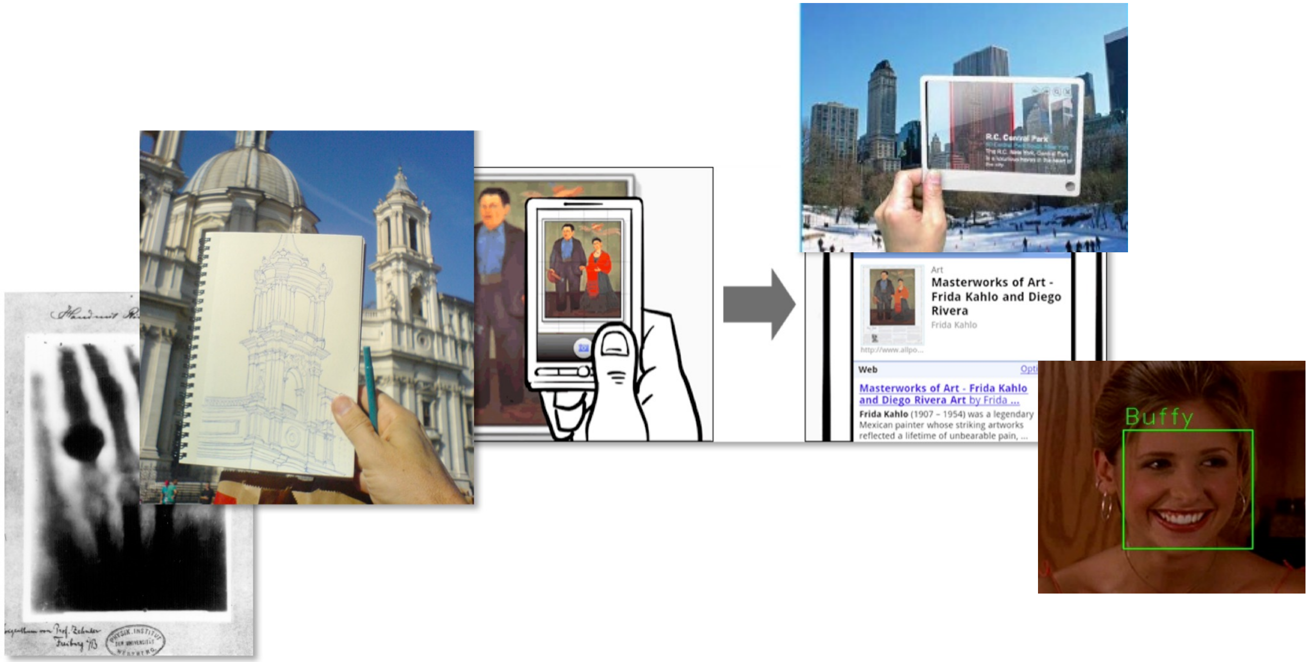


COURS Reconnaissance Visuelle par deep learning

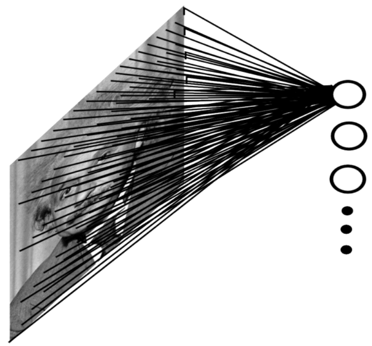
https://cord.isir.upmc.fr/teaching-multimedia/

