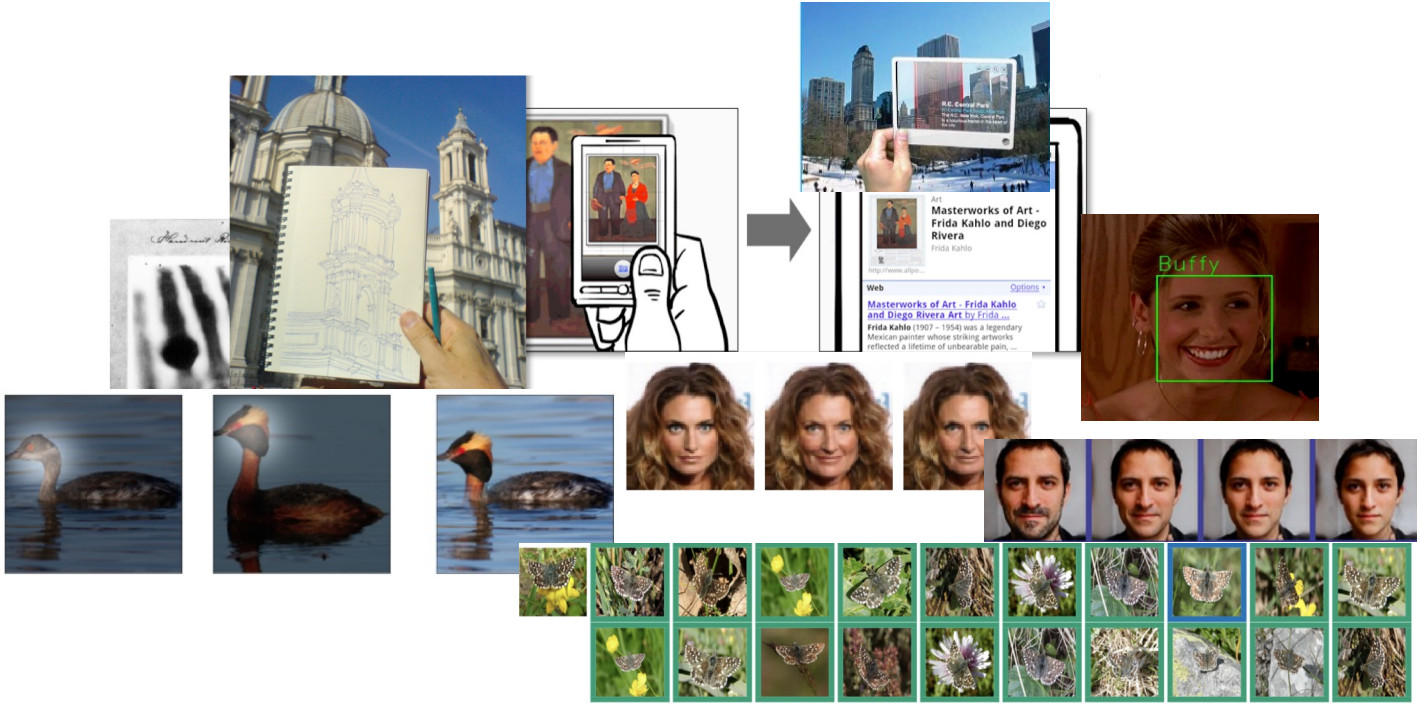


COURS RDFIA deep Image

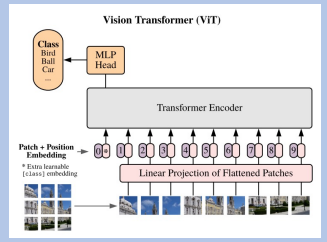
Matthieu Cord
Sorbonne University

Course Outline – Week timeline

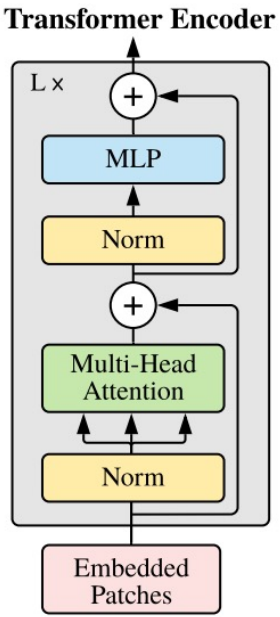
1. Computer Vision basics: Visual (local) feature detection and description, Bag of Word Image representation
2. Supervised learning: Introduction to Neural Networks (NNs)
3. Machine Learning basics: Risk, Classification, Datasets, benchmarks and evaluation, Linear classification (SVM)
4. Convolutional Nets for visual classification
5. Large deep convnets and Vision Transformers
6. Beyond ImageNet: FCNs and Segmentation
7. Transfer Learning and domain adaptation
8. Generative models with (conditional) GANs
9. (cGan) **+ Vision-Language models**
10. Control
11. Explainable AI and applications
12. Bayesian deep learning

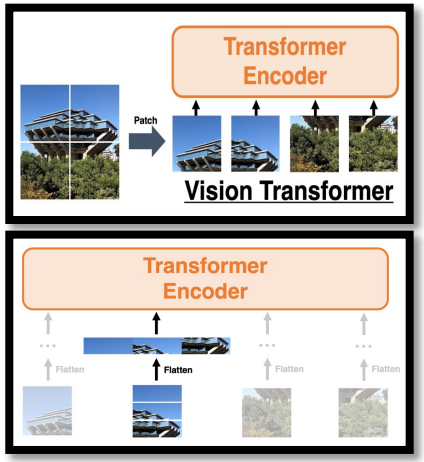
Outline

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





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

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

$$N = HW/P^2$$

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\cdots;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$$

CLS token

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0)$$

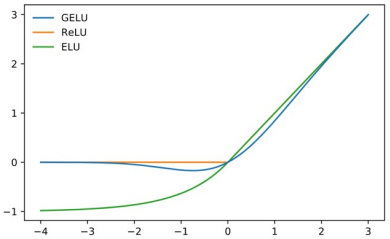
[class=CLS] token: a learnable embedding to the sequence of embedded patches

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

MSA: Multi Head Self Attention

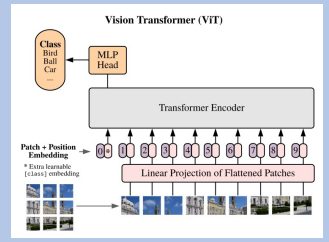
MLP: two layers with a GELU non-linearity

