Outline

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

- **[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 $\rightarrow$ Feature map of a CNN

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

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

$$N = HW / P^2$$
Outline

1. Attention and Vision Transformers (ViT)
   • NLP: Attention is all you need
   • Transformer Encoder ViT with Self Attention for image classification

2. Transformer Decoder for downstream tasks
Outline

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

2. **Transformer Decoder** for downstream tasks
   - Detection
   - Segmentation
   - Continual Learning, ...
General Decoder

[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]
General Decoder

[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]

Output query array / Output array defines the downstream task: detection, segmentation ...
General Decoder

[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]

Output query array / Output array defines the downstream task: detection
Transformer Decoder for detection

Just another scheme for DETR model

End-to-End Object Detection with Transformers
Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko

We present a new method that views object detection as a direct set prediction problem. Our approach streamlines the detection pipe-line designed components like a non-maximum suppression procedure or anchor generation that explicitly encode our prior knowledge. The new framework, called DETection TRansformer or DETR, is a set-based global loss that forces unique predictions via bipartite matching.
General Decoder

[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]

Output query array / Output array defines the downstream task: segmentation ...
General Decoder: or not!
General Decoder

[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]

Output query array / Output array defines the downstream task: continual learning
Video Transformer

[ViViT: A Video Vision Transformer ICCV 2021]
General Encoder / Decoder

Input array = N cameras
General Encoder / Decoder

Input array = N cameras
General Encoder / Decoder

Input array = N cameras
General Encoder / Decoder

Input array = N cameras
Output array = Bird Eye View (BEV) representation
General Encoder / Decoder

Input array = N cameras  Output array = Bird Eye View (BEV) representation
General Encoder / Decoder

Input array = N cameras
Output array = Bird Eye View (BEV) representation
Vision Transformers

**Global Attention** mechanism at every layer of the deep archi

Very **competitive architectures** in image classification with the best Convnets

**Fusion/Merging by mixing** thanks to cross attention process

**Somehow universal** deep structure around encoding/decoding for many vision tasks as classification (1 class token), object detection, segmentation, ...