Matthieu Cord
Sorbonne University
Computer Science - ISIR

Outline

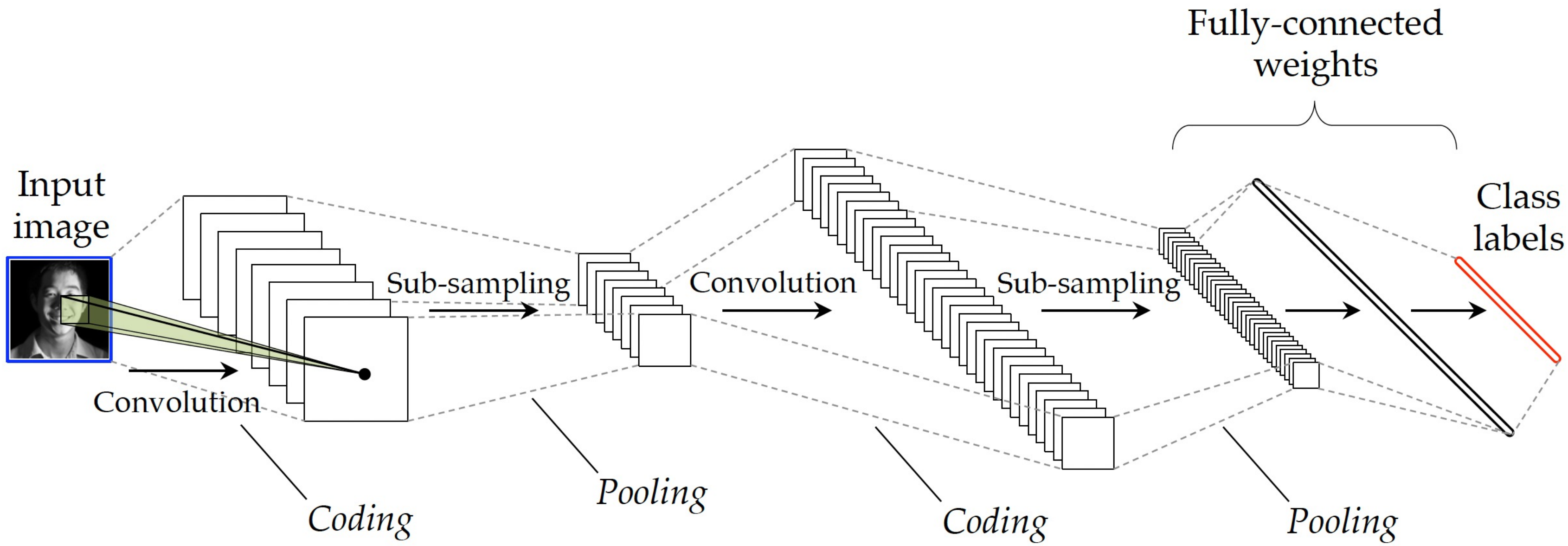
1. Attention and Vision Transformers
   - NLP: Attention is all you need

