Master 2 Internship Proposal: Multimodal Vision-Language Pretraining

Advisors: Emmanuelle Salin, Stephane Ayache, Benoit Favre

October 13, 2022

1 Context

Vision-Language Pretrained models such as UNITER [2] have been developed in recent years to learn to extract visual, textual and multimodal information from text and images.

They are based on the transformer architecture and pretrained on a large corpus of text-image data using vision, language and multimodal tasks, which are called "pretraining" tasks. The pretrained models can then be finetuned and applied to multiple tasks such as visual question answering [1] and image captioning [5].

Figure 1: Visual Question Answering dataset [3] examples

2 Problem Statement

The goal of this internship is to study the multimodal pretraining of transformer-based Vision-Language models. Most models are pretrained using an image-text matching task, while some also have an additional multimodal task such as word-region alignment, like UNITER.

However, multiple studies have shown the weaknesses of state-of-the-art language models have. For example, some multimodal concepts which are less represented in the data are harder to extract. In addition, fine-grained multimodal dependencies are hard to understand, especially for vision-language models trained on large noisy datasets with basic multimodal pretraining.

More specifically, the goal is to study self-supervised multimodal pretraining tasks and their impact on the ability of a model to extract multimodal information. Several new pretraining tasks can be explored:

- Attention-Guided Masked Multimodal Modeling, based on [4]: Language and vision pretraining tasks consist in masking at random part of the input and asking the model to reproduce the masked input. With attention-guided masking, the idea is to mask the most relevant parts of the image or text input using attention. The goal is to build a more efficient masking technique so that the model relies on both modalities to extract multimodal information.

- Grouped Image-Text Matching: The idea is to combine several images in one image input and several captions in a single text input to do a grouped image-text matching. Instead of deciding whether an image corresponds to a caption, the model evaluates how each parts of an image correspond to each part of the text. The goal is to prompt the model to look for finegrained multimodal dependencies.

3 Profile

The intern will propose, implement and analyse multimodal pretraining tasks for Vision-Language models. The work will be implemented using 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 be a six-month internship at at LIS/CNRS in Marseille during spring 2023. It will be held in the context of Emmanuelle Salin's thesis on understanding the generality of multimodal representations.

4 Contact

Please send a CV and letter of application to benoit.favre@lis-lab.fr, emmanuelle.salin@lis-lab.fr, and stephane.ayache@lis-lab.fr. Do not hesitate to contact us if you have any question.

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. 2019.

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022.

[5] 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.