Post-doc position (CEA List and LISN) :

Injecting Knowledge into Multimedia Entity Representation

In the context of the research project *MultimEdia Entity Representation and Question Answering Tasks* (ANR 2020-2024), a postdoctoral position is proposed for highly motivated candidates interested in multimedia understanding and natural language processing.

Context

Exploiting multimedia content often relies on the correct identification of entities in text and images. A major difficulty for understanding a multimedia content lies in its ambiguity with regard to the actual user needs, for instance when identifying an entity from a given textual mention or matching a visual object to a query expressed through language [ADJ20a,ADJ20b].

The MEERQAT (https://www.meerqat.fr) project addresses the problem of analyzing ambiguous visual and textual content by learning and combining their representations and by taking into account the existing knowledge about entities. It aims at solving the Multimedia Question Answering (MQA) task, which requires answering a textual question associated with a visual input like an image, given a knowledge base (KB) containing millions of unique entities and associated text. The post-doc specifically addresses the problem of injecting knowledge into multimodal entities to be able to answer questions that relate to them. Other partners of the project work on the visual, textual and KB representation, and on the entity disambiguation.

Main activities

Entities can be represented by different modalities, in particular by visual and textual content. In a common space, an entity can thus be represented by several vectors, that need to be combined into a unique representation that reflects the similarity of the related entities. In such a context, a promising approach consists of learning a visual representation from natural language supervision [RAD21] relying on large datasets by a simple learning strategy based on contrastive predictive coding [OOR18], adapted to text and visual modalities [ZHA20]. The learned representation allows to address multiple cross-modal tasks and provide a large-scale vocabulary that is adapted to general audience in a given language. It exhibits state of the art performance on several tasks and can even exceed humans on certain tasks. However, it does not include any structural information from a knowledge base. The main task of the post doc will thus consist in injecting such prior knowledge into the entity representation to address Multimedia Question Answering. Some approaches were recently proposed to do so in the context of caption generation [GOE20].

We consider entities such as persons, places, objects or organizations (NGOs, companies...). Depending on the type of an entity, the information to take into account in its representation is not obvious. If a person can probably be associated with a couple of mentions and images, it becomes less obvious for other types of entities. For instance, a company can be associated with its logo, but also with its main products or even its main managers (CEO, CTO . . .). In the same vein, a location can be represented by many pictures, and a large one such as a city by

some emblematic buildings or places. The second task of the post-doc will consist to determine the appropriate information to include in the representation of a given entity, depending on its type.

Position and Application

The post-doc will be supervised by CEA and LISN. The candidate will be hired by CEA (Palaiseau, near Paris, France) for a 18-months post-doc. The LISN is located close to CEA on the Paris-Saclay University Campus.

The salary depends on qualifications and experience. This will include social coverage (health, unemployment, retirement).

The postdoc will have access to large supercomputers equipped with multiple GPUs and large storage for experiments, in addition to a professional laptop.

To apply to the position, send a CV (including publication list or a URL pointing to it, such as Google Scholar) and a cover letter to Hervé Le Borgne <herve.le-borgne@cea.fr>, Olivier Ferret <olivier.ferret@cea.fr>, Sahar Ghannay <Sahar.Ghannay@limsi.fr> and Anne Vilnat <Anne.Vilnat@limsi.fr>.

Profile

- PhD in Natural Language Processing, Computer Vision, Machine Learning or other relevant fields
- Strong publication record, with accepted articles in top-tier conferences and journals of the domain
- Solid programming skills (pytorch/tensorflow). Publicly available project will be appreciated
- Ability to communicate and collaborate at the highest technical level
- Experience on using GPUs on a supercomputer (e.g. with SLURM or similar tool) will be appreciated

References

[ADJ20a] O. Adjali, R. Besançon, O. Ferret, H. Le Borgne and B. Grau, Multimodal Entity Linking for Tweets, 42nd European Conference on Information Retrieval (ECIR): Advances in Information Retrieval, 2020

[ADJ20b] O. Adjali, R. Besançon, O. Ferret, H. Le Borgne and B. Grau, Building a Multimodal Entity Linking Dataset From Tweets, 12th International Conference on Language Resources and Evaluation (LREC), 2020

[GOE20] A. Goel, B.Fernando, T-S. Nguyen, H. Bilen. Injecting Prior Knowledge into Image Caption Generation. ECCV 369-385, 2020

[OOR18] Oord, A. v. d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. arXiv:1807.03748, Jul 2018.

[RAD21] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. ArXiv 2103.00020, Feb 2021.

[ZHA20] Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.;Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747, 2020.