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*

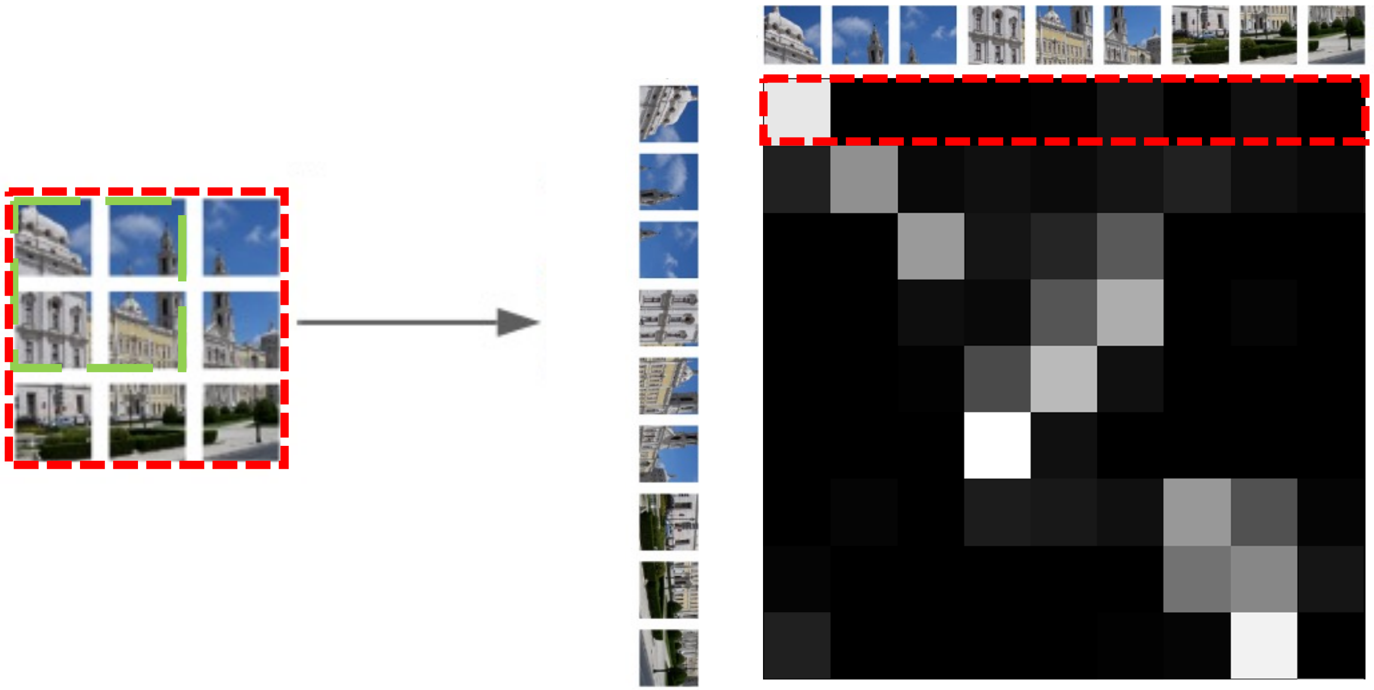
Rq: less local after many layers



Global (Self) attention

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

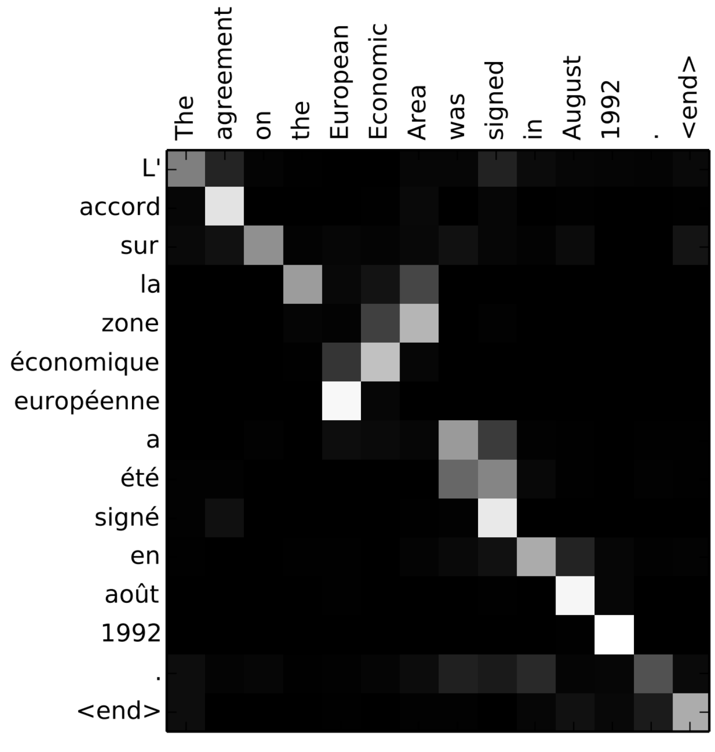
=> Different than convNet!



Let’s see what they do in Natural Language Processing (NLP):

Attention between words in Machine translation process:

1. Computing of weights
2. Use them to compute new features

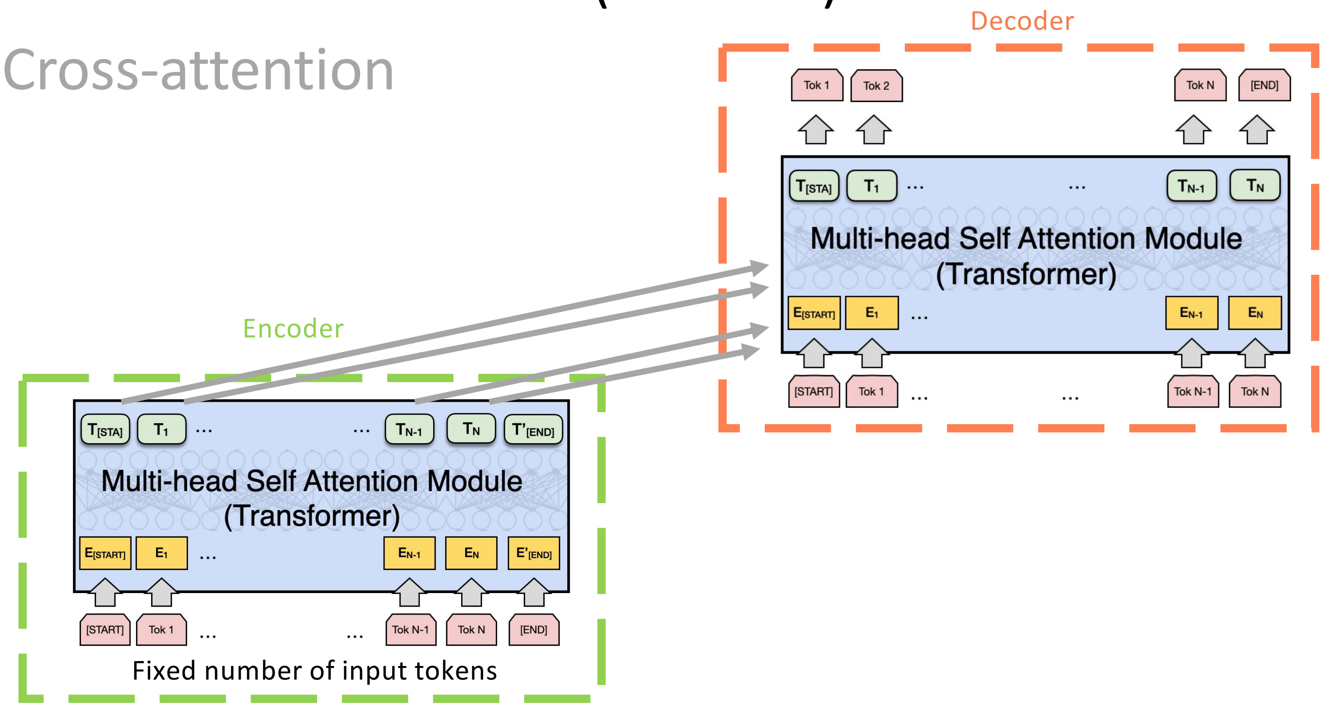


Attention process in NLP

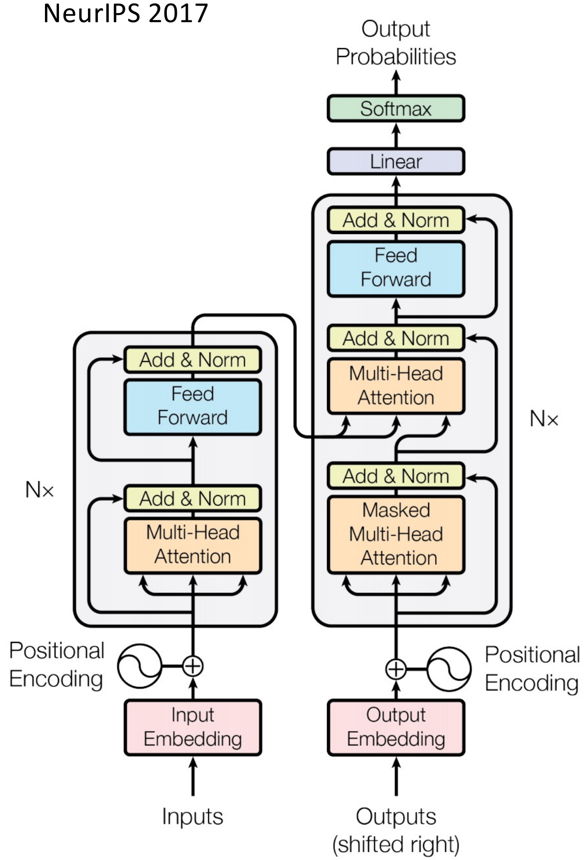
Basic language translation models: Encoder/Decoder

Transformer architecture (no RNNs)

- Cross-attention



[Vaswani et al. Attention is all you need]
https://arxiv.org/abs/1706.03762
NeurIPS 2017

