Transformers in Vision

Camille Guinaudeau
Université Paris Saclay, LISN / CNRS
guinaudeau@limsi.fr

Part III

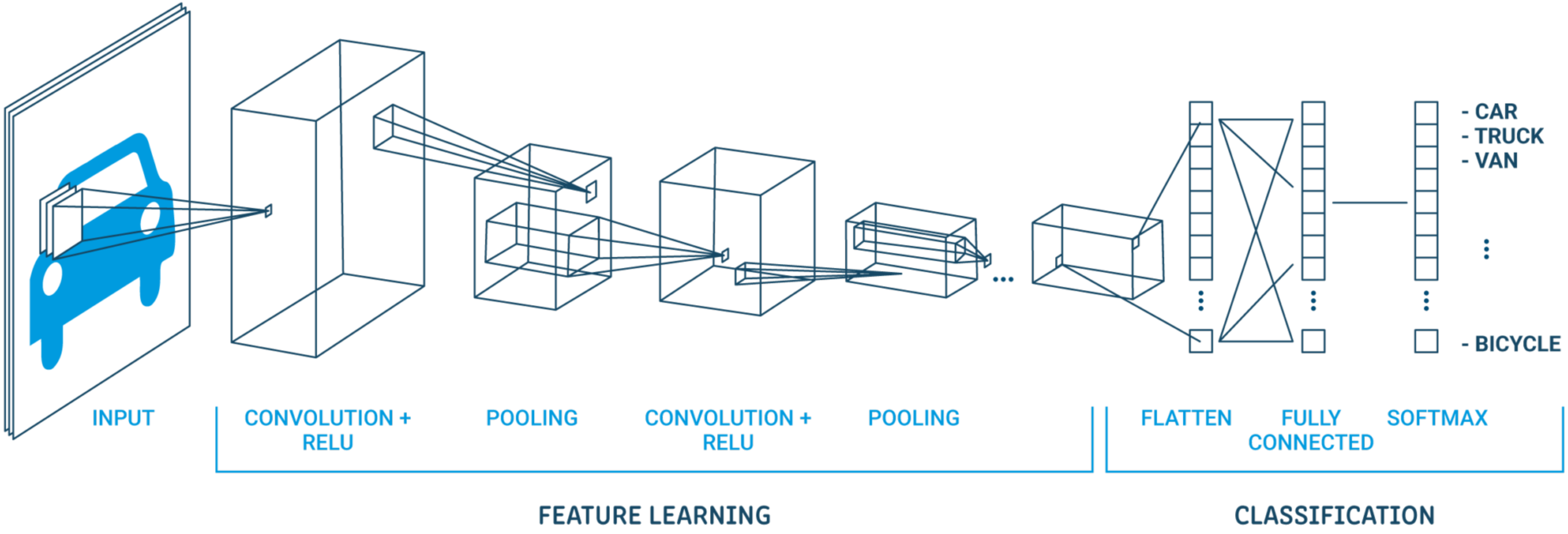
Transformers in Vision

Outline

- From CNN to Vision Transformers
  - Convolutional models
  - Self-attention layers
- Vision Transformers
  - Explicit positional encoding [Dosovitskiy et al., 2020], [Touvron et al., 2021]
  - Implicit positional encoding [Chu et al., 2021]
  - Introducing Convolutions to Vision Transformers [Wu et al., 2021]
- Multi-Modal Transformers
  - Text + Image [Radford et al., 2021]
  - Text + Video [Gabeur et al., 2020]

