Investigating the benefits of combining CNNs and Transformer architecture for rail domain perception task

Vatsal Harshadbhai Bambhania

August 24, 2023

Based on the transformer used for natural language processing [2], the Vision Transformer (ViT) [1] has transferred the self-attention mechenism for processing the image data in the visual domain. ViT has out-performed the state-of-the-art pure convolutional models in the classification task. However, this comes at the cost of requiring an even larger dataset as it lacks "inductive bias" that is present intrinsically in convolutional nerual networks (CNNs).

This thesis aims to explore various approaches to combine the long-range dependency mapping ability of transformer with inductive bias of the CNN. In the architecture proposed by Xie Y. et al. [3], the visual features extracted by CNN blocks are used by subsequent transformers-based segmentation of medical data. We want to perform a similar segmentation task on the real-world dataset called "RailSem19" which includes the image of rail tracks, platforms, tram tracks, other vehicles and trains, people, vegetation, sky, obstacles, etc. The primary aim of this thesis is to investigate the improvement achieved by pairing CNN with transforms. The goal of this thesis is defined as follow:

1. Training efficiency and training time of the model.
2. The overall improvement in the model's performance.
3. Improvement in model complexity.
4. Improvement in terms of interpretability/explainability.
5. Out-of-distribution and uncertainty performance.

The thesis involves the following key steps:

- Creating the model architecture based on papers [3] and [4] which incorporates similar approach of combining the CNN and transformers. The training will be done using RailSem19 dataset.
- Optimize the model architecture for our use case.

- Create the pipeline for interpretation of trained model.
- Evaluate the results and compare them with the pure CNN benchmark model.
- Analyse its out-of-distribution performance.

This thesis aims to better understand the potential of pairing CNN and transformer and explore the efficient techniques of combining them.

References

[1] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[2] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[3] Yutong Xie et al. "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer. 2021, pp. 171–180.

[4] Ning Zhang et al. "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 18537–18546.