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



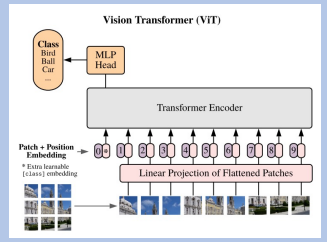
Outline

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

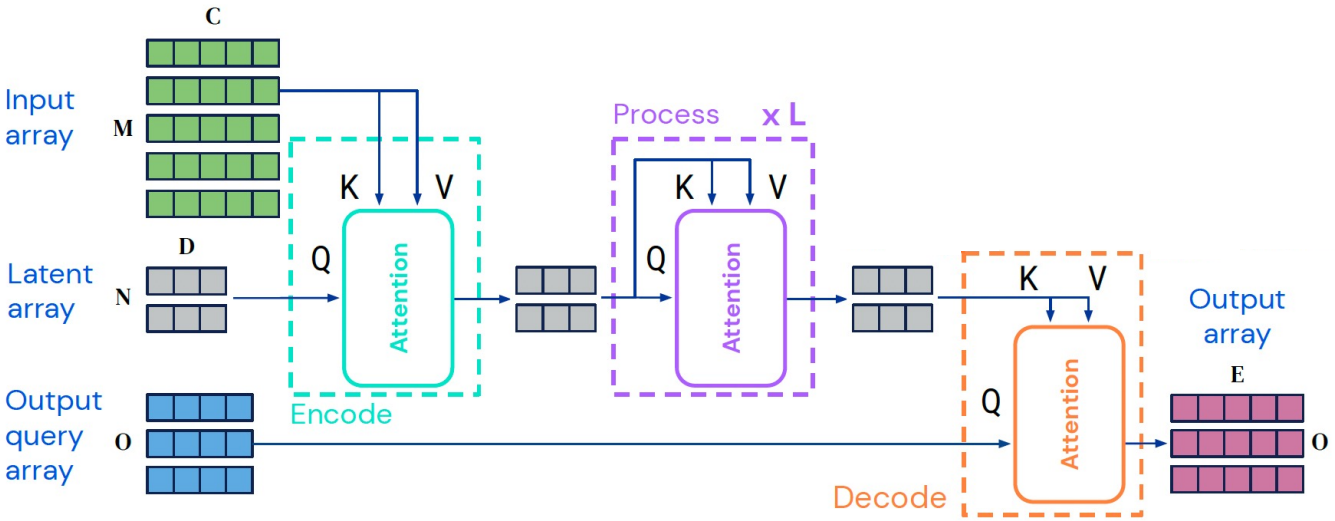
1. Attention and Vision Transformers (ViT)
   - NLP: Attention is all you need
   - Transformer Encoder ViT with Self Attention for image classification



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

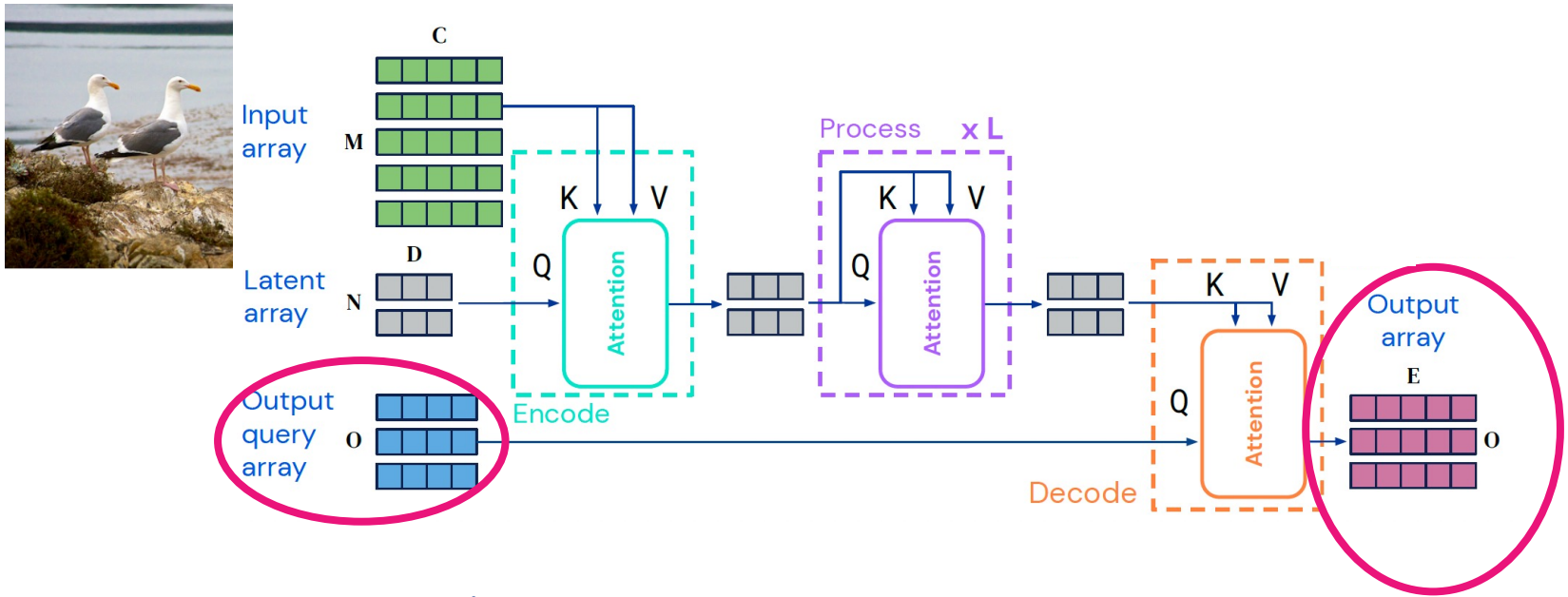
General Decoder

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



General Decoder

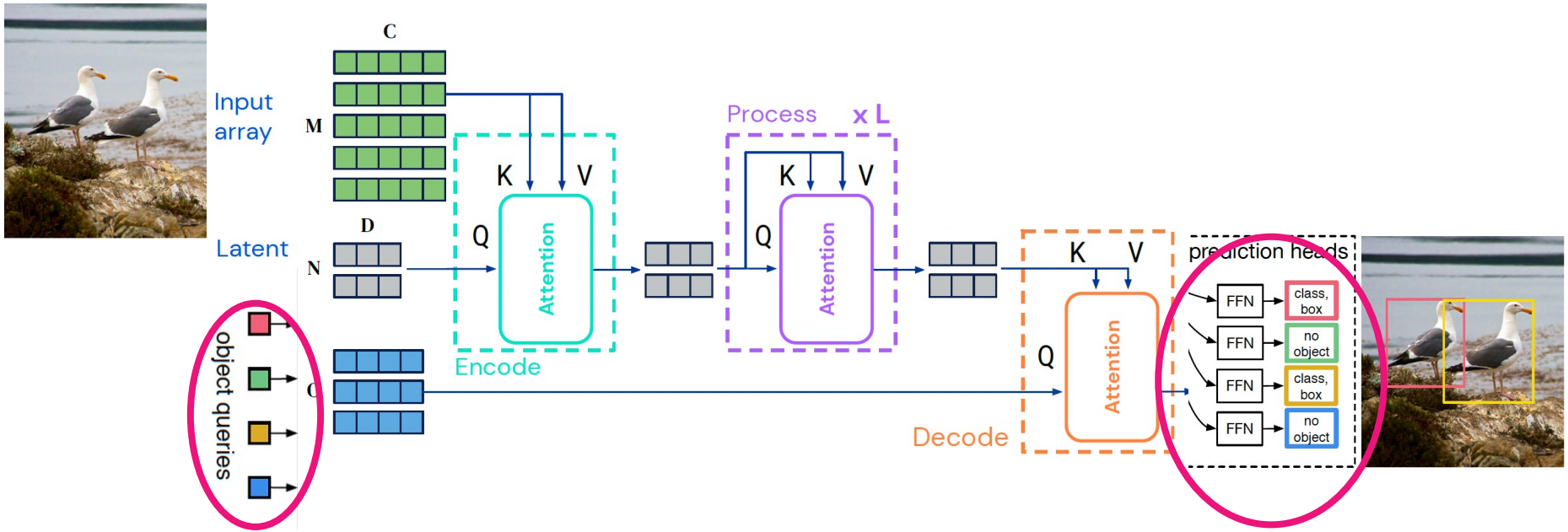
[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]