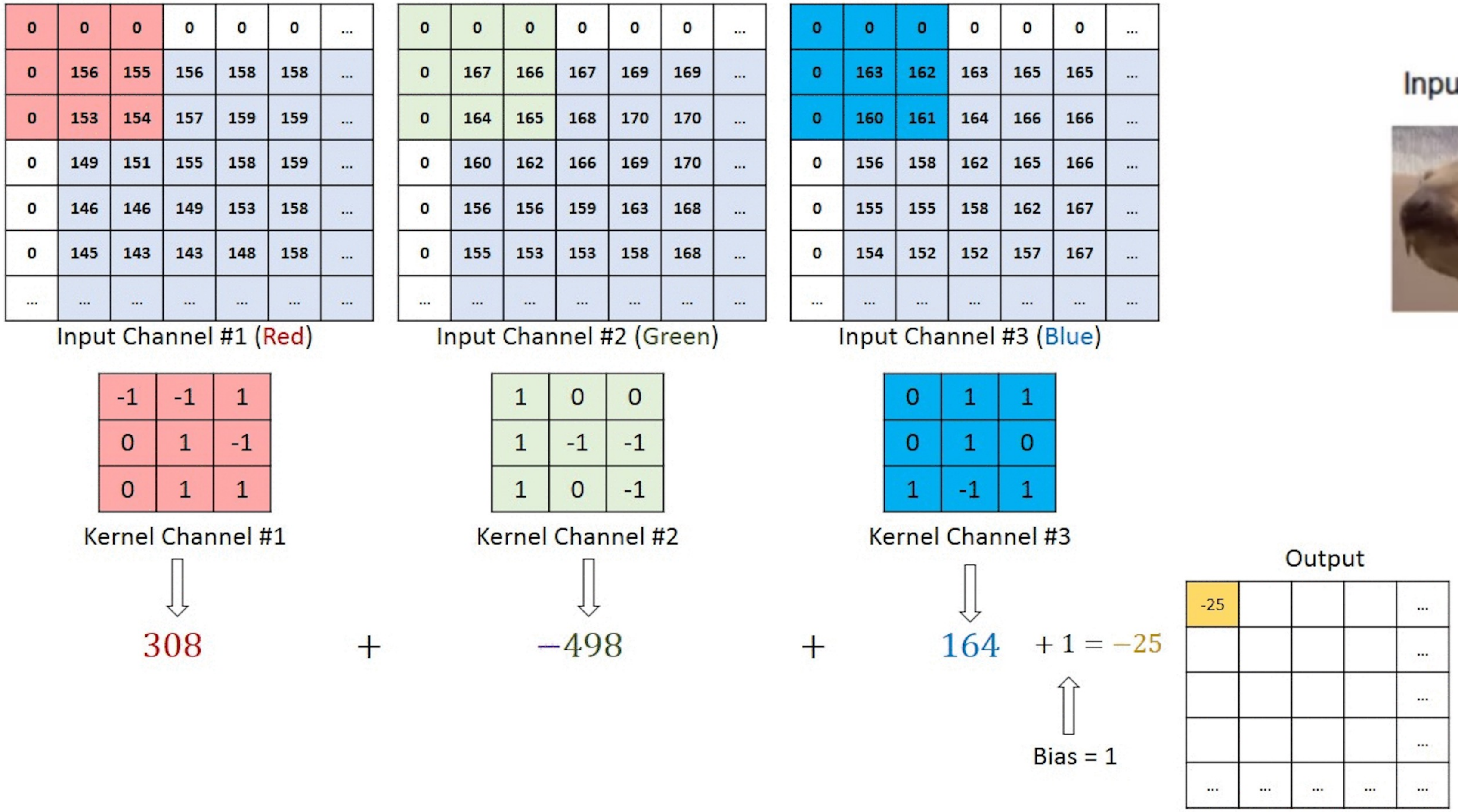
From CNN to Vision Transformers

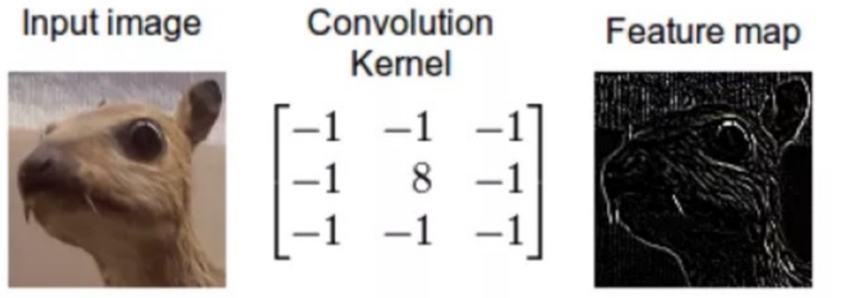
Convolutional models



From CNN to Vision Transformers

Convolutional layer





Example

From CNN to Vision Transformers

Convolutional models

Convolutional operations has two important spatial constraints:

