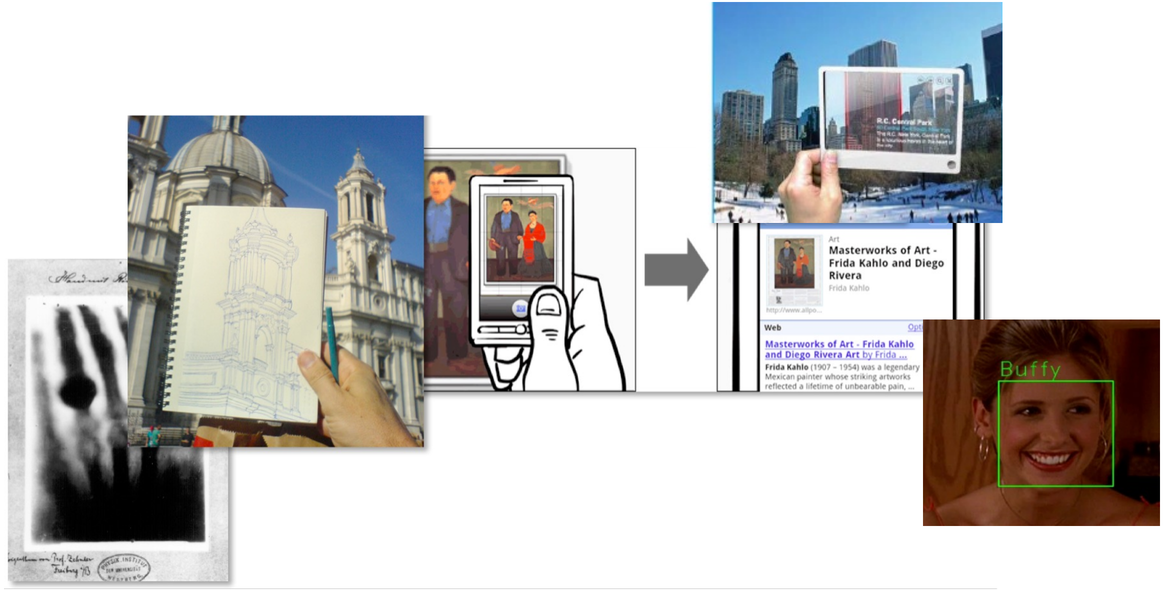


COURS Reconnaissance Visuelle par deep learning

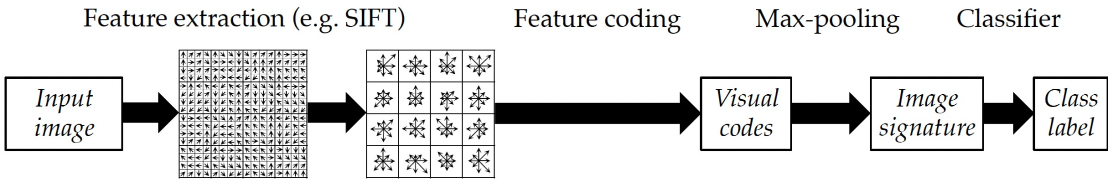
https://cord.isir.upmc.fr/teaching-multimedia/

Course Outline

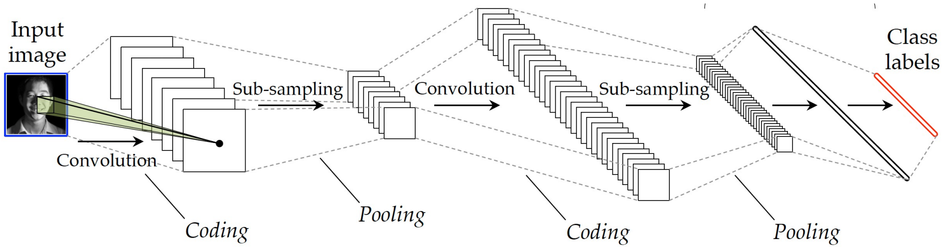
1. Intro to Computer Vision and Machine Learning
2. Intro to Neural Networks
3. Machine Learning complements
4. Neural Nets for Image Classification
   1. Recap MLP
   2. Convolutional Neural Networks
   3. Examples: LeNet5, AlexNet, GoogLeNet, VGG, ResNet
5. **Vision Transformers**
   1. **NLP: Attention is all you need**
   2. Transformer for Image Classification

Image classification: where we are and what is missing?

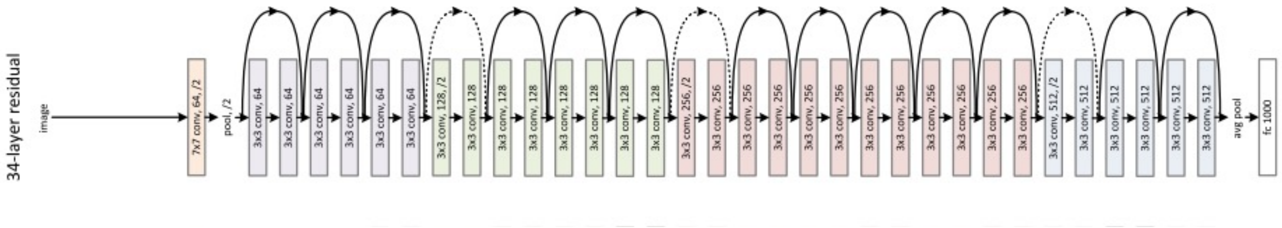
The 2000s: *BoWs + SVM*



The 2010s: Very *Large* ConvNets



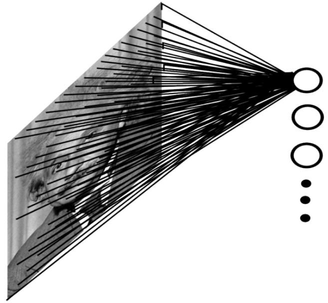
2020: The star: **ResNet**



Next step?