Output query array / Output array defines the downstream task: detection, segmentation …

General Decoder

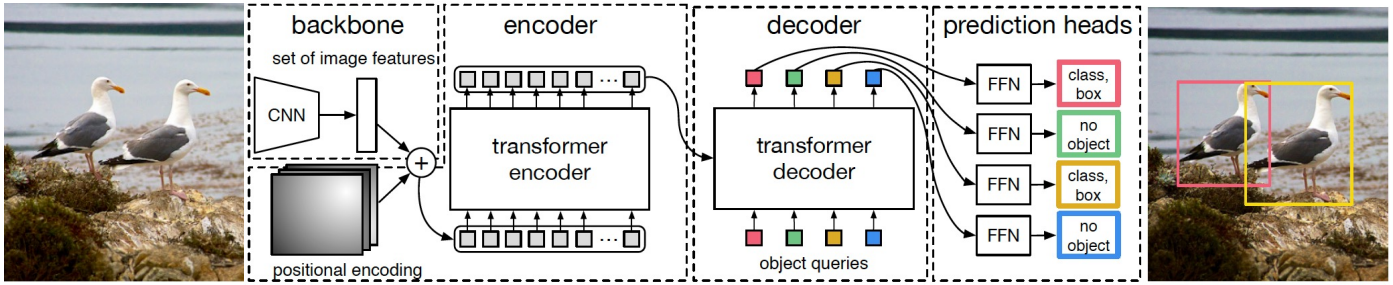
[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]

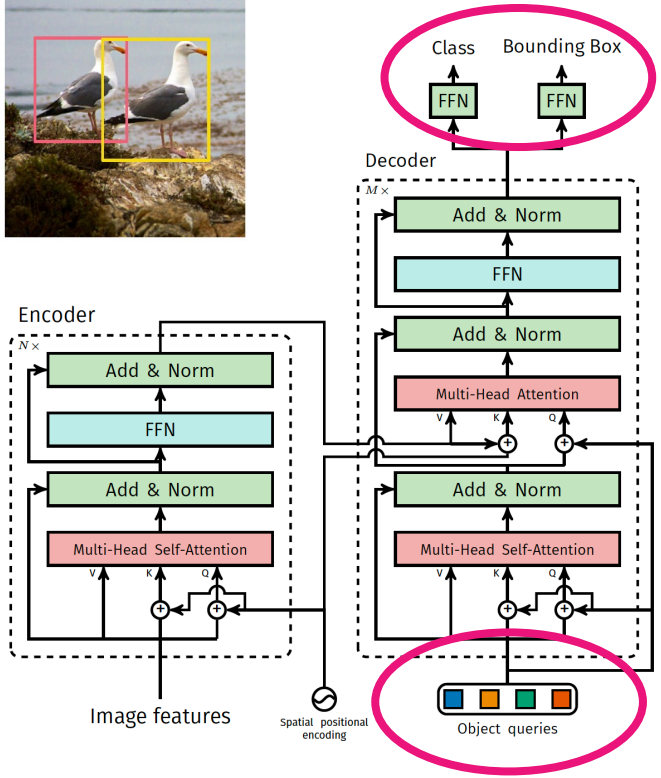


Output query array / Output array defines the downstream task: detection

Transformer Decoder for detection

Just another scheme for DETR model





Cornell University

arXiv > cs > arXiv:2005.12872

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 26 May 2020 (v1), last revised 28 May 2020 (this version, v3)]

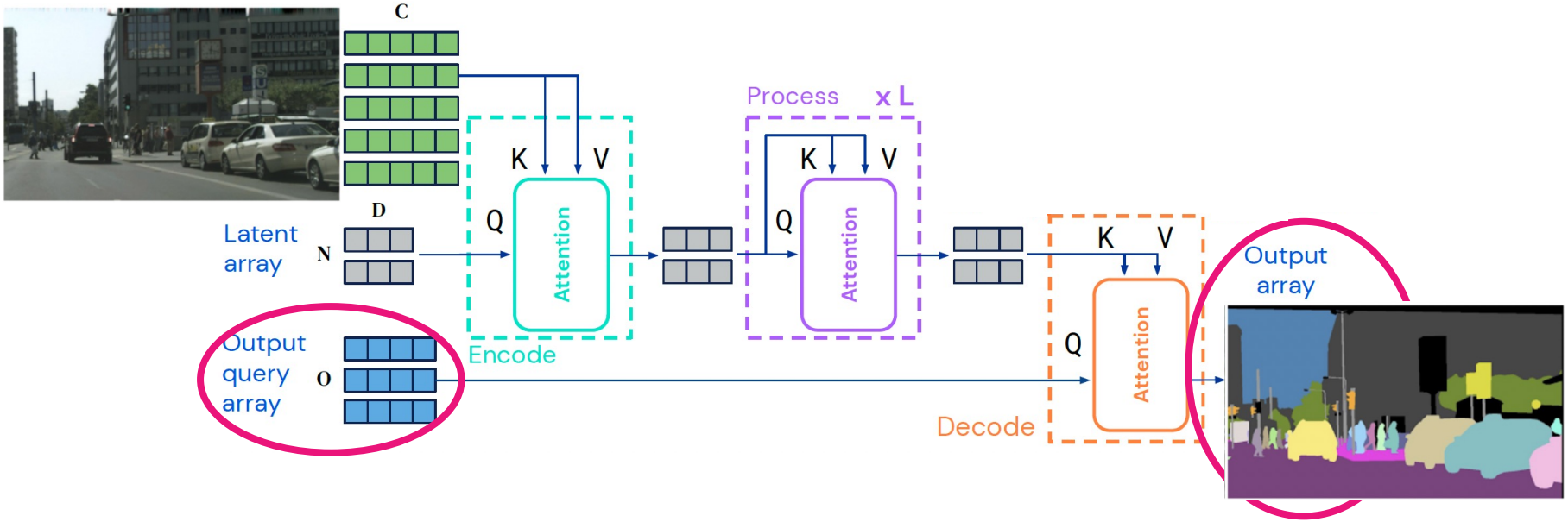
End-to-End Object Detection with Transformers

