Beyond ImageNet: recent advances

Attention process and Vision Transformers
Attention process in ConvNets

In ConvNets what information is shared between pixels (or features) in one block? => 2D spatial locality (typically 3x3) => attention is done locally

Attention is done locally => less local after many layers
Global (Self) attention

How to build a deep architecture with \textit{local global} attention inside? Meaning that one patch may interact with all others!

$\Rightarrow$ Different than conv!
Let’s see what they do in Natural Language Processing (NLP):

Attention between words in Machine translation process:

Computing of weights
Use them to compute feat.
Attention process in Natural Language Processing (NLP)

Attention between words

Basic models: Recurrent neural Nets (RNNs) as encoders => classification

Classify

[English, German, Swiss German, Gaelic, Dutch, Afrikaans, Luxembourgish, Limburgish, other]
Attention process in Natural Language Processing (NLP)

Attention between words

Basic models: Recurrent neural Nets (RNNs) as decoders => text generation
Attention process in Natural Language Processing (NLP)

Attention between words/language

Basic models: Cross-attention for language translation: Seq2Seq -- RNNs2RNNs
Perhaps an even better idea is to compute the average h vector across all steps and pass this to the decoder at each time step in the decoder but using a weighted average with learned weights, and the weights are specific for each time step!!!

\[ \bar{h}_j = \sum a_{j,i} h_i \]

such that:

\[ a_{j,i} = \frac{\exp(h_j v_{j-1})}{\sum \exp(h_i v_{i-1})} \]
Attention process in Natural Language Processing (NLP)

**Transformer** architecture: no RNNs
Attention is All you Need
Encoder/Decoder with attention

Fixed number of input tokens
[but hey! we can always define a large enough length and add mask tokens]

[Vaswani et al. Attention is all you need] [https://arxiv.org/abs/1706.03762]
Attention process in Natural Language Processing (NLP)

**Transformer** architecture:
- no RNNs
- Attention is All you Need
- Encoder/Decoder with attention

Fixed number of input tokens
[but hey! we can always define a large enough length and add mask tokens]
Fixed number of input tokens
[but hey! we can always define a large enough length and add mask tokens]
Attention process in Natural Language Processing (NLP)

**Transformer architecture:**
- no RNNs
- Attention is All you Need
- Encoder/Decoder with attention
Attention process in Vision

Is it possible to mimic this attention-based architecture for vision processing?
ViT for vision image Transformer (credit slide to Jeamin Jong)

Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable “classification token” to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).
ViT - Input Section

\[ \mathbf{x} \in \mathbb{R}^{H \times W \times C} \]

\[ \mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)} \]

\[ N = HW / P^2 \]
ViT - Transformer Encoder

• Similar to BERT’s [class] token, we prepend a learnable embedding to the sequence of embedded patches \( z_0^0 = x_{\text{class}} \), whose state at the output of the Transformer encoder \( z_L^0 \) serves as the image representation \( y \) (Eq. 4).

• Layernorm (LN) is applied before every block, and residual connections after every block.

• The MLP contains two layers with a GELU non-linearity.

\[
\begin{align*}
z_0 &= [x_{\text{class}}; x_1^1 E; x_2^2 E; \cdots; x_N^N E] + E_{\text{pos}}, \\
z_{\ell} &= \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1}, \quad \ell = 1 \ldots L \\
z_{\ell} &= \text{MLP}(\text{LN}(z_{\ell}')) + z_{\ell}', \quad \ell = 1 \ldots L \\
y &= \text{LN}(z_L^0)
\end{align*}
\]

Hybrid Architecture: Raw image patches $\rightarrow$ Feature map of a CNN
ResMLP: Feedforward networks for image classification with data-efficient training
Perceiver IOA General Architecture for Structured Inputs & Outputs