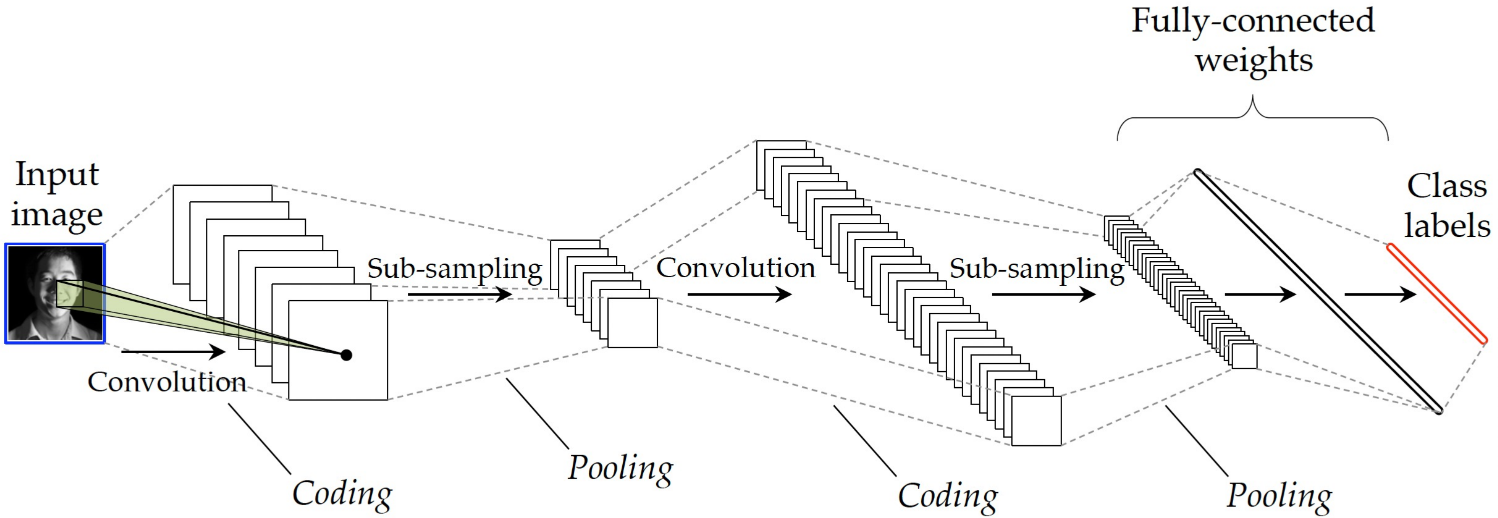
What is missing?