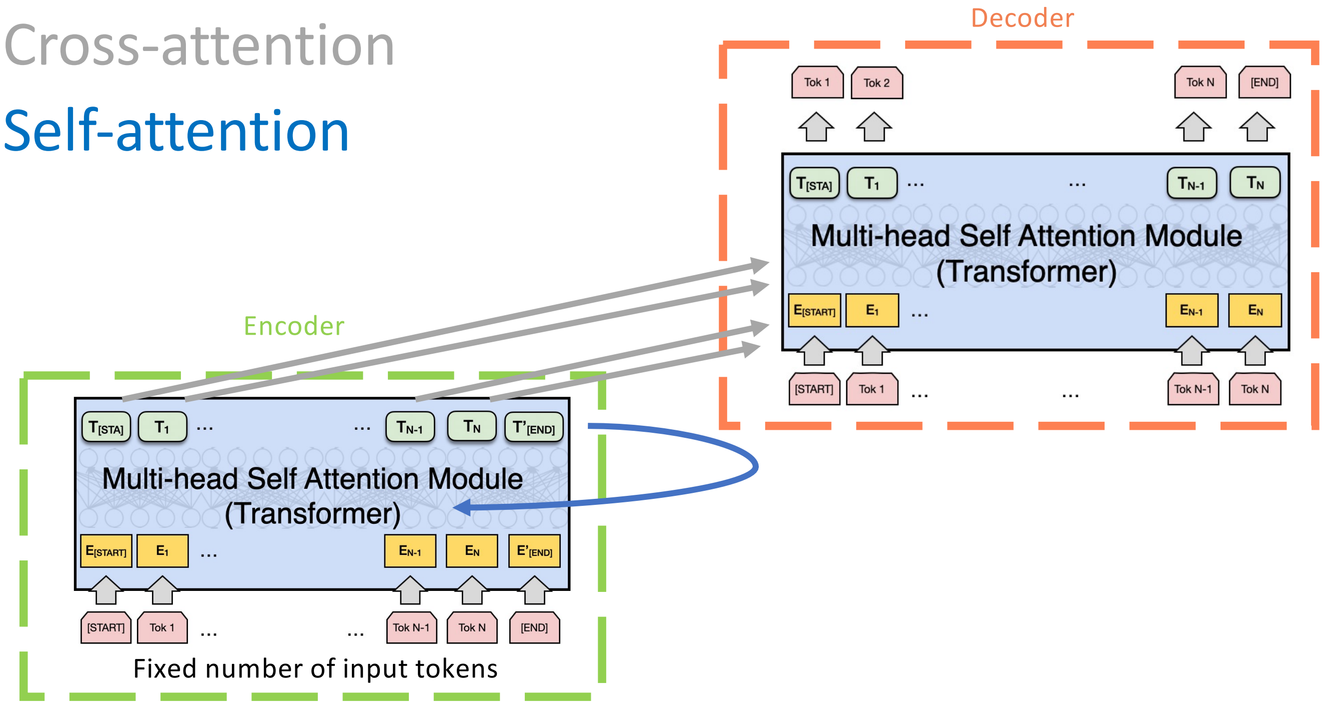


Attention process in NLP

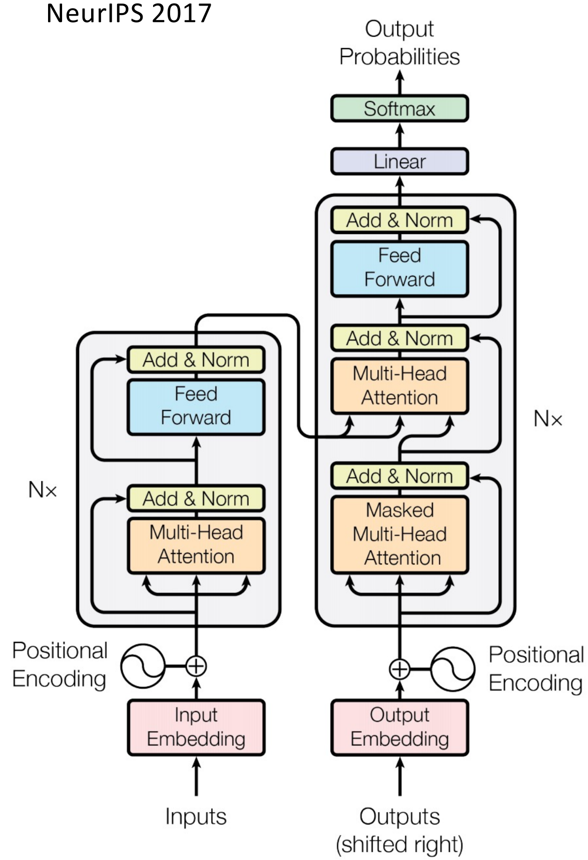
Basic language translation models: Encoder/Decoder

Transformer architecture (no RNNs)