- Translation invariance
- Local sensitivity

Limitations:

- Lack a global understanding of the image
- Complex models

From CNN to Vision Transformers

Self-attention

As opposed to convolution layers whose receptive field is the $K \times K$ neighborhood grid, the self-attention's receptive field is always the full image

Self-attention layers take a feature map as input

- compute attention weights between every pair of features
- each position has information about any other
- can replace or be combined with convolutions

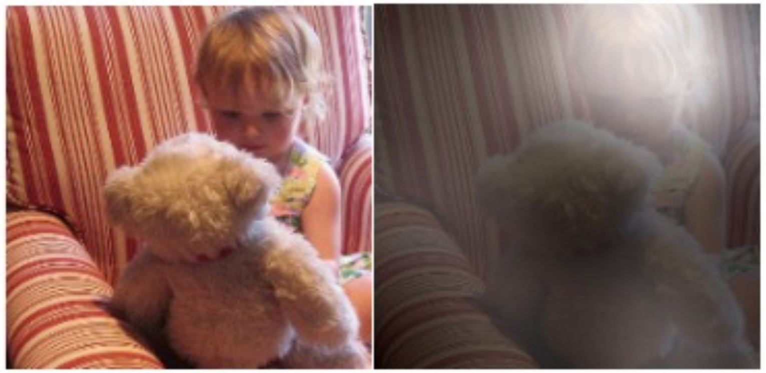
From CNN to Vision Transformers

[Xu et al., 2015]

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

From CNN to Vision Transformers

Self-attention

Soft Attention: the alignment weights are learned and placed over all patches in the source image

Pro: the model is smooth and differentiable

Con: expensive when the source input is large

Hard Attention: only selects one patch of the image to attend to at a time.

Pro: less calculation at the inference time

Con: the model is non-differentiable and requires more complicated techniques to train