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*

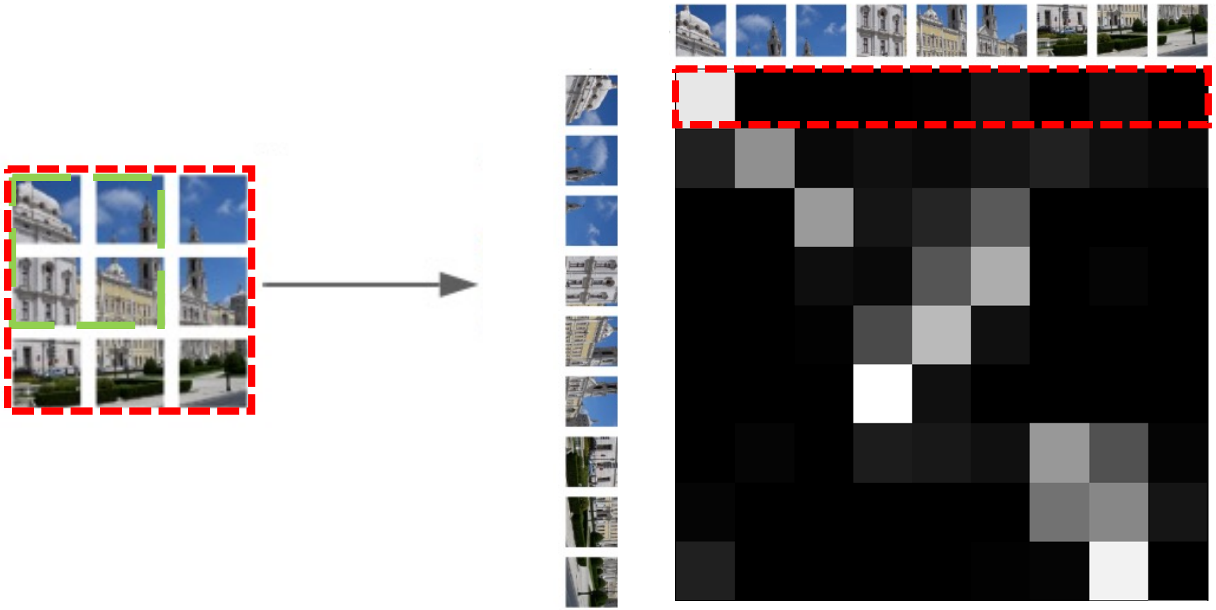
Rq: less local after many layers



Global (Self) attention

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

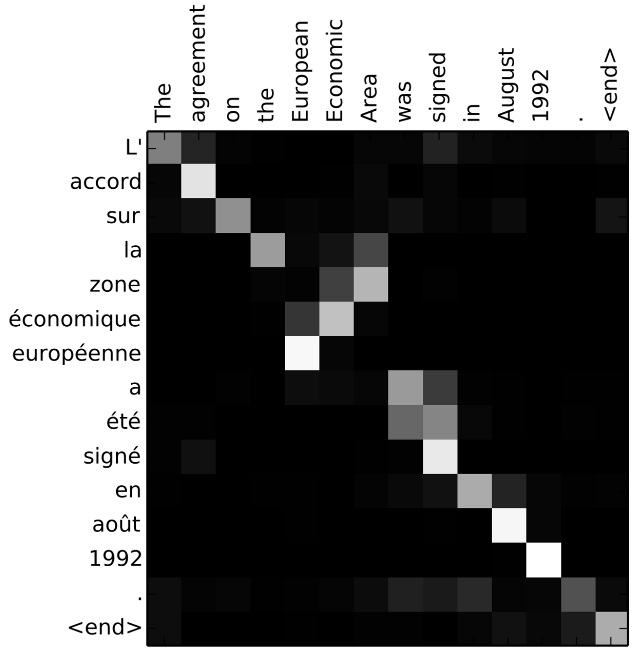
=> Different than convNet!



