Internship: probing joint vision-and-language representations

Advisors: Emmanuelle Salin, Stephane Ayache & Benoit Favre

October 27, 2020

Context Recent advances in deep learning have enabled exciting applications in the context of multimodal processing involving images and texts, such as visual question answering [1], visual dialog [4], image captionning [14], text undersanding in multimodal context [5]... This internship is focused on exploring representations trained to perform such tasks.

Vision-and-language representations are typically extracted with neural networks drawing from the transformers architecture, pre-trained with self-supervision on large datasets, such as Conceptual Captions [11] or MSCOCO [10]. The resulting family of architectures generally involve representing objects extracted from the image as embeddings, and concatening them with word embeddings associated to the text, before feeding them to multiple layers of attention mechanisms [12, 13, 9, 2].

Problem statement The central question of the internship is to verify whether those multimodal representations lead to the emergence of high-level semantic and structural properties from the simple self-supervision tasks used in pre-training, such as predicting masked inputs across modalities.

To explore this question, the intern will rely on the methodology developped by the natural language processing commnuity for probing representations from transformers from the BERT family for text. In that monomodal context, proposed probes were simple canonical tasks such as predicting the length of the sentence, part-of-speech tags, or syntactic structures from the representations with linear models, and without any fine-tuning [7, 8, 3, 6].

Expected contribution The candidate will propose probes and analysis methods for testing high-level structural and semantic concepts in various vison-and-language representations. In particular, the candidate will explore how data, neural architecture, and self-training tasks affect such probes.

The work will be implemented within the MMF framework[1] which itself relies on Pytorch. It is assumed that the candidate has the following qualities:

- Excellent knowledge of deep learning methods
- Extensive experience with implementing Pytorch models
- Great scientific writing skills
- A hunch for the challenges of doing research

The internship will take place at LIS[2]/CNRS in Marseille during spring 2021 and will be held in the context of Emmanuelle Salin's thesis on understanding the generality of multimodal representations. Additional pointers to relevant literature are available[3]. Salary will be around 500 euros/month for a duration of 5-7 months.

[1] `https://github.com/facebookresearch/mmf`
[2] `https://www.lis-lab.fr/`
[3] `https://github.com/yuewang-cuhk/awesome-vision-language-pretraining-papers`

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[3] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[4] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[5] Sebastien Delecraz, Leonor Becerra-Bonache, Benoit Favre, Alexis Nasr, and Frederic Bechet. Multimodal machine learning for natural language processing: Disambiguating prepositional phrase attachments with images. *Neural Processing Letters*, pages 1–27, 2020.

[6] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. When bert forgets how to pos: Amnesic probing of linguistic properties and mlm predictions. *arXiv preprint arXiv:2006.00995*, 2020.

[7] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

[8] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[9] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[11] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[12] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[13] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[14] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016.