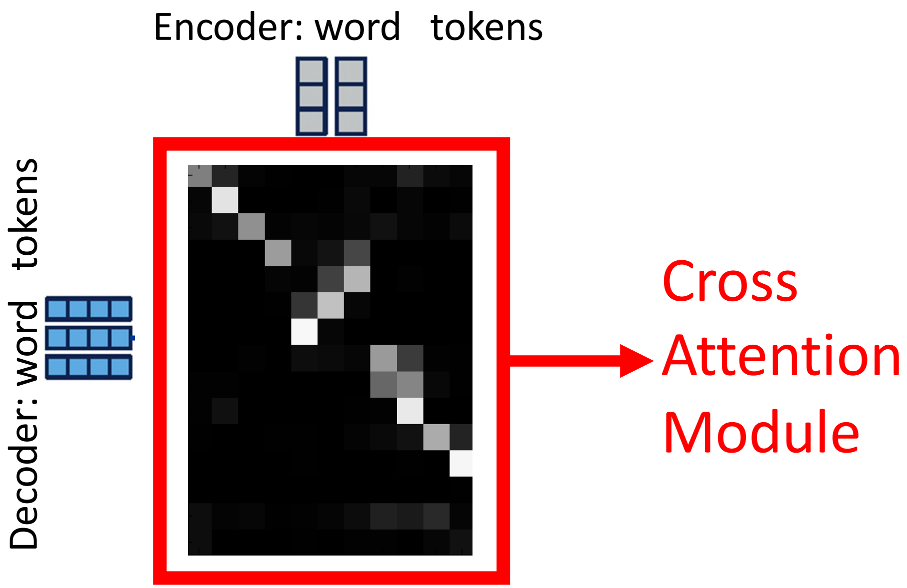
- Cross-attention
- Self-attention

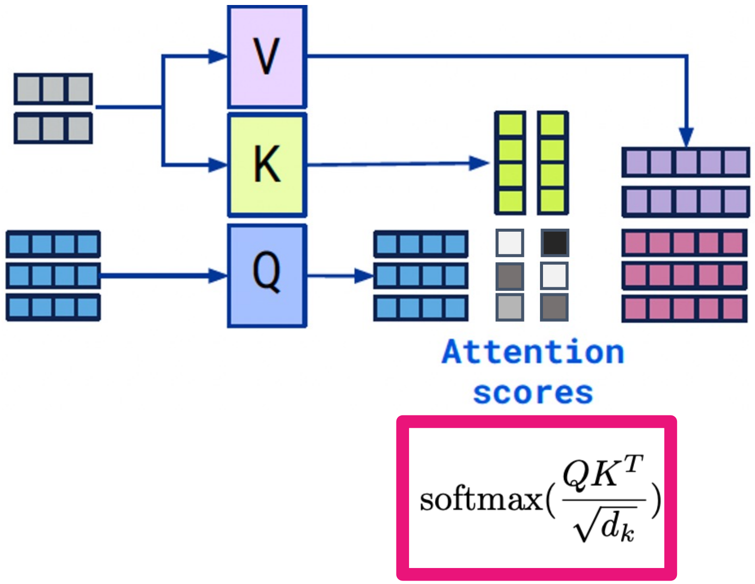


[Vaswani et al. Attention is all you need]
https://arxiv.org/abs/1706.03762
NeurIPS 2017



Attention process in NLP



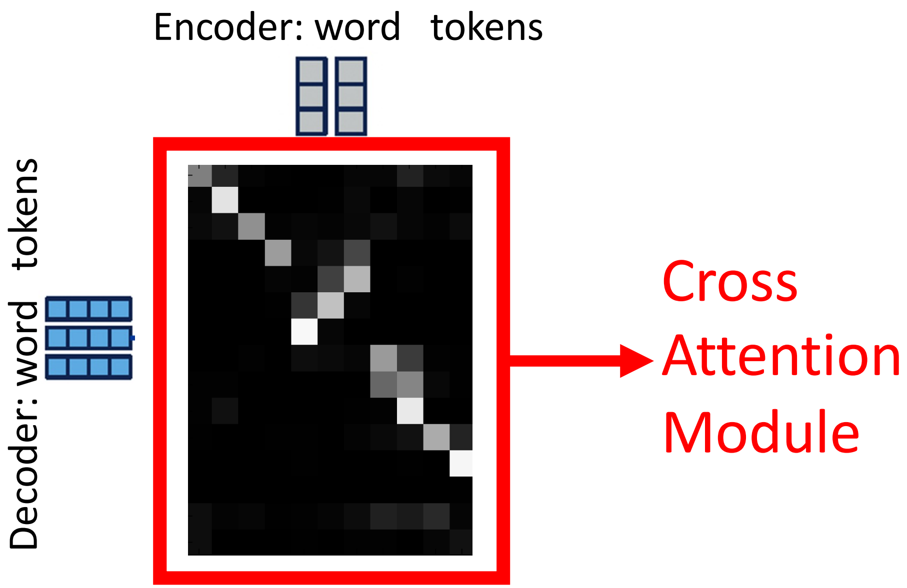


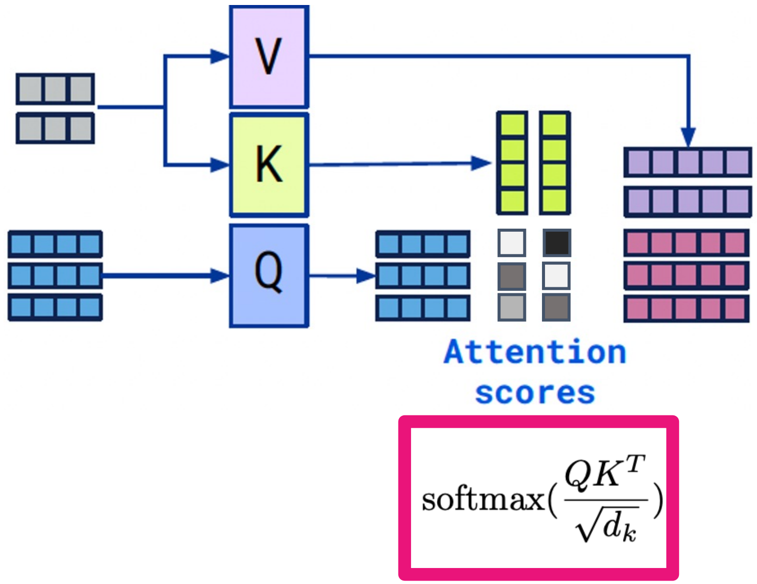
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Outline