From CNN to Vision Transformers

Self-attention

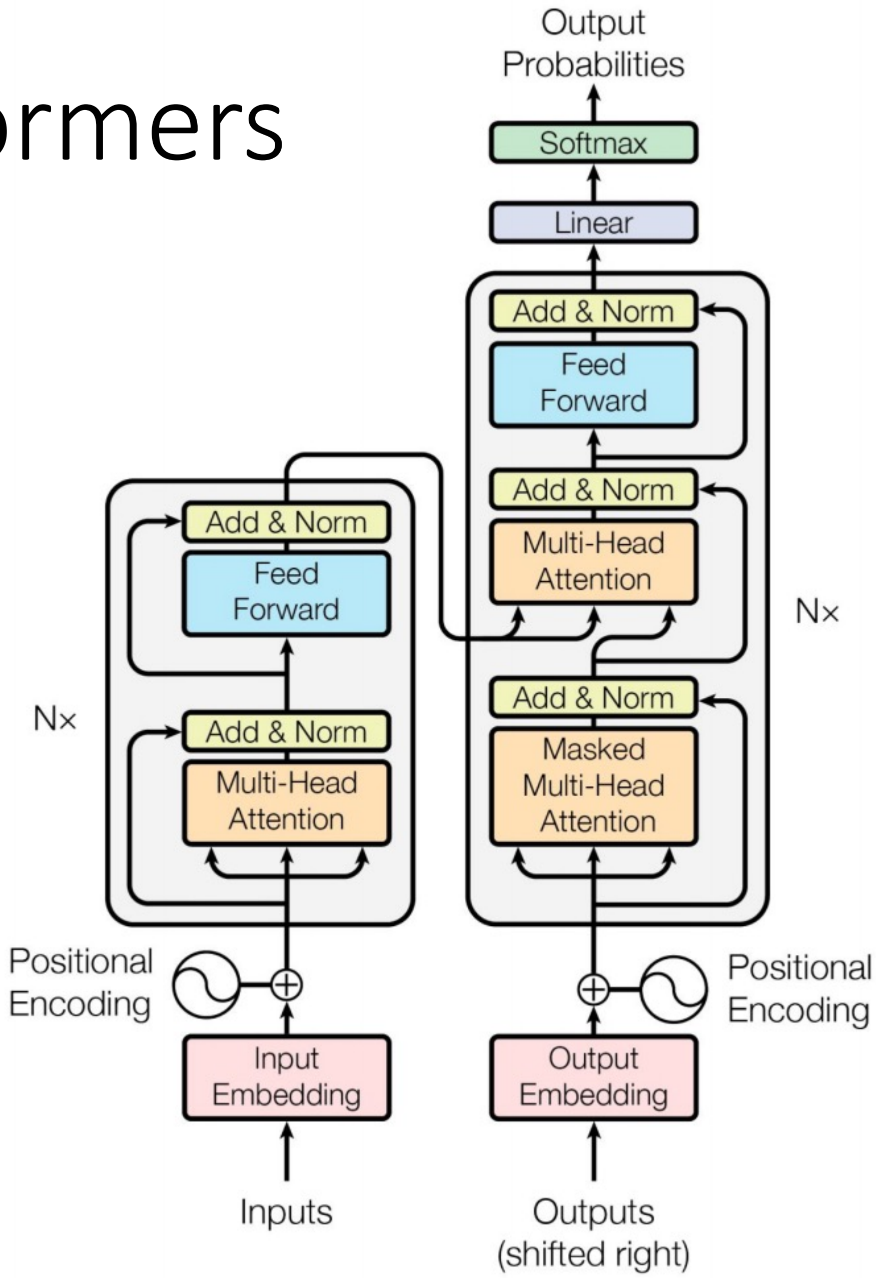
augmenting convolution models with self-attention:

- Video classification and object detection [Wang et al., 2018]
- Video action recognition [Chen et al., 2018]
- Object detection and image classification [Bello et al., 2019]

Limitations: computation cost can be expensive for high resolution input

- Attention computation along the two spatial axis sequentially instead of the whole image [Wang et al., 2020]
- Patches of feature maps instead of the whole spatial dimensions [Ramachandran et al., 2019]

Vision Transformers



[Vaswani et al., 2017]

Vision Transformers

How to deal with images in Transformer?

Two strategies:

Modify the input

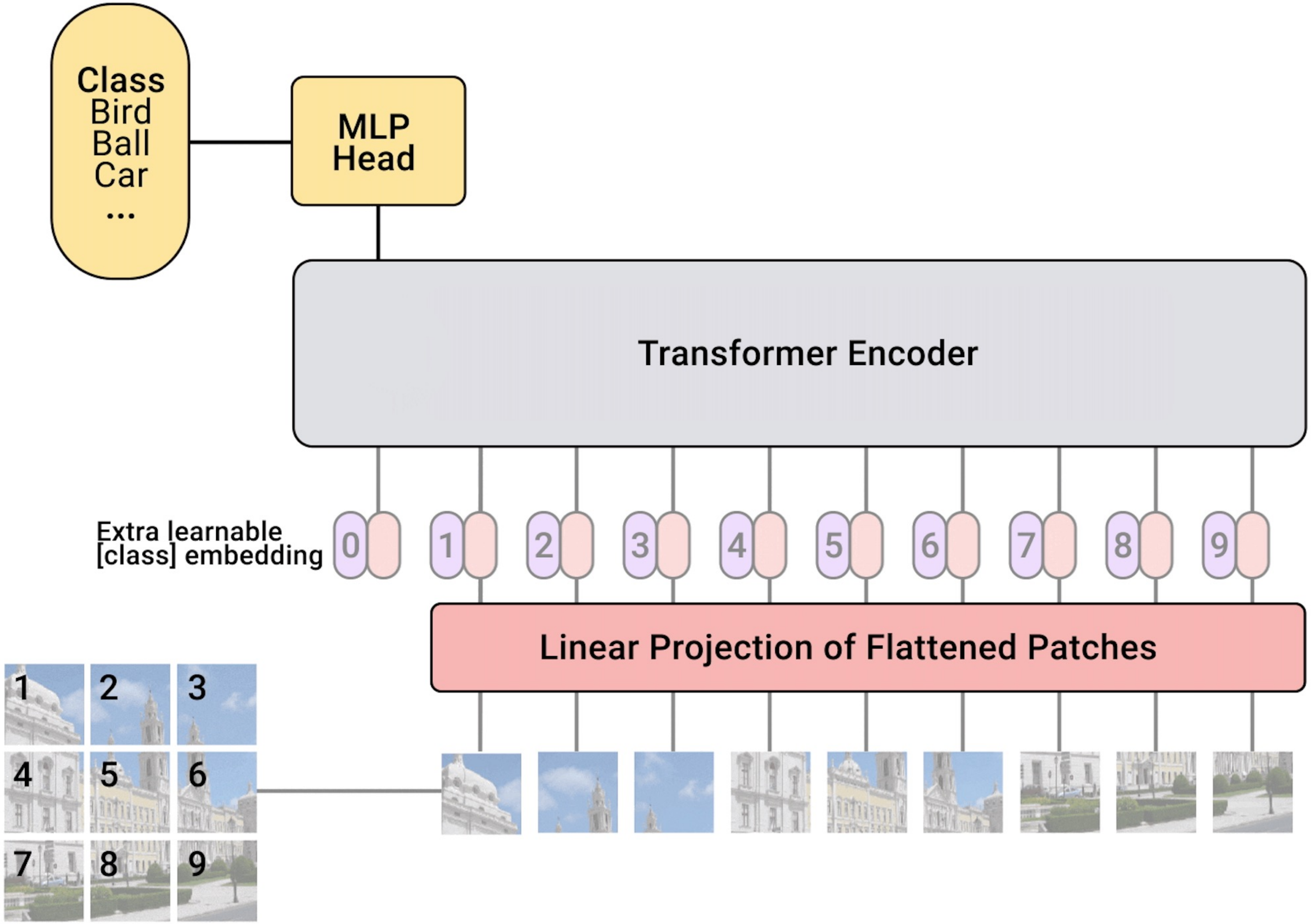
- Vision Transformer [Dosovitskiy et al., 2020]
- DeiT: Data-efficient Image Transformers [Touvron et al., 2021]
- Conditional Positional Encodings for Vision Transformers [Chu et al., 2021]

Modify the architecture

- CvT: Introducing Convolutions to Vision Transformers [Wu et al., 2021]

