Master 2 Internship Proposal: Using interpretability methods to explain Vision-Language models for medical applications

Advisors: Emmanuelle Salin, Stephane Ayache, Benoit Favre

October 4, 2021

1 Context

Recent advances in Deep Learning have led to the growth of interpretable Machine Learning, which seeks to help understand the decisions of a model. Indeed, in various fields such as Medicine, Finance and Security, it is important for models to be trustworthy and reliable. As part of this internship, we want to develop a method to explain the decisions of medical Vision-Language models.

State-of-the-Art Vision-Language models such as UNITER [3] are built based on the transformer architecture to extract representations from texts and images. Those representations are then used in multimodal applications such as visual question answering [1] and image captioning [7]. However, due to the complex architecture of those models, explaining them remains a challenge. Applying interpretability methods to those models can be a way to make them more reliable. In the context of medical data, we want to be able to explain why a radiology report does or does not fit a X-ray. To this end, we will rely on medical datasets such as MIMIC-CXR [4]. This internship is focused on the development of interpretability methods for Vision-Language models.

2 Problem Statement

The goal of this internship is to explain transformer-based Vision-Language models such as UNITER. Current explainability methods for transformer-based models mostly rely on attention weights. However, studies show that attention weights by themselves are a limited tool for transformer model interpretability [2], and additional tools are necessary to explain model predictions.

We decide to focus on model-agnostic methods for this internship, as they don't use model internals such as attention weights. We will study how local model-agnostic interpretability methods such as LIME [6] can explain Vision-Language models by attributing the model decision to parts of the input. In particular, the intern will focus on explaining the predictions of the model on the Image-Text Matching task. The goal is to explain why the model predicts that a image-caption pair is matching or not, using text tokens and image superpixels. The interpretability method should help :

- Highlight matching textual and visual information such as objects
- Show if concepts such as color, number, position and size are understood by the model at a multimodal level
- Establish how the image and text contradict each other if they do not match
- Determine the importance of the language and vision modalities in the model prediction
- Study how dataset bias, and in particular textual bias, impacts the model prediction
- Study how the model reacts to perturbations (e.g textual descriptions that are similar yet distinct from the visual information)
- Show if simple logical operations (or, and ...) are understood by the model

To that end, the intern will use a dataset based on Clevr [5] to evaluate the interpretability method on true and adversarial examples. The work will first be evaluated on a carefully designed synthetic dataset before being tested on real world data such as chest X-rays and their reports.

3 Profile

The intern will propose, implement and analyse interpretability methods 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 2022. It will be held in the context of Emmanuelle Salin's thesis on understanding the generality of multimodal representations. Pointers on Interpretable Machine Learning are available[1].

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 before the 05/11/21. 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] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*, 2019.

[3] 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.

[4] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[5] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[1] https://christophm.github.io/interpretable-ml-book/

[7] 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.