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 hand-designed components like a non-maximum suppression procedure or anchor generation that explicitly encode our prior knowl the new framework, called DEtection TRansformer or DETR, are a set-based global loss that forces unique predictions via bipartite ma

General Decoder

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

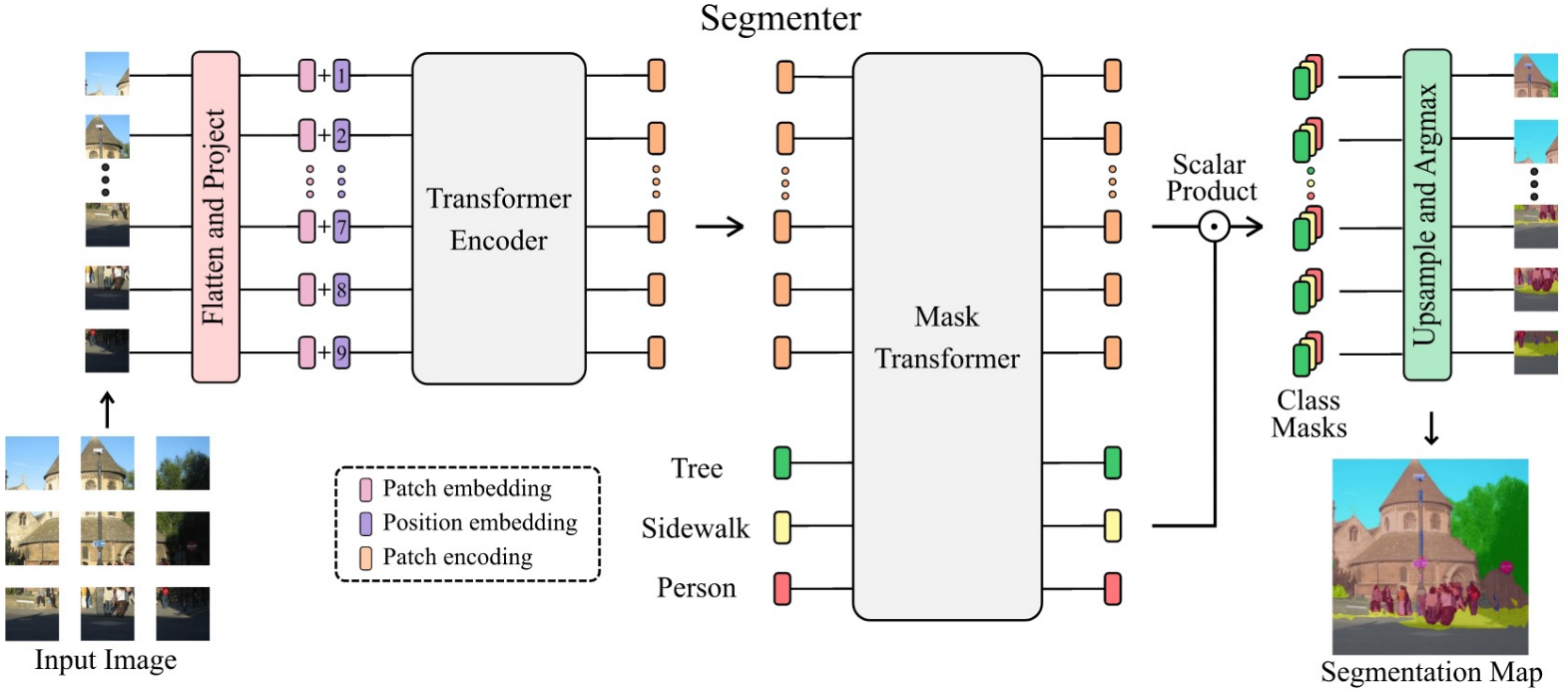


Output query array / Output array defines the downstream task: segmentation …

General Decoder: or not!

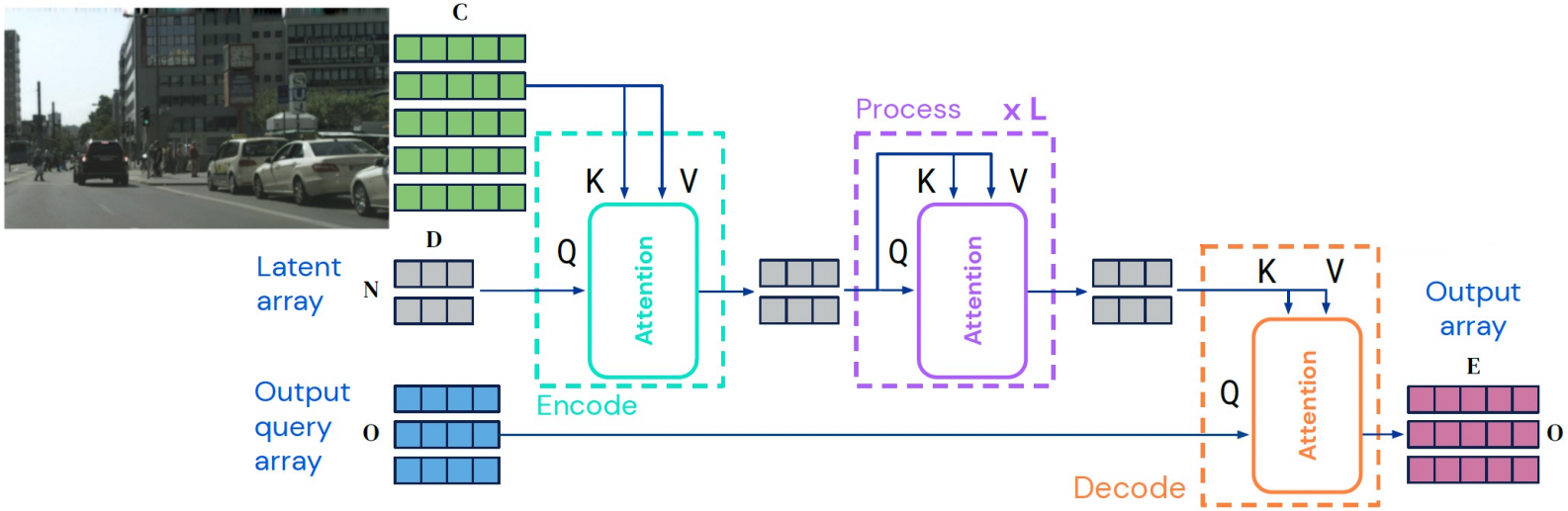
Segmenter: Transformer for Semantic Segmentation