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

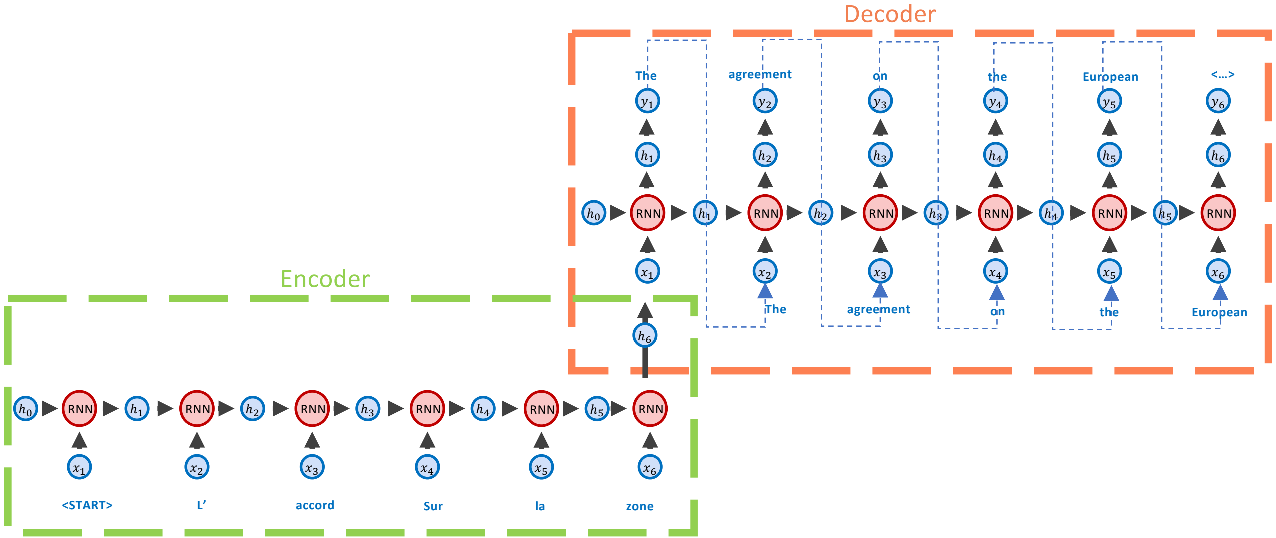


Attention process in NLP

Basic language translation models: Encoder/Decoder

Ex.: Seq2Seq -- RNNs2RNNs

Cross-attention for language translation in at the end of Encoder

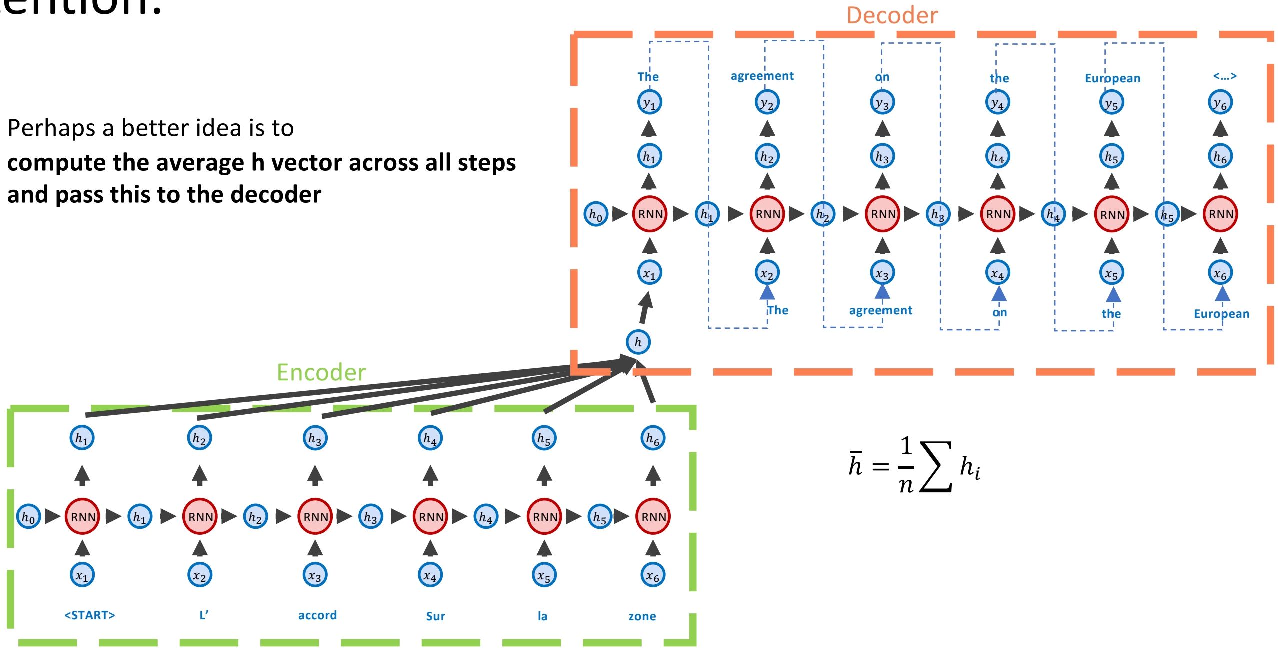


Attention process in NLP

Basic language translation models: Encoder/Decoder

Cross-attention:

Perhaps a better idea is to
compute the average h vector across all steps and pass this to the decoder



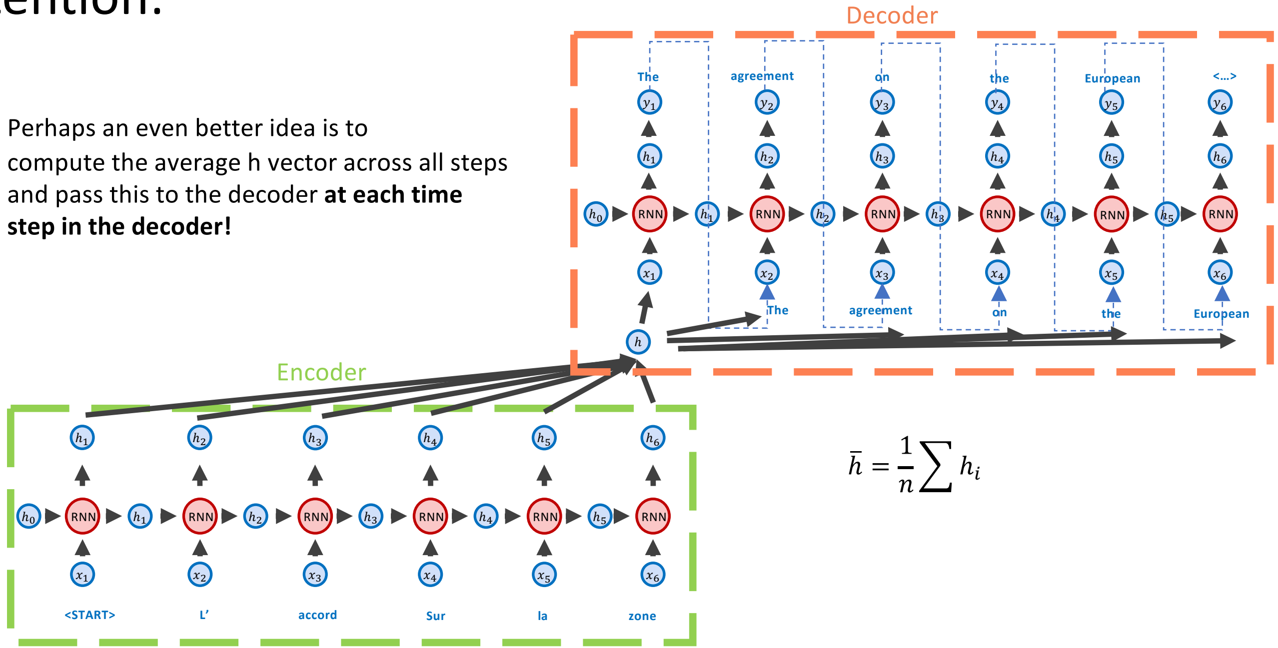
$$\bar{h} = \frac{1}{n}\sum h_i$$

Attention process in NLP

Basic language translation models: Encoder/Decoder

Cross-attention:

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!**



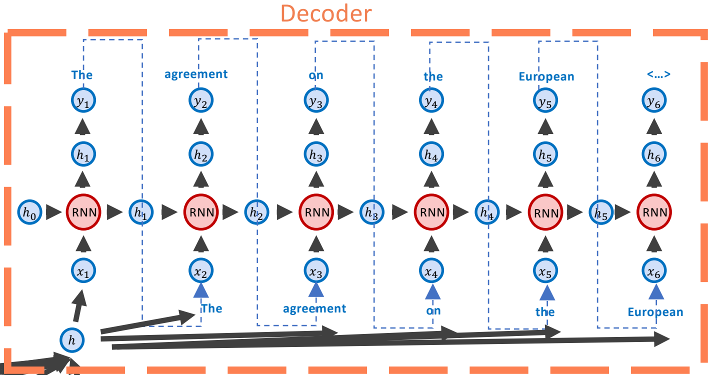
$$\bar{h} = \frac{1}{n}\sum h_i$$

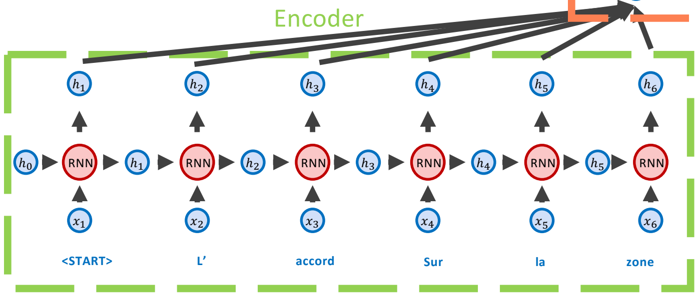
Attention process in NLP

Basic language translation models: Encoder/Decoder

Cross-attention:

