Transformers in Computer Vision

Marc Evrard, Camille Guinaudeau, François Yvon

LISN — CNRS and Université Paris-Saclay

Transformers in Computer Vision

Humane-AI

Convolutional Inductive Biases[1]

Convolutional models

- have dominated the field of Computer Vision for years
- provide suitable inductive biases when extracting features from images

But...

- ... convolutional inductive biases lack a global understanding of the image
- ... large receptive fields are required in order to track long-range dependencies within an image

[1] Inspired from "Transformers in Computer Vision: Farewell Convolutions!" by Victor Perez

Self-attention layers

Self-attention used in early stages of a model can learn to behave similarly to a convolution

Self-attention layers take a feature map as input

- compute attention weights between every pair of features
- –> updated feature map where each position has information about any other feature within the same image
- can replace or be combined with convolutions

Self-attention layers

Basic approach

- flattening spatial dimensions of input feature map –> sequence of features with shape *HW* x *F*
  HW = flattened spatial dimensions, F = feature map's depth
- self attention used directly over the sequence to obtain updated representations

Computation cost can be expensive for high resolution input

- **Wang et al. [2020]:** computes attention along the two spatial axis sequentially instead of dealing directly with the whole image
- **Ramachandran et al. [2019]:** process patches of feature maps instead of the whole spatial dimensions

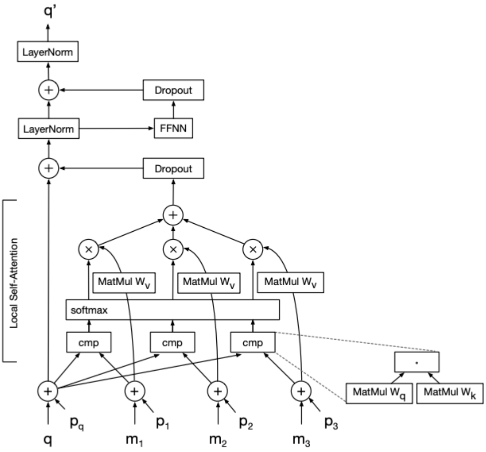
Vision Transformers

Image Transformer - Parmar et al. [2018]

- Task: image generation
- New pixel generated by taking into account previously known pixel values within the image
- Feature generation: self-attention takes into account a flattened patch of m features as context and produces a representation for the unknown pixel value
- RGB value converted into a tensor of d dimensions using 1D convolutions and the m features of the context patch are flattened to be 1 dimensional.

Vision Transformers

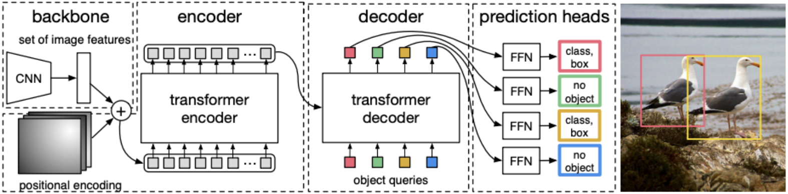
Image Transformer - Parmar et al. [2018]



Vision Transformers

DEtection TRansformer - Carion et al. [2020]

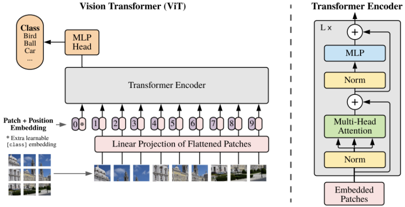
- Task: Object detection
- Visual features extracted from a convolutional backbone.
- Feature maps are flattened over their spatial dimensions (h x w x d) –> (hw x d)
- Learnable positional encoding added to each dimension



Vision Transformers

Vision Transformer (ViT) - Dosovitskiy et al. [2020]

- Task: Image recognition
- Input sequence: flattened vector of pixel values extracted
- Flattened vector fed to a linear projection layer –> patch embeddings
- Learnable positional embedding added to each embeddings
- Learnable embedding attached to the beginning of the sequence



Bibliography I

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European Conference on Computer Vision, pages 213–229. Springer, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In International Conference on Machine Learning, pages 4055–4064. PMLR, 2018.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909, 2019.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In European Conference on Computer Vision, pages 108–126. Springer, 2020.