Robin Strudel* Inria†
Ricardo Garcia* Inria†
Ivan Laptev Inria†
Cordelia Schmid Inria†



General Decoder

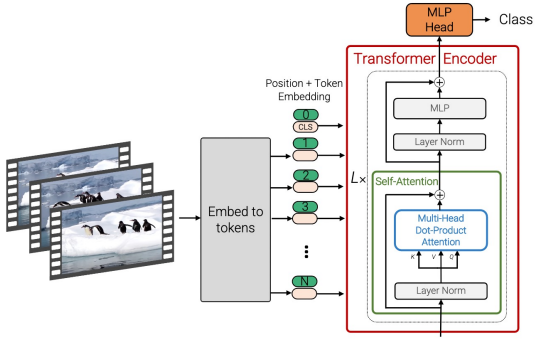
[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]

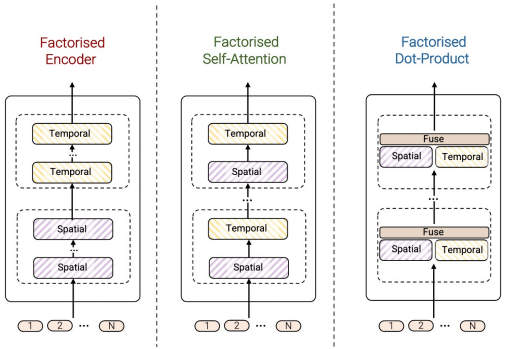


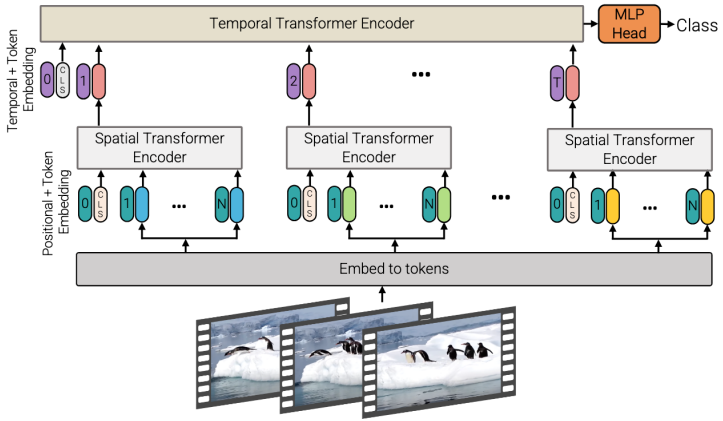
Output query array / Output array defines the downstream task: continual learning

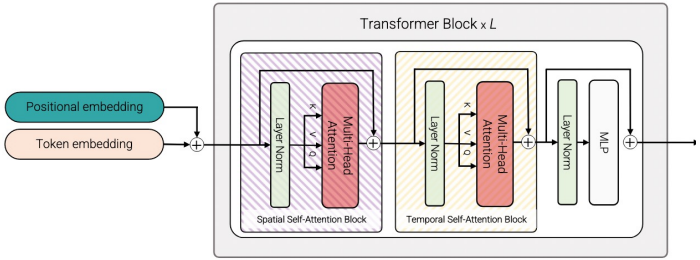
Video Transformer

[ViViT: A Video Vision Transformer ICCV 2021]