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!**

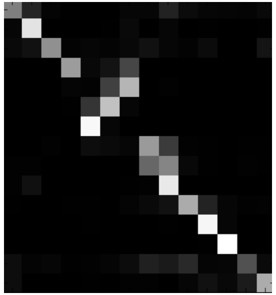




$$\bar{h}_i = \sum \alpha_{i,j} h_j \qquad \alpha_{i,j} = \frac{\exp(< h_j, s_{i-1} >)}{\sum \exp(< h_k, s_{i-1} >)}$$

Cross Attention

Encoder/ Decoder

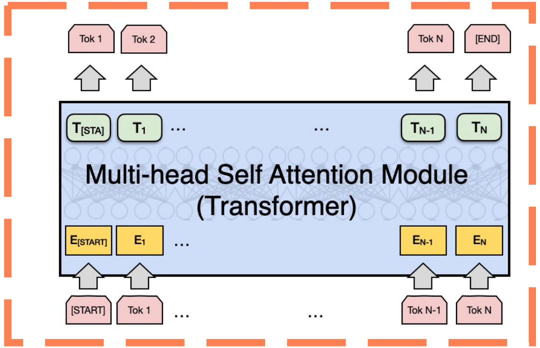


Attention process in NLP

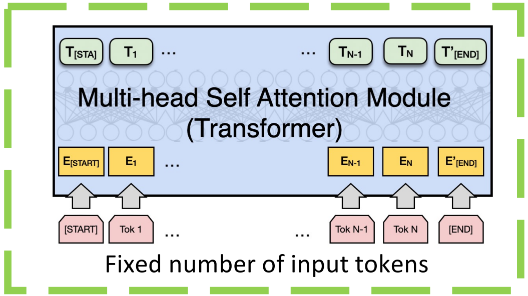
Basic language translation models: Encoder/Decoder

Transformer architecture (no RNNs)