1. Attention and Vision Transformers (ViT)
    - NLP: Attention is all you need
    - **Transformer for image classification**

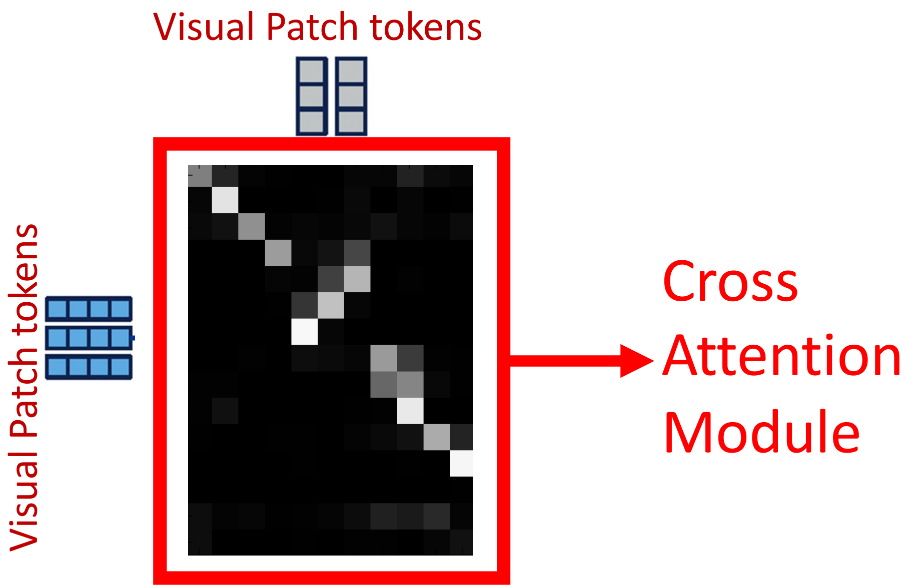
Attention process in NLP

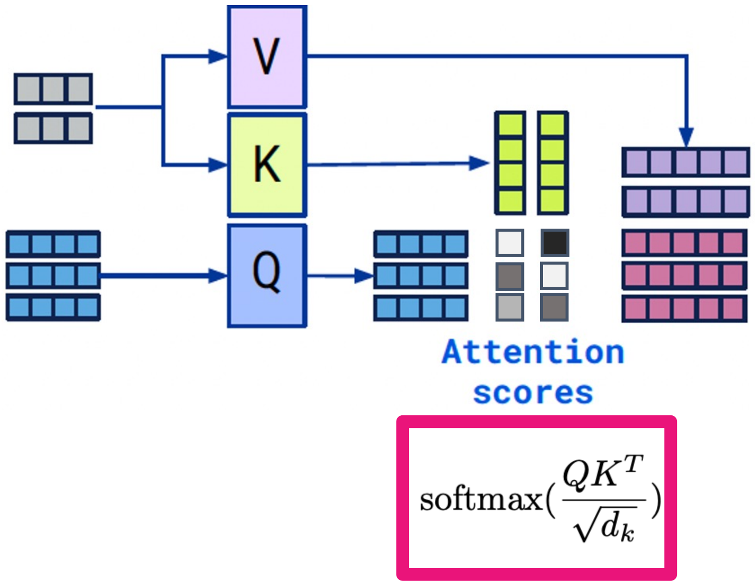




$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Attention process in Vision





$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Very similar except that Visual token is definitively less natural than word for NLP

