Vision Transformers

Konstantin A. Maslov
k.a.maslov@utwente.nl

University of Twente
Faculty of Geo-Information Science and Earth Observation

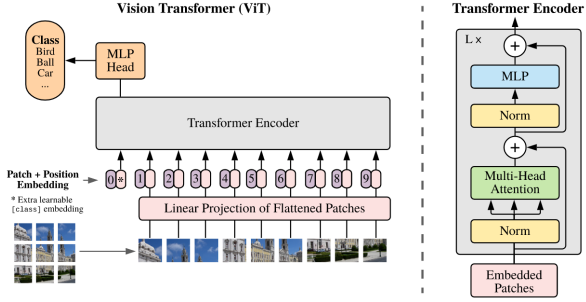
29 Oct 2022

Outline

Layer normalization

Multi-head self attention

ViT



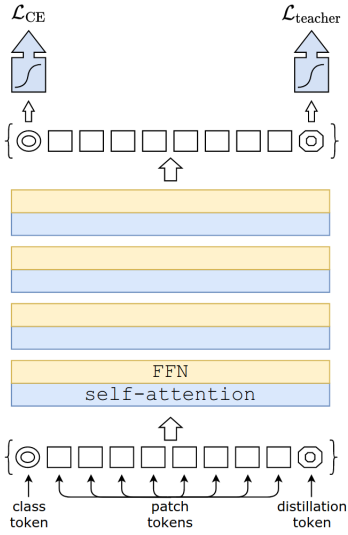
- Transformer design from NLP with minimal changes
- Achieves performance close to the state-of-the-art (CNNs) in classification tasks
- Can learn long-range relations in the very first layers due to the multi-head self-attention (which has quadratic time complexity)
- Requires pre-training on huge datasets to achieve good performance

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
https://doi.org/10.48550/arxiv.2010.11929

ViT DEMO

https://github.com/konstantin-a-maslov/transformers-seminar

DeiT



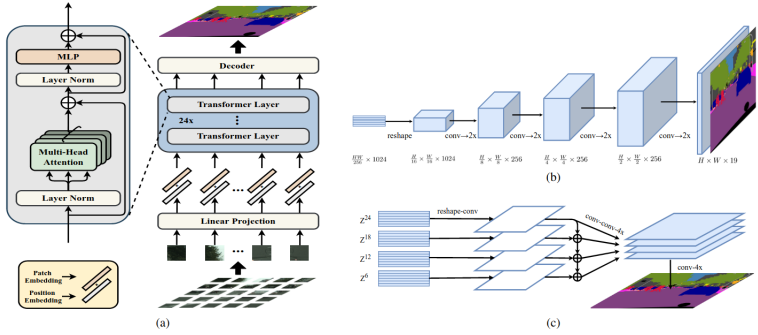
- ViT trained with distillation
- The teacher model is a CNN
- The authors claim that it reduces the amount of data required to train a transformer

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., & Ai, F. (2020). Training data-efficient image transformers & distillation through attention. https://doi.org/10.48550/arxiv.2012.12877

DPT

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision Transformers for Dense Prediction. https://doi.org/10.48550/arXiv.2103.13413

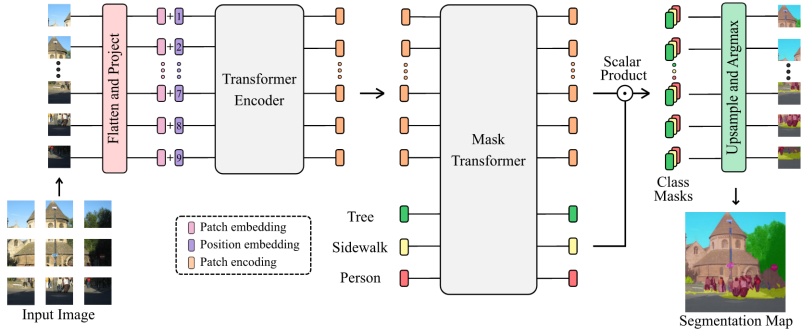
SETR



- ViT with a typical upsampling decoder as in FCNs
- The model has shown state-of-the-art performance in some tasks

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., & Zhang, L. (2020). Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 6877–6886. https://doi.org/10.48550/arxiv.2012.15840

