimodal neural language models. In arXiv:1411.2539, 2014.

(Krishna et al 2016) Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia-Li, David Ayman Shamma, Michael Bernstein, Li Fei-Fei, Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, http://arxiv.org/abs/1602.07332

(Krizhevsky 2012) Krizhevsky, A., Sutskever, I. and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks NIPS 2012: Neural Information Processing Systems

(Kumar 2016) Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., … Socher, R. (n.d.). Ask Me Anything: Dynamic Memory Networks for Natural Language Processing., Arxiv.

(Lecun 2015) LeCun, Y., Bengio, Y. and Hinton, G. E.,Deep Learning**.** Nature, Vol. 521, pp 436-444.

(Mikolov 2013) Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS). pp. 3111–3119 (2013)

(Mnih 2014) Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent Models of Visual Attention. In Advances in Neural Information Processing Systems 27 (Vol. 27, pp. 1–9).

(Silberer 2014) Silberer, C., Lapata, M.: Learning grounded meaning representations with autoencoders*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 721–732 (2014)

(Socher 2014) Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics (TACL) 2*, 207–218 (2014)

(Sukhbaatar 2015) Sukhbaatar, Sainbayar; Szlam, Arthur; Weston, Jason; Fergus, Rob, End to End Memory Networks, NIPS 2015

(Vinyals 2014) Vinyals O., Toshev A., Bengio S., Erhan D.: Show and tell: a neural image caption generator, ArXiv 2014.

(Weston 2015) Weston J., Bordes A., Chopra S., Mikolov T.: Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. CoRR abs/1502.05698 (2015)