Post-doc position (CEA List and INRIA) :

Multimodal entity representation and disambiguation

In the context of the research project “MultimEdia Entity Representation and Question Answering Tasks” (MEERQAT – ANR 2020-2024), a postdoctoral position is proposed for highly motivated candidates interested in computer vision and multimedia understanding.

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.

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 representing multimodal entities at large scale to disambiguate them. Other partners of the project work on the visual, textual and KB representation, as well as on question answering based on the three modalities..

Main activities

We consider entities such as a person, a place, an object or an organization (NGO, company...). Entities can be represented by different modalities, in particular by visual and textual content. However, a given mention of this entity is often ambiguous. For example, the mention «Paris» refers not only to the city of France (and a dozen of other cities in the world), but also to the model Paris Hilton and the Greek hero of the Trojan War. An additional visual content linked to the mention can greatly help to disambiguate, although the visual content itself carries other ambiguities. We also consider a third type of information, namely links between entities within a knowledge base. The task of Multimedia Question Answering needs all these three modalities to be solved.

The postdoctoral associate will work on the representation of entities described by several modalities, with a particular emphasis on the use of visual data to help in search and linking of entities. The goal is to not only disambiguate one modality by using another [ROS18,KAM21], but also to jointly disambiguate both by representing them in a common space. Most of state of the art representation of visual and textual content rely on neuronal models. There also exist embeddings that reflect the links in a knowledge base [WAN17]. Many works address cross-modal tasks between two of these modalities, relying on such representations projected in a common space, in order to minimize a loss corresponding to the task of interest, such as visual question answering (VQA) [MAL14, ANT15, BEN17, SHA19] or zero-shot learning [LEC19, SKO21]. Other approaches identify attributes in the visual content through a pre-trained model, then query a knowledge base to map it to the textual modality and learn a knowledge-based

VQA model [WU16, WAN17]. Such approaches have been extended to include structural facts that link the attributes [WAN18] and common-sense knowledge [MAR21, WU21]. Other works address VQA involving some knowledge on named entities, although still limited to the sole type of persons [SHA19b]. These last approaches require a quite structured knowledge, but others allow more general sources of knowledge, including free-form text found on the Web [MAR19]. For more specific use cases, it is also possible to create an ad-hoc knowledge base [GAR20].

However, to tackle the MQA task of interest in the MEERQAT project, one must address these issues at large scale, with a high level of ambiguity requiring fine reasoning on the entities. Depending on the type of an entity, the information to take into account in its representation is not obvious. A person may be associated with just a couple of mentions and images, but the situation becomes more complex for other types of entities. For instance, a company may be associated with its logo, but also with its main products or even its managers (CEO, CTO...). In the same vein, a location may be represented by many pictures, and a city by landmark buildings or places.

We aim at determining the appropriate information to include in the representation of a given entity. Hence, in a common space, an entity can be represented by several vectors, that need to be combined into a unique representation that reflects the similarity to 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, which is crucial for visual reasoning.

Position and Application

The candidate will be hired by CEA (Palaiseau, near Paris, France) for a 18-month post-doc. A stay of 6 months at INRIA (Rennes, France) is planned during this period, provided that the health context allows it. The additional costs resulting from this stay will be covered by the CEA.

The salary depends on qualifications and experience.

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) and a cover letter to Hervé Le Borgne <herve.le-borgne@cea.fr>, Yannis Avrithis <yannis@avrithis.net>, Laurent Amsaleg <Laurent.Amsaleg@irisa.fr> and Ewa Kijak <Ewa.Kijak@irisa.fr>.

Profile

- PhD in Computer Vision, Machine Learning, Natural Language Processing 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

[ANT15] Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. VQA: Visual Question Answering. In Proc. ICCV, 2015.

[BEN17] Ben-Younes, H.; Cadene, R.; Cord, M.; and Thome, N. MUTAN: Multimodal tucker fusion for visual question answering. In Proc. ICCV, 2017.

[CHE20] Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations, In Proc. ICML, 2020.

[GAR20] Garcia, N.; Otani, M.; Chu, C.; Nakashima, Y. KnowIT VQA: Answering Knowledge-Based Questions about Videos, In Proc. AAAI, 2020.

[KAM21] Kamath, A.; Singh, M.; LeCun, Y.; Misra, I.; Synnaeve, G.; Carion, N. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. arXiv preprint arXiv:2104.12763, 2021.

[LEC19] Le Cacheux, Y.; Le Borgne, H.; Crucianu, M. Modeling Inter and Intra-Class Relations in the Triplet Loss for Zero-Shot Learning. In Proc. ICCV, 2019.

[MAL14] Malinowski, M.; Fritz, M. A multi-world approach to question answering about real-world scenes based on uncertain input. In Proc. NIPS, 2014.

[MAR19] Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. OK-VQA: A visual question answering benchmark requiring external knowledge. In Proc. CVPR, 2019.

[MAR21] Marino, K.; Chen, X.; Parikh, D.; Gupta, A.; Rohrbach, M. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. In Proc. CVPR, 2021.

[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.

[ROS18] Rosenfeld, A.; Biparva, M.; Tsotsos, J. K. Priming Neural Networks. In Proc. CVPR, 2018.

[SHA19] Shah, M.; Chen, X.; Rohrbach, M.; Parikh, D. Cycle-consistency for robust visual question answering. In Proc. CVPR, 2019.

[SHA19b] Shah, S.; Mishra, A.; Yadati, N.; Talukdar, P. P. KVQA: Knowledge-aware visual question answering. In Proc. AAAI, 2019.

[SKO21] Skorokhodov, I.; Elhoseiny, M. Class Normalization for (Continual)? Generalized Zero-Shot Learning. arXiv:2006.11328, 2021.

[WAN17] Wang, Q.; Mao, Z.; Wang, B.; Guo, L. Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12):2724–2743, 2017.

[WAN17b] Wang, P.; Wu, Q.; Shen, C.; Dick, A.; Van Den Henge, A. Explicit knowledge-based reasoning for visual question answering. In Proc. IJCAI, 2017.

[WAN18] Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and van den Hengel, A. FVQA: Fact-based visual question answering. IEEE Trans. PAMI 40(10):2413–2427, 2018.

[WU16] Wu, Q.; Wang, P.; Shen, C.; Dick, A.; van den Hengel, A. Ask me anything: Free-form visual question answering based on knowledge from external sources. In Proc. CVPR, 2016.

[WU21] Wu, J.; Lu, J.; Sabharwal, A.; Mottaghi, R. Multi-Modal Answer Validation for Knowledge-Based VQA. arXiv preprint arXiv:2103.12248, 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.