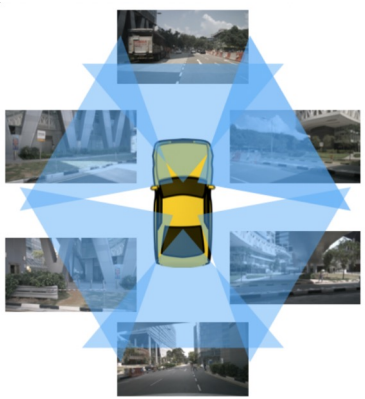






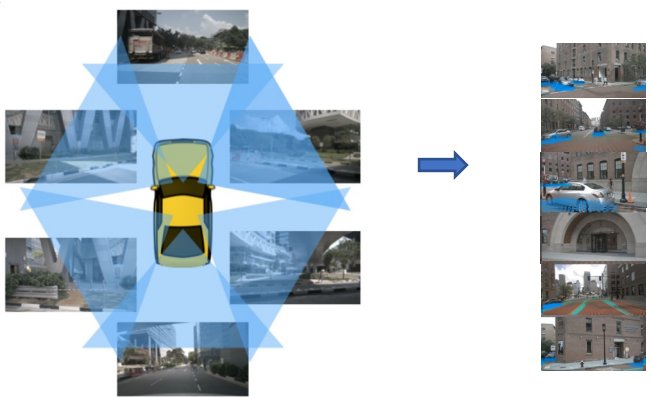
General Encoder / Decoder

Input array = N cameras



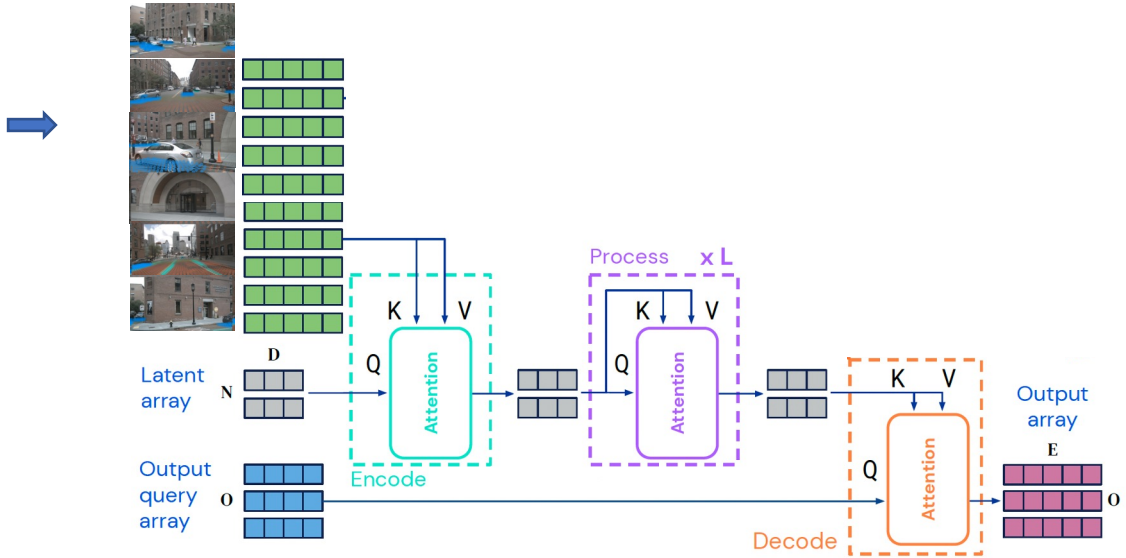
General Encoder / Decoder

Input array = N cameras



General Encoder / Decoder

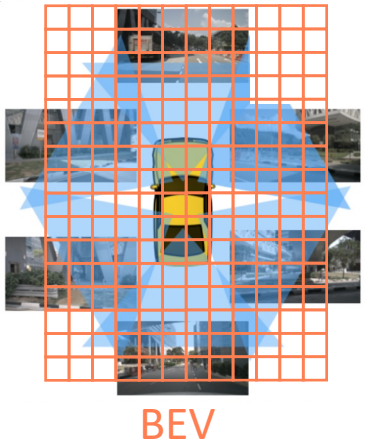
Input array = N cameras

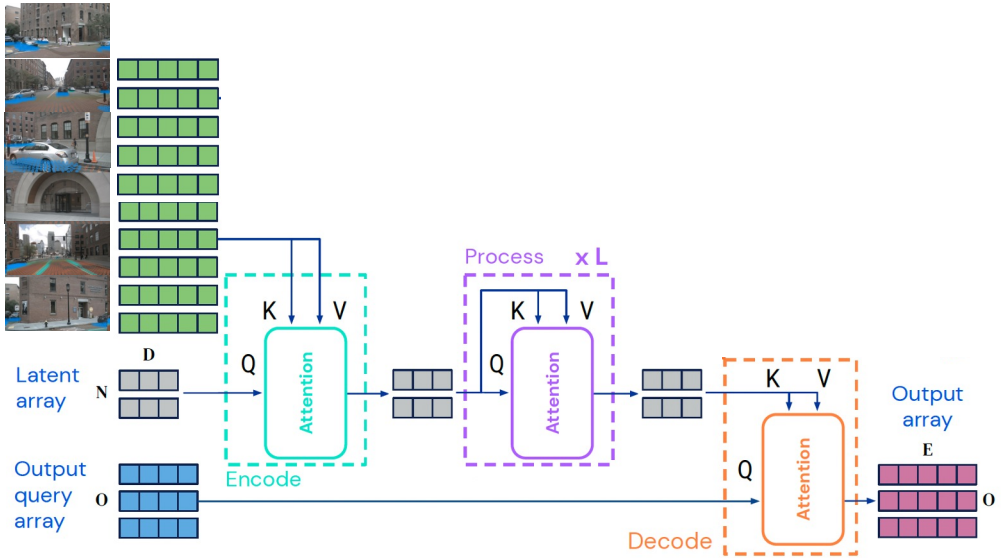


General Encoder / Decoder

Input array = N cameras

Output array = Bird Eye View (BEV) representation

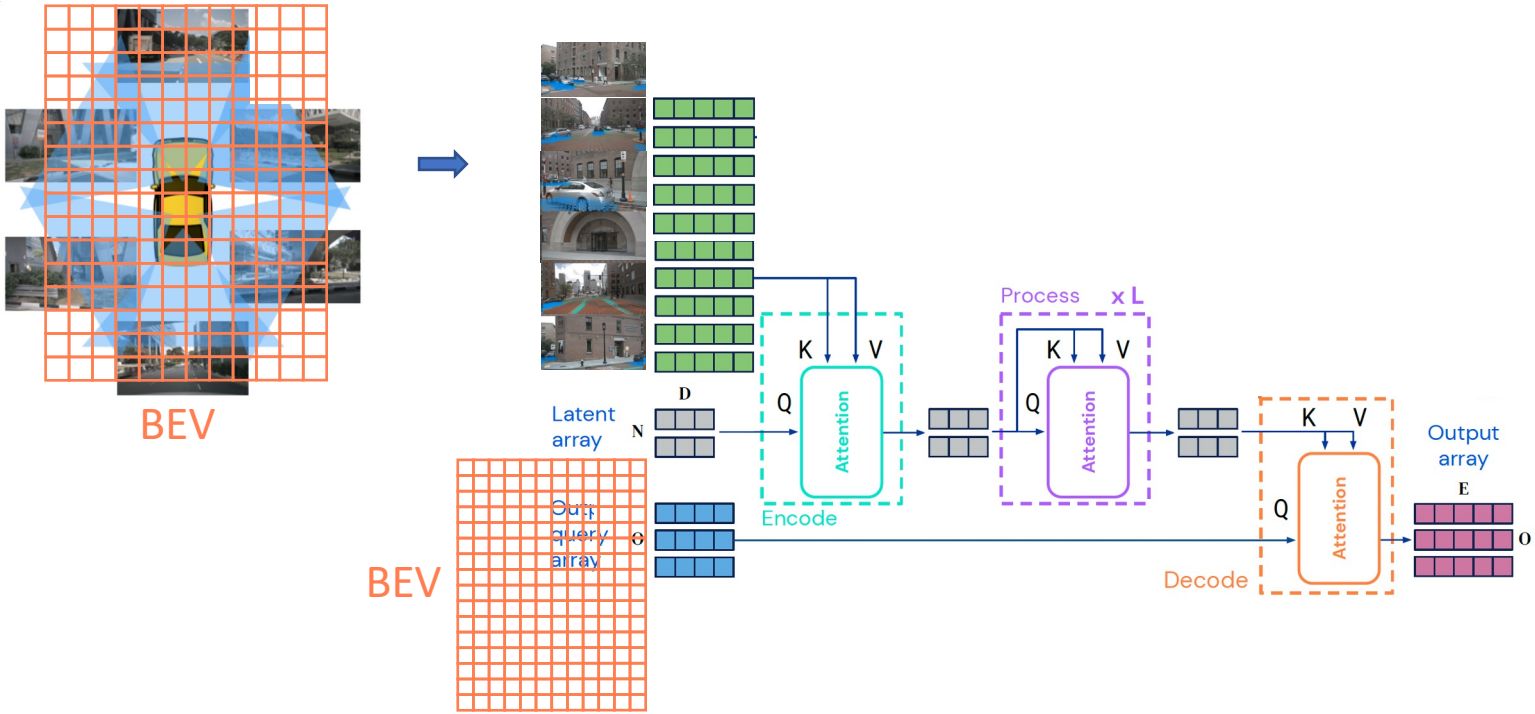




General Encoder / Decoder

Input array = N cameras

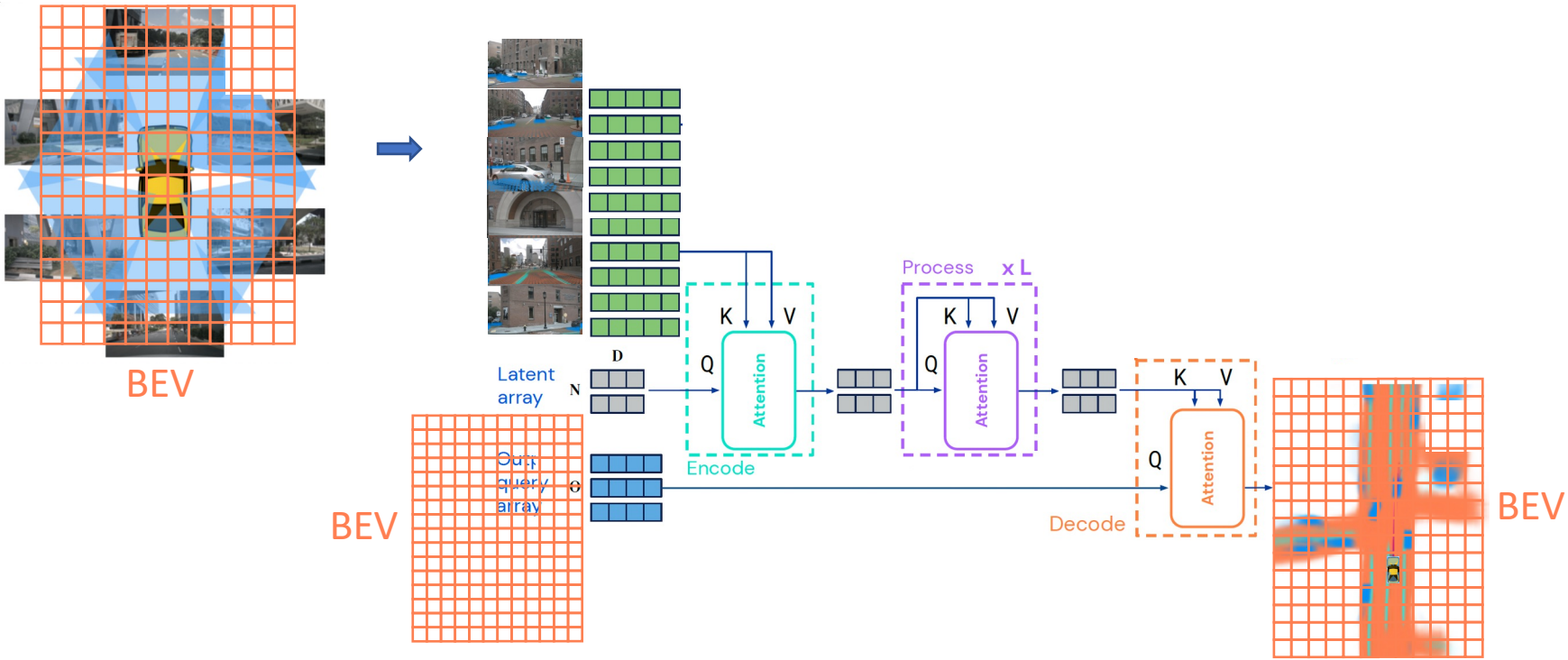
Output array = Bird Eye View (BEV) representation



General Encoder / Decoder

Input array = N cameras

Output array = Bird Eye View (BEV) representation