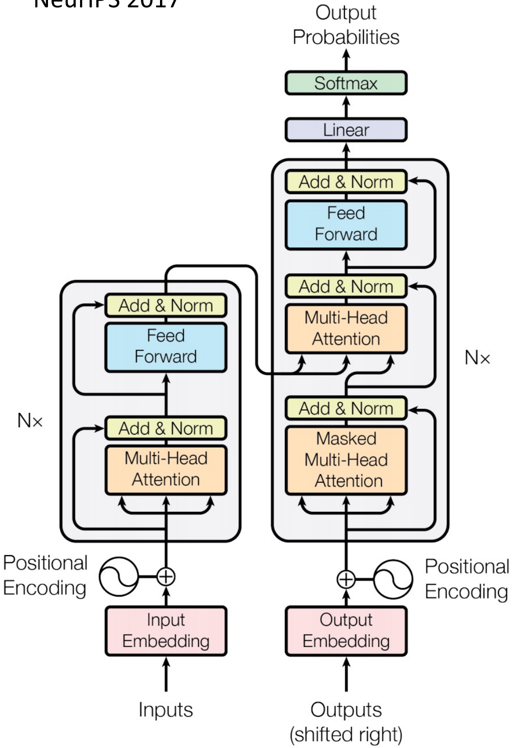
Decoder



Encoder



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

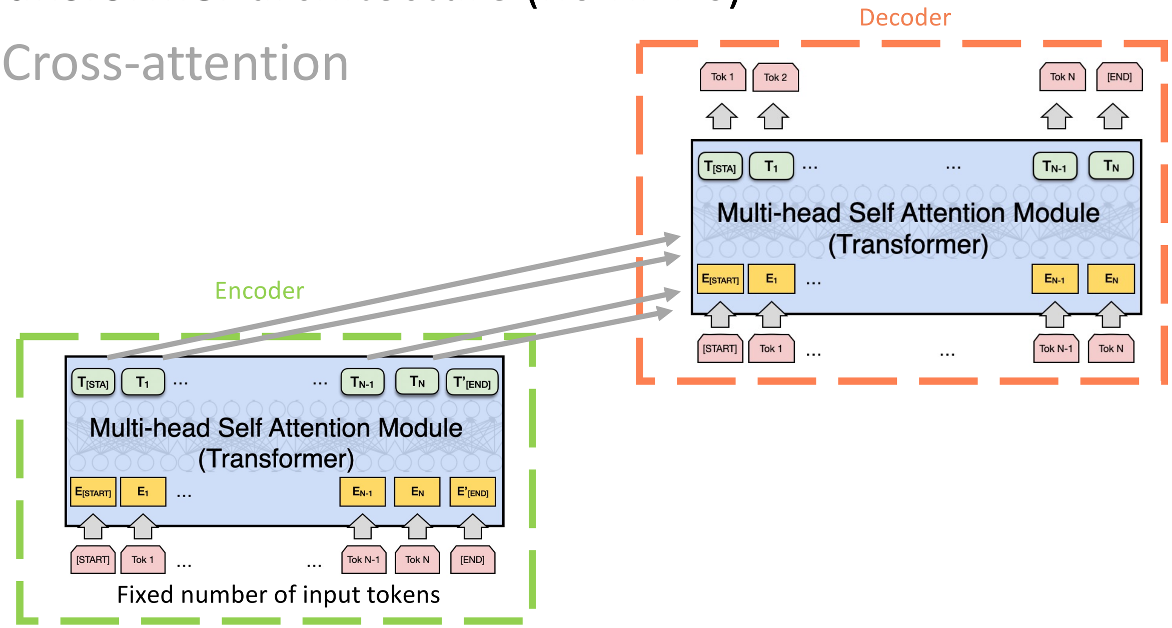


Attention process in NLP

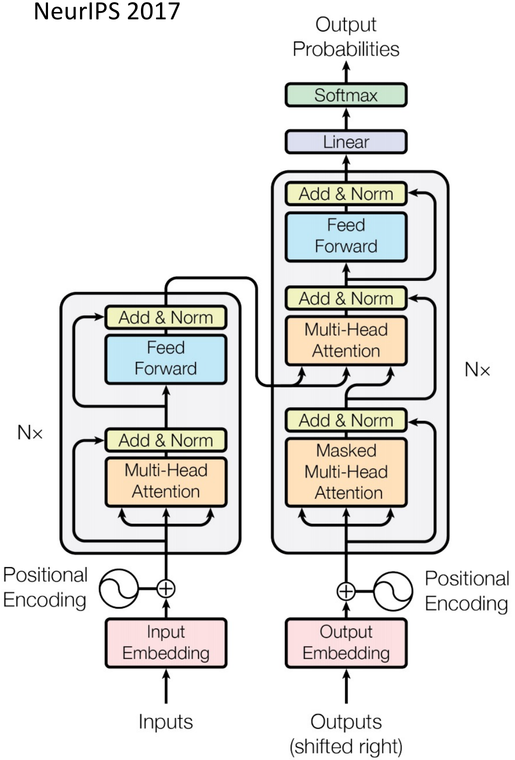
Basic language translation models: Encoder/Decoder

Transformer architecture (no RNNs)

- Cross-attention



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

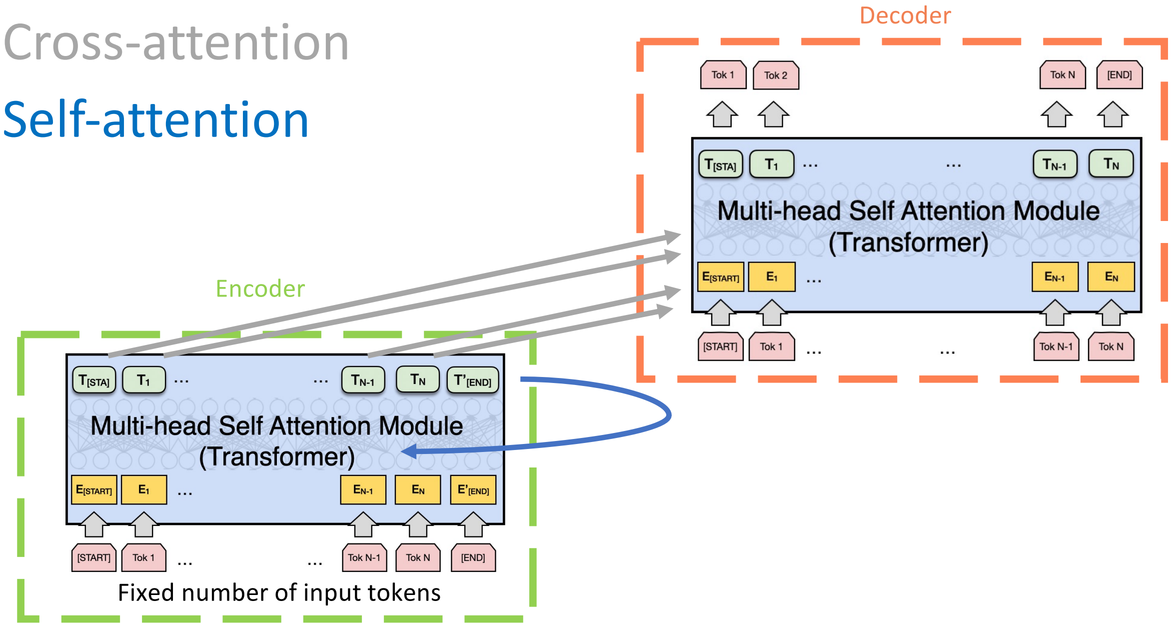


Attention process in NLP

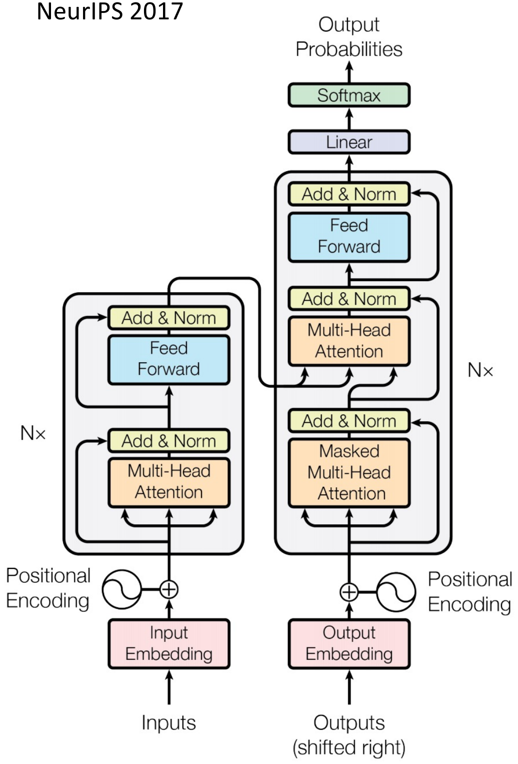
Basic language translation models: Encoder/Decoder

