

Lecture 8: Attention and Transformers

Image Captioning using Transformers

Input: Image I

Output: Sequence $y = y_1, y_2, \dots, y_T$

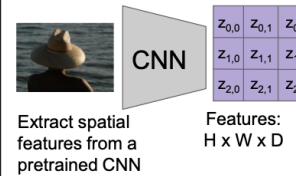


Image Captioning using Transformers

Input: Image I

Output: Sequence $y = y_1, y_2, \dots, y_T$

Encoder: $c = T_W(z)$
where z is spatial CNN features
 $T_W(\cdot)$ is the transformer encoder

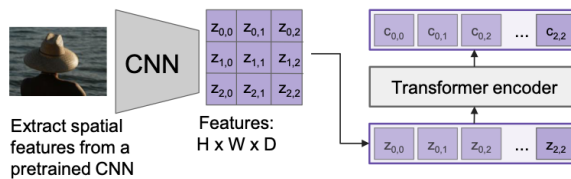


Image Captioning using Transformers

Input: Image I

Output: Sequence $y = y_1, y_2, \dots, y_T$

Decoder: $y_t = T_D(y_{0:t-1}, c)$
where $T_D(\cdot)$ is the transformer decoder

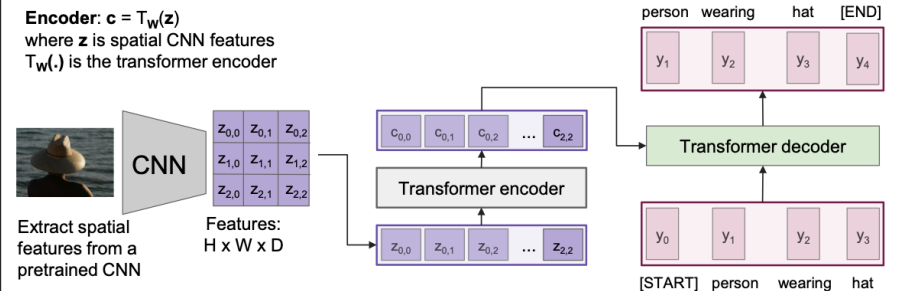
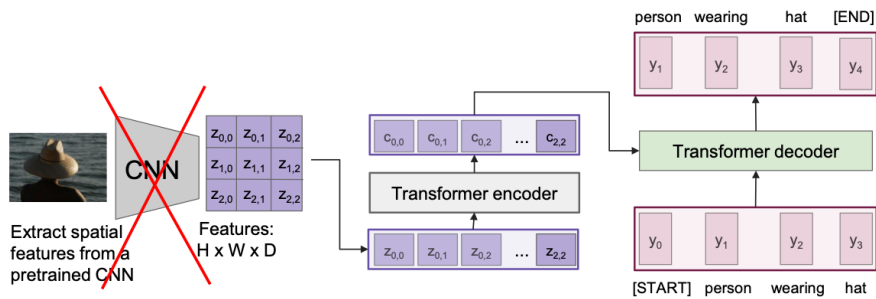


Image Captioning using transformers

- Perhaps we don't need convolutions at all?



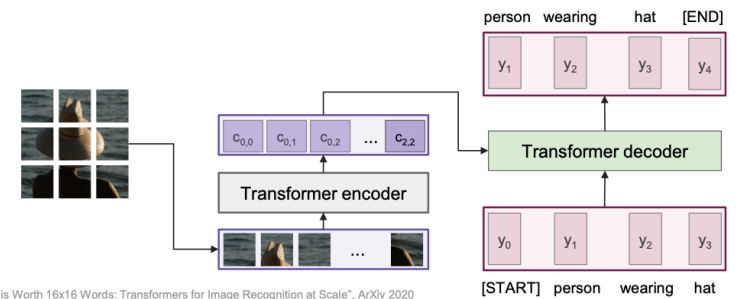
Fei-Fei Li, Ehsan Adeli, Zane Durante

Lecture 9 - 99

April 25, 2024

Image Captioning using **ONLY** transformers

- Transformers from pixels to language



Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ArXiv 2020
[Colab link](#) to an implementation of vision transformers