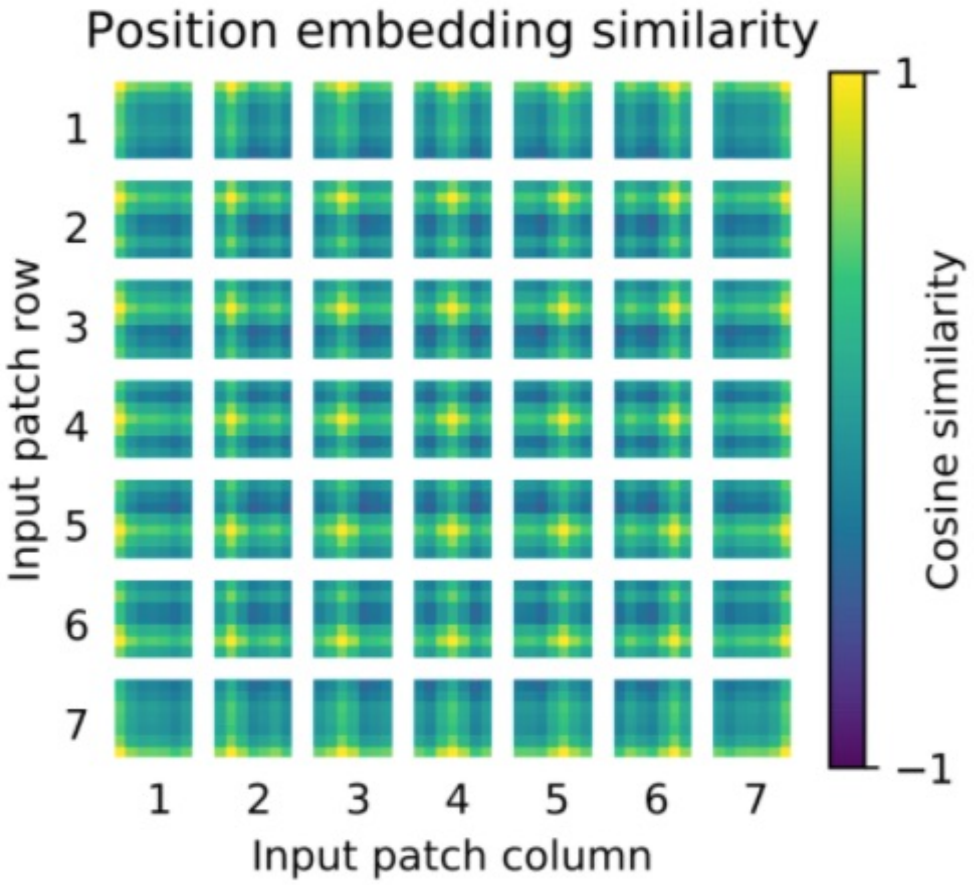
Vision Transformers

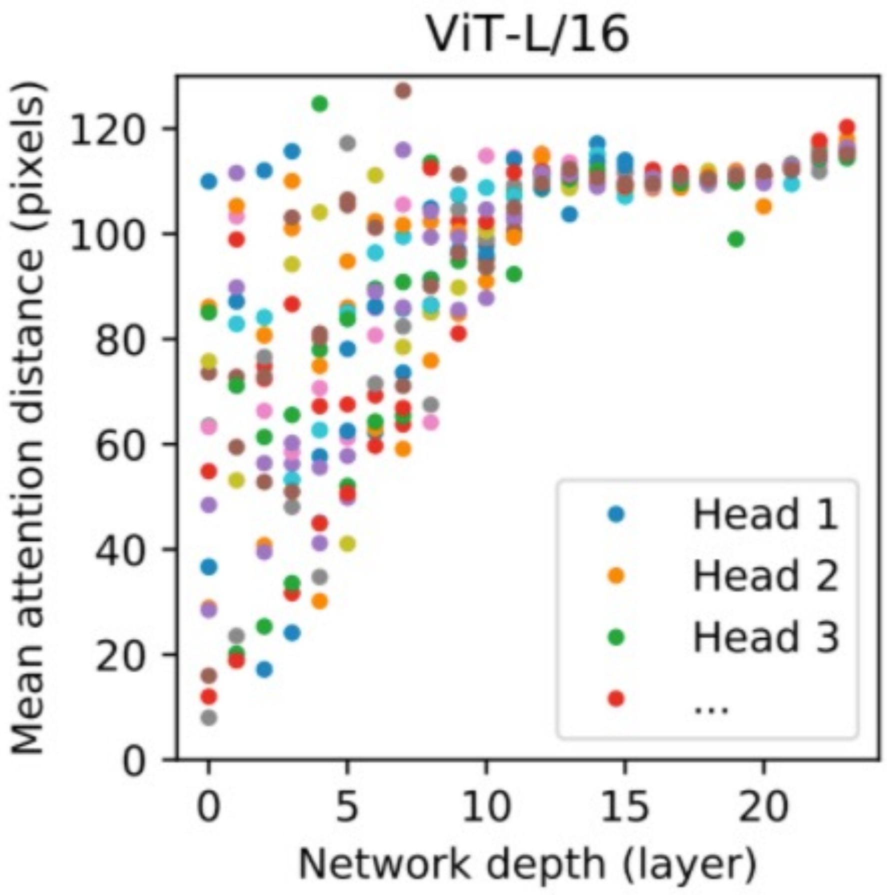
[Dosovitskiy et al., 2020]