Attention process in Vision

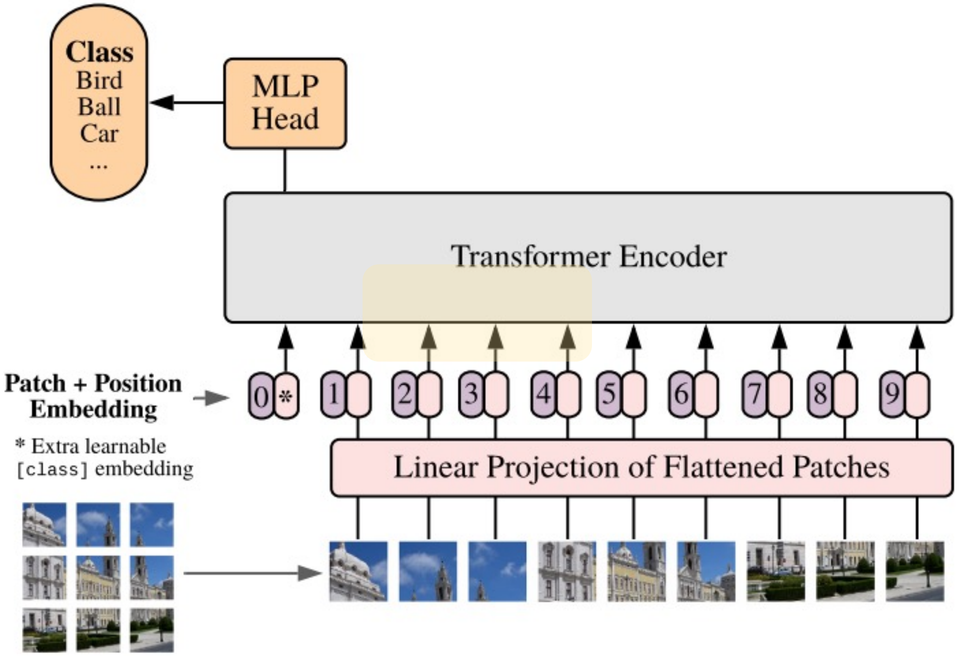
Is it possible to mimic this attention-based architecture for vision processing?

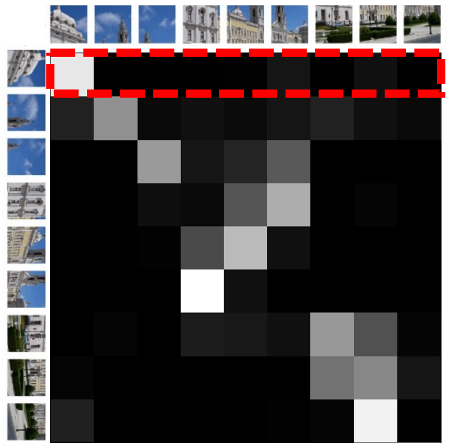
Yes! **ViT** (Vision image Transformers) architecture

Published as a conference paper at ICLR 2021

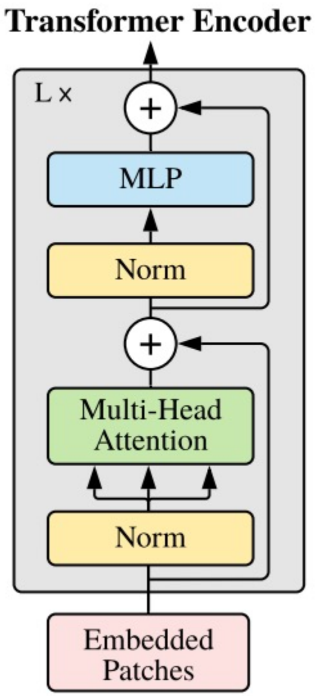
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

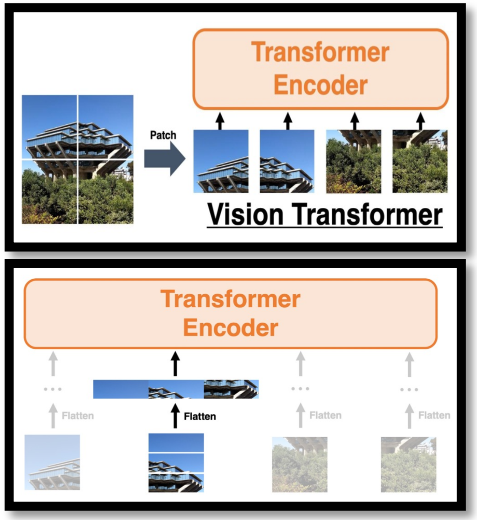
Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*],
Xiaohua Zhai[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,**
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com





Attention process in Vision





$x \in \mathbb{R}^{H \times W \times C}$

$x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$

$N = HW/P^2$

CLS token

$$\begin{aligned}
\mathbf{z}_0 &= [\mathbf{x}_{\text{class}};\, \mathbf{x}_p^1\mathbf{E};\, \mathbf{x}_p^2\mathbf{E};\cdots;\, \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, && \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C)\times D},\, \mathbf{E}_{pos} \in \mathbb{R}^{(N\boxed{+1})\times D} \\
\mathbf{z'}_\ell &= \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, && \ell = 1\ldots L \\
\mathbf{z}_\ell &= \text{MLP}(\text{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, && \ell = 1\ldots L \\
\mathbf{y} &= \text{LN}(\mathbf{z}_L^0)
\end{aligned}$$

[class=CLS] token: a learnable embedding to the sequence of embedded patches

Layernorm (LN) before every block, and residual connections after every block

MSA: Multi Head Self Attention

MLP: two layers with a GELU non-linearity

Hybrid Architecture : Raw image patches --> Feature map of a CNN