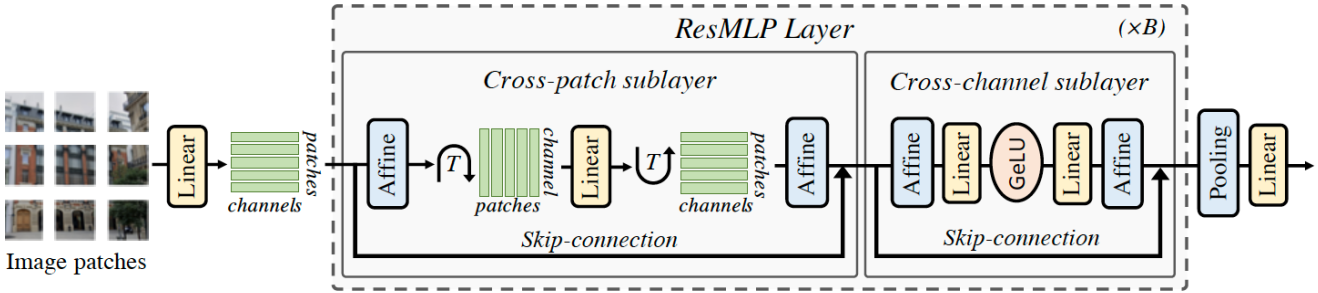
Vision Transformers

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, …



Outline

1. Attention (from NLP) and Vision Transformers (ViT)
2. Transformer Decoder for downstream tasks
3. **Mixing Visual and Language Models**

CLIP: Vision + Language Models (VLM)

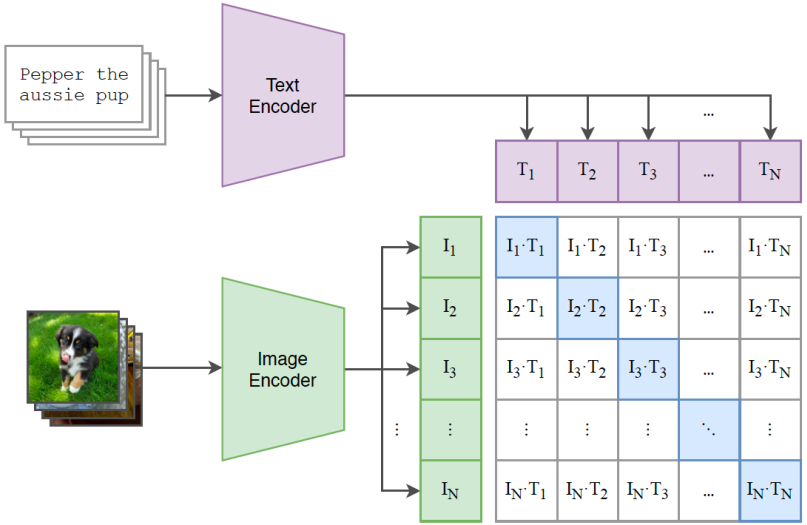
[Learning transferable visual models from natural language supervision. Radford/Sutskever ICML, 2021]

CLIP: Vision + Language Models (VLM)

[Learning transferable visual models from natural language supervision. Radford/Sutskever ICML, 2021]

Massive Text+Image =**500M pairs** pre-trained model (from Internet = no manual labeling)

Contrastive loss for pre-training

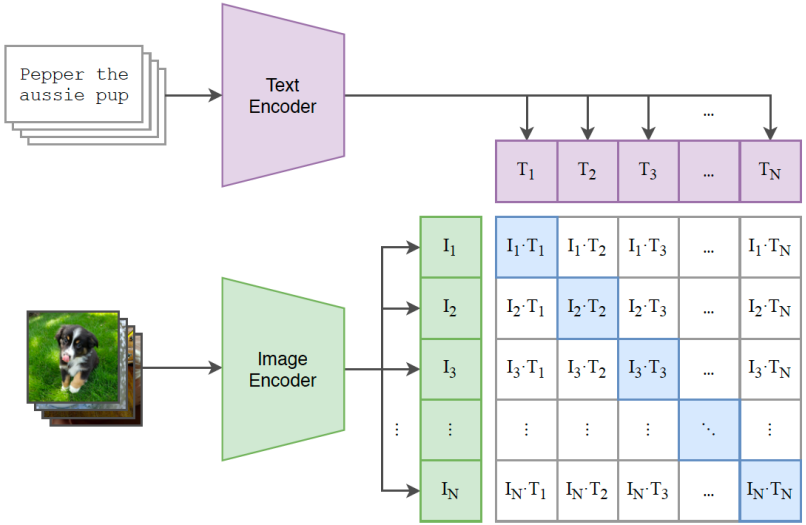


CLIP: Vision + Language Models (VLM)