Transformer architecture (no RNNs)

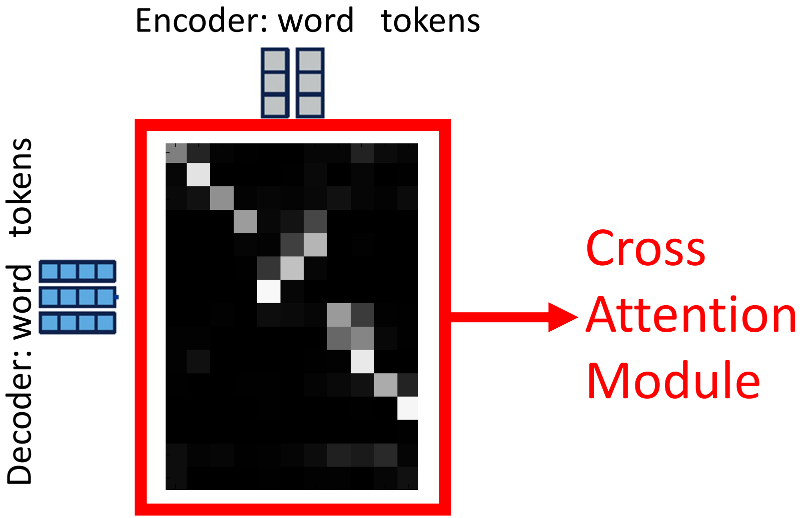
- Cross-attention
- Self-attention

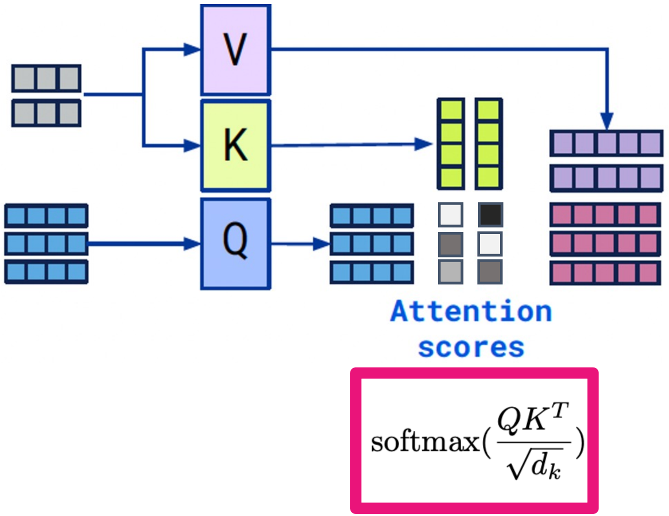


[Vaswani et al. Attention is all you need]
https://arxiv.org/abs/1706.03762
NeurIPS 2017



Attention process in NLP



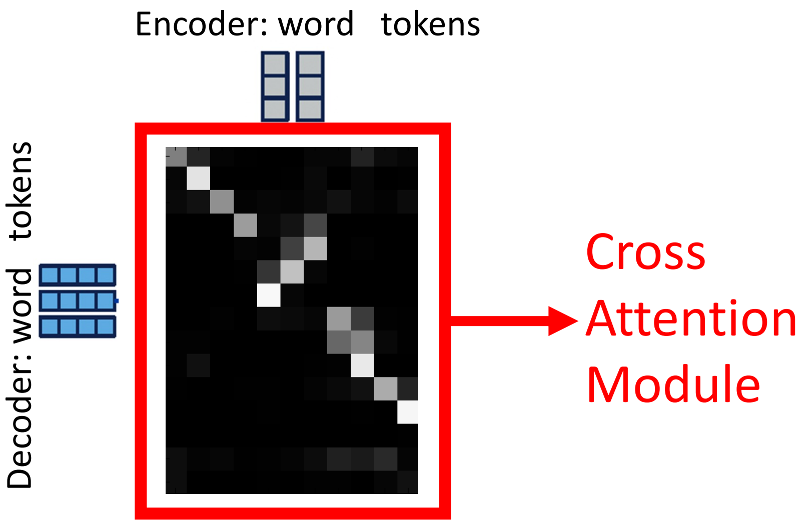


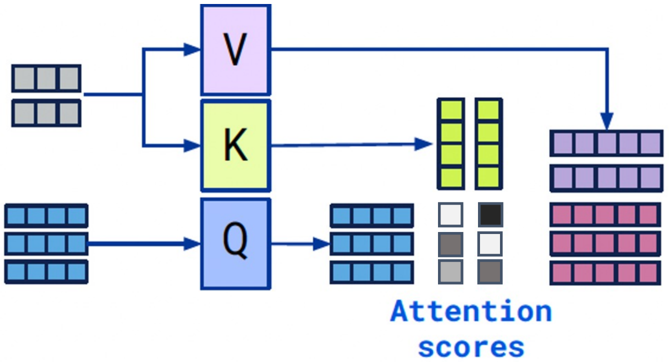
$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Course Outline

1. Intro to Computer Vision and Machine Learning
2. Intro to Neural Networks
3. Machine Learning complements
4. Neural Nets for Image Classification
   1. Recap MLP
   2. Convolutional Neural Networks
   3. Examples: LeNet5, AlexNet, GoogLeNet, VGG, ResNet
5. **Vision Transformers**
   1. NLP: Attention is all you need
   2. **Transformer for Image Classification**