Fei-Fei Li, Ehsan Adeli, Zane Durante

Lecture 9 - 100

April 25, 2024

ViTs – Vision Transformers

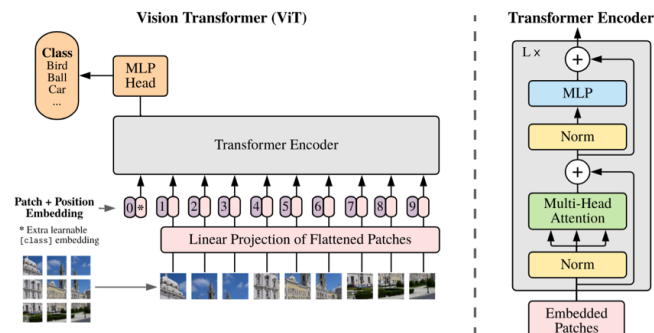


Figure from:
 Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ArXiv 2020

Fei-Fei Li, Ehsan Adeli, Zane Durante

Lecture 9 - 101

April 25, 2024

Vision Transformers vs. ResNets

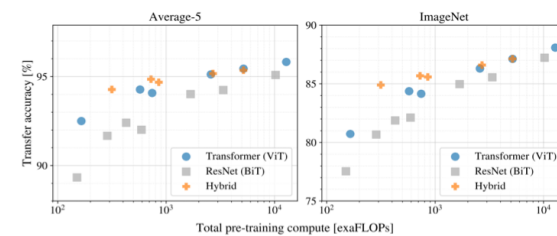


Figure 5: Performance versus cost for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

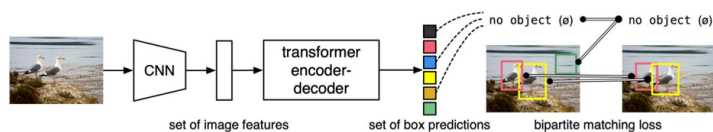
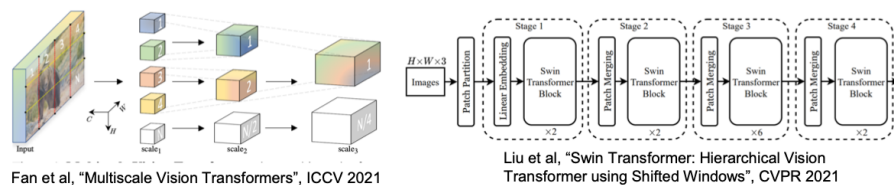
Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ArXiv 2020
[Colab link](#) to an implementation of vision transformers

Fei-Fei Li, Ehsan Adeli, Zane Durante

Lecture 9 - 102

April 25, 2024

Vision Transformers

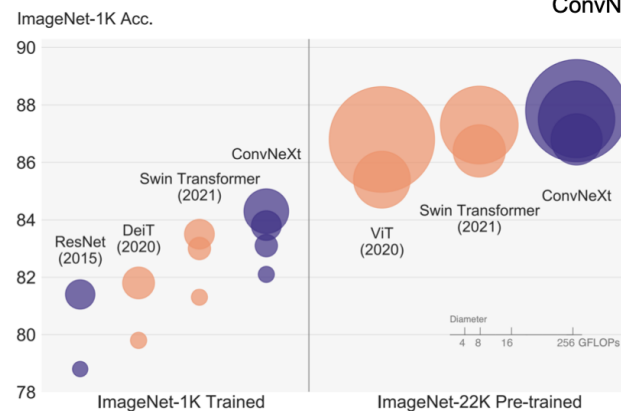


Fei-Fei Li, Ehsan Adeli, Zane Durante

Lecture 9 - 103

April 25, 2024

ConvNets strike back!



A ConvNet for the 2020s. Liu et al. CVPR 2022

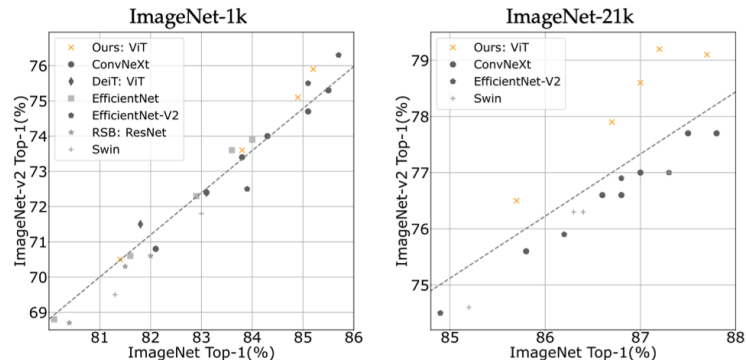
Fei-Fei Li, Ehsan Adeli, Zane Durante

Lecture 9 - 104

April 25, 2024

DeiT III: Revenge of the ViT

Hugo Touvron^{*†} Matthieu Cord[†] Hervé Jégou^{*}



Fei-Fei Li, Ehsan Adeli, Zane Durante

Lecture 9 - 105

April 25, 2024