Self-Supervised Learning in Vision Transformers

1. General Info

Project Title: Self-Supervised Learning in Transformers

Contact Person: Yousef Yeganeh, Azade Farshad

Contact Email: y.yeganeh@tum.de, azade.farshad@tum.de

2. Project Abstract

The goal of this project is to examine techniques of self-supervised learning in Transformers.[1] We specifically aim to use the physics of OCT imaging modality to design the most efficient model for this modality.[2] The most recent advancements in the implementation of transformers in computer vision (ViT)[3] challenges state-of-the-art CNN architectures. We try to get familiarized with different techniques and use and modify the most appropriate one or propose a novel approach. We use either PyTorch or Tensorflow Federated as our platform.

3. Background and Motivation

Fully attention modules that later adopted the name of Transformer have been state-of-the-art in sequential data, like language. One of the most well-known models of this type was BERT that was introduced by Google.[4]
There have been some attempts to use attention modules in combination with CNN layers in vision tasks, but last November, Google, inspired by BERT, introduced a fully attention-based architecture that beat state-of-the-art CNN architectures.
The shortcomings of ViT were massive computational and data requirements, however, since then many research papers were published to make the training of such networks more efficient for even small computational power. We usually have two steps of training in Transformers: pre-training, which is done by a self-supervised learning approach in a general dataset and a fine-tuning step in a specific dataset.
It would be exciting to find the best approach and even architecture to be used in different tasks for OCT dataset.

4. Technical Prerequisites

- Good background in statistics
- Good background in machine learning, deep learning
- Good skills in Python
- Good skills in PyTorch

5. Benefits:

- Possible novelty of the research
- Possible publication

6. Students' Tasks Description

Students' tasks would be the following:

Groups 1 & 2:

- Understanding the underlying methods
- Learning the tasks at hand and investigate the best techniques and architectures for them
- Modification and execution of the chosen techniques
- Running the evaluation metrics on the dataset and providing ablation studies

7. Work-packages and Time-plan:

	Description	#Students	From	To
WP1	Familiarize with the literature.	4	22.04	29.04
WP2	Familiarize with the required frameworks. Come up with a detailed time-plan (gantt)	4	29.04	06.05
WP3	Familiarize with clinical data, data pre-processing	4	06.05	13.05
WP4	Implementing and adapting the techniques for the tasks	4	13.05	27.05
WP5	Evaluation of the implemented method	4	27.05	03.06
WP6	Modification and customization of techniques for OCT dataset	4	3.06	10.06
M1	Intermediate Presentation II	4	10.06.2021	
WP7	Modification and customization for the tasks	4	10.06	17.06
WP8	Implement and Evaluation	4	17.06	01.06
WP9	Testing and Documentation	4	01.07	15.07
M2	Final Presentation	4	15.07.2021	

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.
2. Tran, A., Weiss, J., Albarqouni, S., Roohi, S.F. and Navab, N., 2020, October. Retinal Layer Segmentation Reformulated as OCT Language Processing. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 694-703). Springer, Cham.
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
4. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.