Attention process in NLP

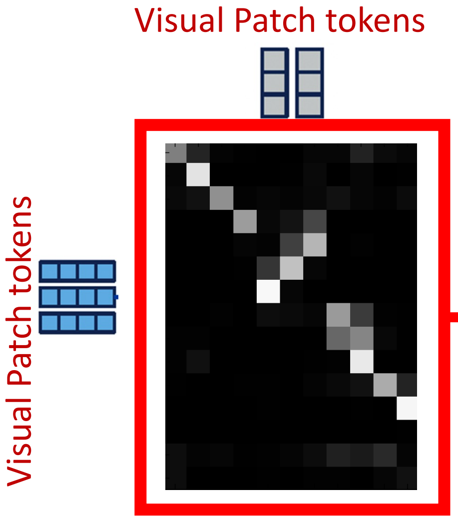




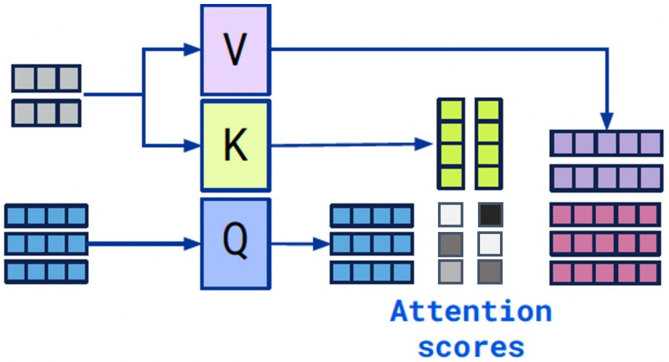
$$\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

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

Attention process in Vision



Cross Attention Module



$$\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

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

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

Attention process in Vision

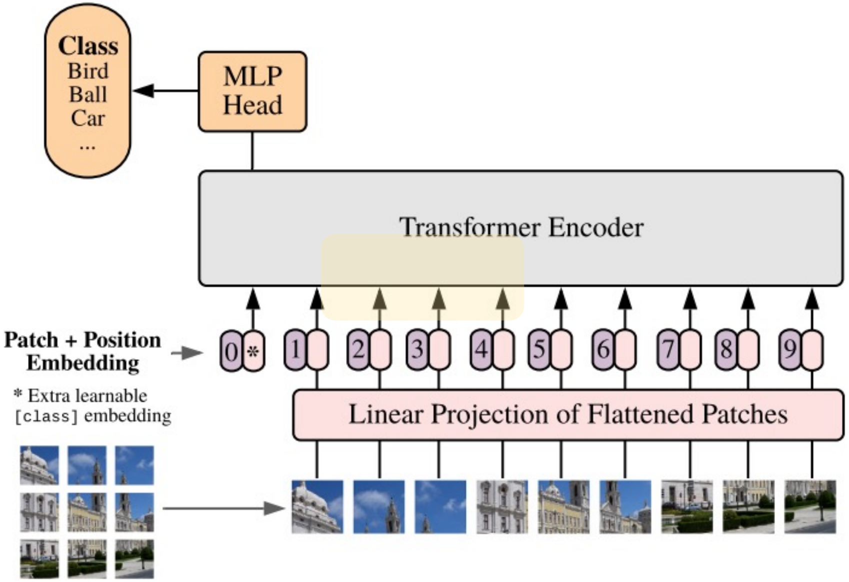
Is it possible to mimic this attention-based architecture for vision processing?

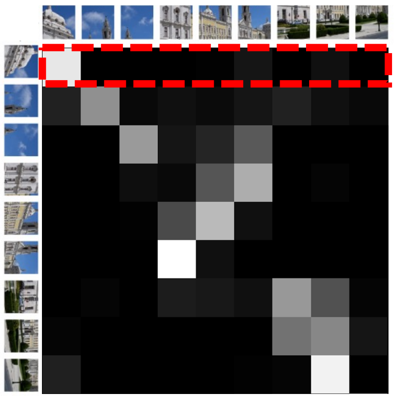
Yes! **ViT** (Vision image Transformers) architecture

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



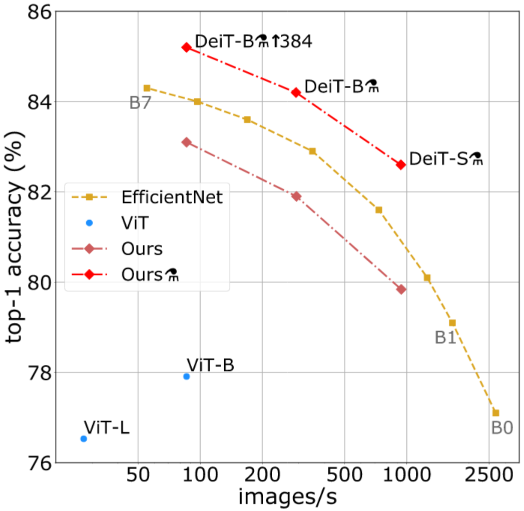


Attention process in Vision

Experiments with ViT (and variants DeiT, CaiT) transformers for image classification

State-of-the-art performance on ImageNet1k classification!

From ViT paper, **many tricks/discussions to simplify learning** in DeiT, CaiT, …



DeiT

Published as a conference paper at ICML 2021

Training data-efficient image transformers & distillation through attention

Hugo Touvron [1 2] Matthieu Cord [1 2] Matthijs Douze [1]
Francisco Massa [1] Alexandre Sablayrolles [1] Hervé Jégou [1]