[Learning transferable visual models from natural language supervision. Radford/Sutskever ICML, 2021]

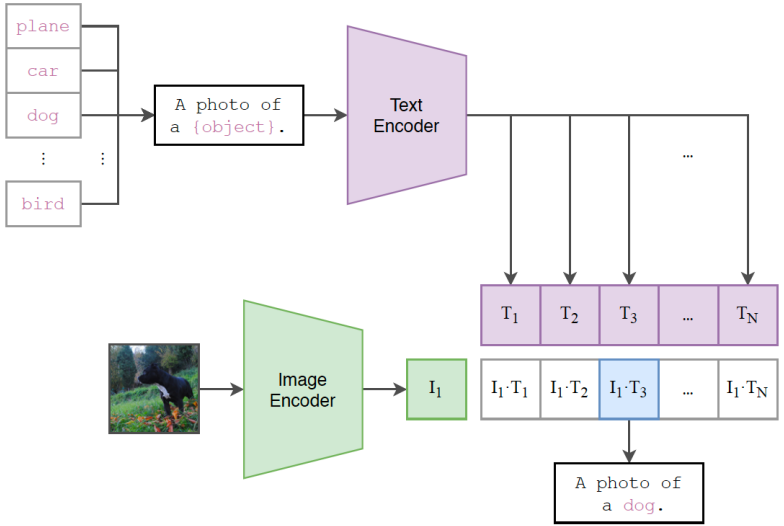
Massive Text+Image =**500M pairs** pre-trained model (from Internet = no manual labeling)

Contrastive loss for pre-training



Pre-trained encoders = **dual encoders** (Text/Image)

used for Zero-shot classifier, and other downstream tasks

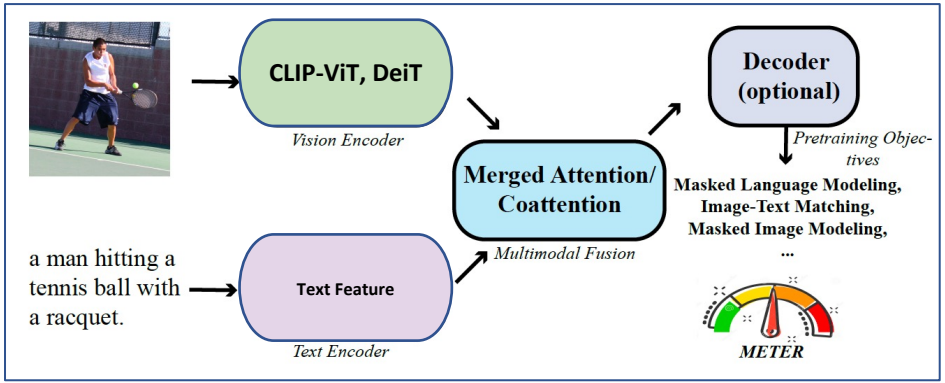


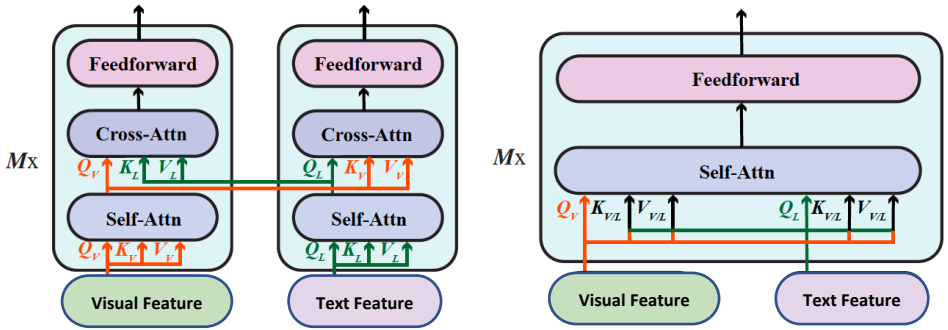
Training End-to-End Vision-and-Language Transformers

From **dual** encoders (Text/Image)

Training VLM Multimodal Fusion on specific datasets and pretext tasks

Commonly used datasets (4M images in total): COCO, Conceptual Captions, SBU Captions, and Visual Genome





(a) Co-attention model. **(b)** Merged attention model.

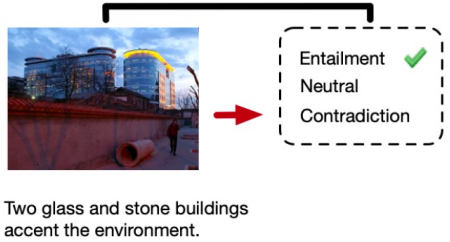
VLM experiments

Several Complex visual understanding tasks
VLMs give the **best** results

Visual reasoning (NLVR2)
Answer: yes or no

Visual Question Answering

Visual Entailment

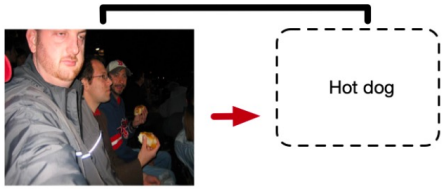


VLM experiments

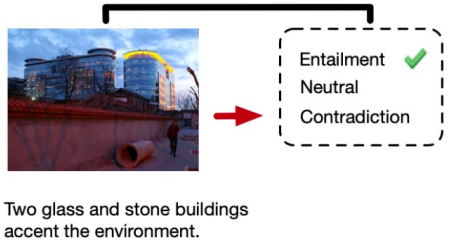
Several Complex visual understanding tasks
VLMs give the **best** results

Visual reasoning (NLVR2)
Answer: yes or no

Visual Question Answering



Visual Entailment



VLM => Foundation models

What is a foundation model? " *In recent years, a new successful paradigm for building AI systems has emerged: Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model*" (from Stanford Center for Research on Foundation Models).

Language/vision/etc models (BERT/GPT3+ViT…) with (huge) training data
From Web pages => Sequence of tokens (text,images, video,audio) no labeling required

To sum up

2010s: Convnets with imagenet (1.2M images 1000 class/words) per-sample labels

2020s?: Transformers => All is Token and transformers for all tokens

- Starting with 500M pairs of images+text => free language description

Many ongoing research challenges:

- Training foundation Models
- Building extra models on top of VLM that can be rapidly adapted to numerous tasks using only a handful of annotated examples (or even none)
- Better understanding the latent structure of these multimodal spaces and interaction functions