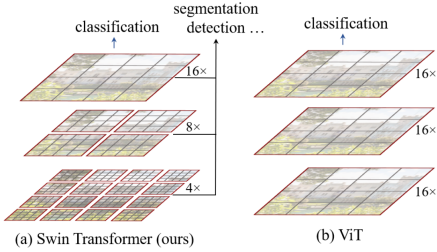
Segmenter

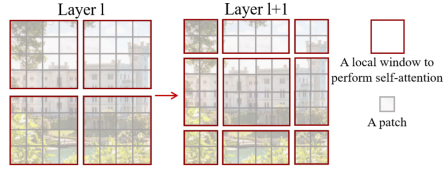


- Transformer-based decoder
- Introduced class embeddings
- The authors emphasized that transformer-based models are not so good at generating sharp object boundaries
- Does not seem to be a popular choice nowadays

Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021). Segmenter: Transformer for Semantic Segmentation. Proceedings of the IEEE International Conference on Computer Vision, 7242–7252. https://doi.org/10.48550/arxiv.2105.05633

Swin transformer

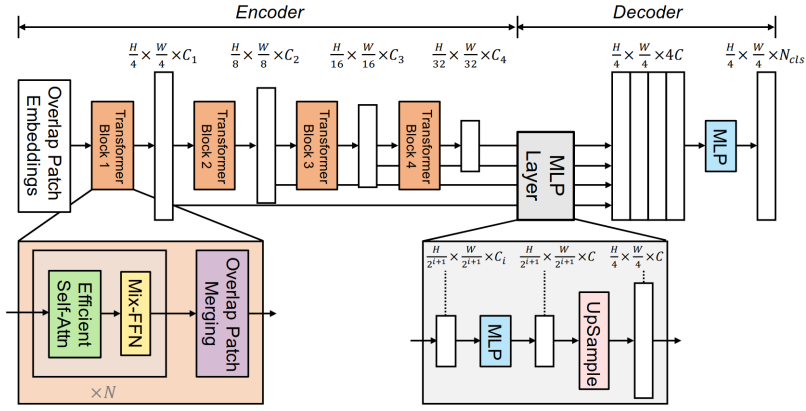




- Has linear time complexity due to the hierarchical design
- Introduced shifting windows

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE International Conference on Computer Vision, 9992–10002. https://doi.org/10.48550/arxiv.2103.14030

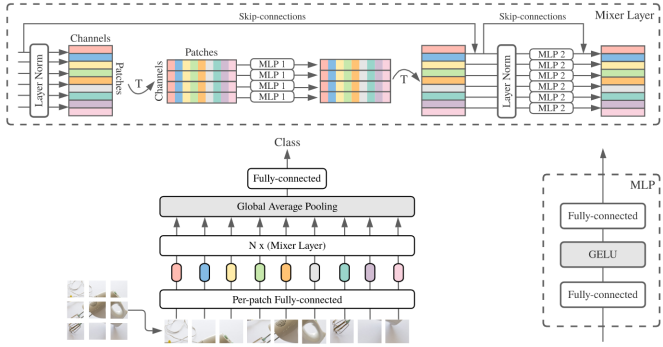
SegFormer



- The authors focused on an efficient design
- Seems to be the state-of-the-art among transformers for semantic image segmentation (general tasks) nowadays

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. Advances in Neural Information Processing Systems, 15, 12077–12090. https://doi.org/10.48550/arxiv.2105.15203

MLP-Mixer



- Not a transformer!
- Replaces multi-head self-attentions with simple MLPs
- The authors have shown that it still possible to obtain good results with this design

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., & Dosovitskiy, A. (2021). MLP-Mixer: An all-MLP Architecture for Vision. Advances in Neural Information Processing Systems, 29, 24261–24272. https://doi.org/10.48550/arxiv.2105.01601

Common practices

- 1
- 2
- 3

Summary

- Transformers require a lot of data to train
  - Which can be not our case
  - Pre-training is complicated due to the absence of huge datasets for multispectral/SAR/DEM data
  - Using the weights from more common datasets (ImageNet, JFT, ...) is still an option though, but it requires studies on how to better 'generalise' them for non-RGB images
- Seems like transformers are bad at restoring sharp boundaries in segmentation maps
  - Can be crucial as the spatial resolution of the imagery we use is very different from the resolution of the images from ImageNet or similar datasets
  - Smaller patch sizes improve the situation (if one has enough memory...)
  - Perhaps, there is a space to explore hybrid CNN-Transformer models
- However, as it will be shown later, a simple SETR-like transformer can still generate interesting results