Vision Transformers

[Dosovitskiy et al., 2020]





Vision Transformers

[Dosovitskiy et al., 2020]

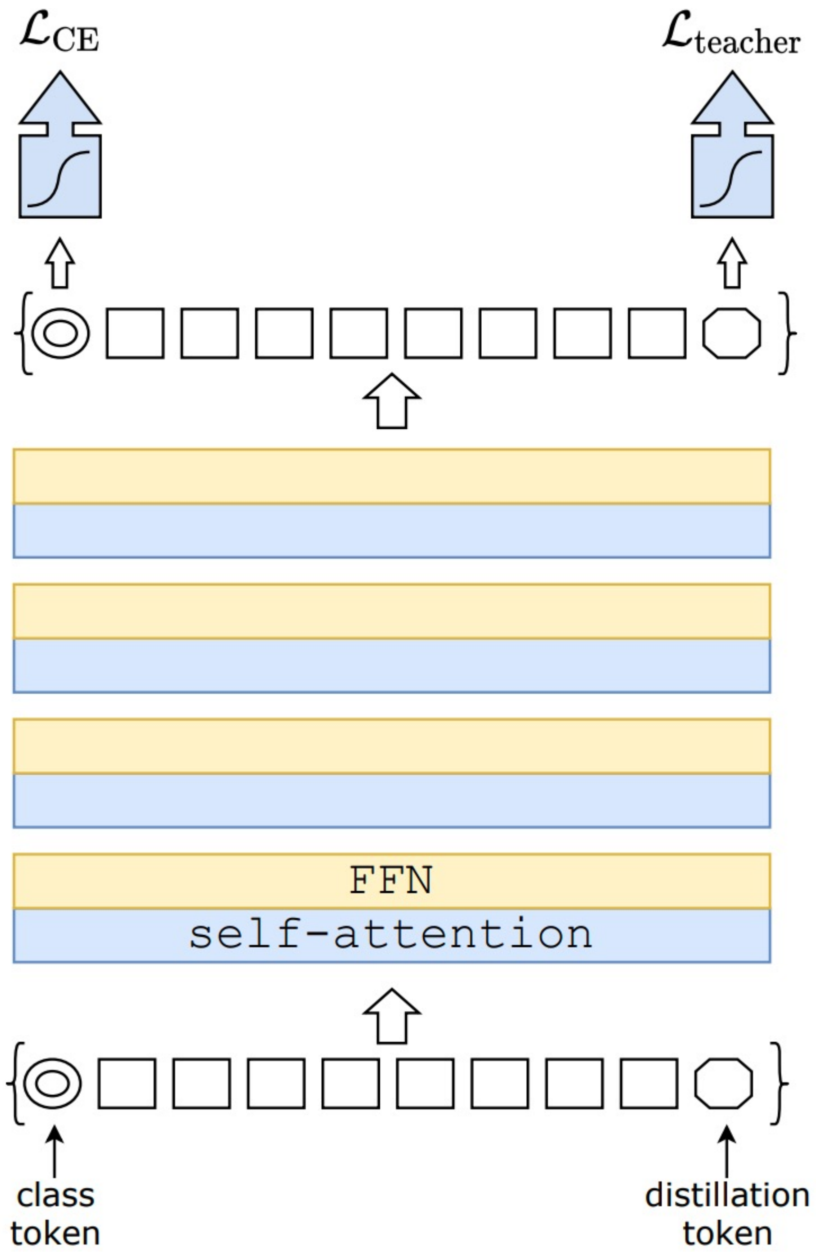
	ViT-H	Previous SOTA
ImageNet	88.55	88.5
ImageNet-ReaL	90.72	90.55
Cifar-10	99.50	99.37
Cifar-100	94.55	93.51
Pets	97.56	96.62
Flowers	99.68	99.63

Vision Transformers

Data-efficient Image Transformers

Introduction of a **knowledge distillation** procedure specific for vision transformers

Training one neural network (the student) on an output of another network (the teacher)



[Touvron et al., 2021]

Vision Transformers

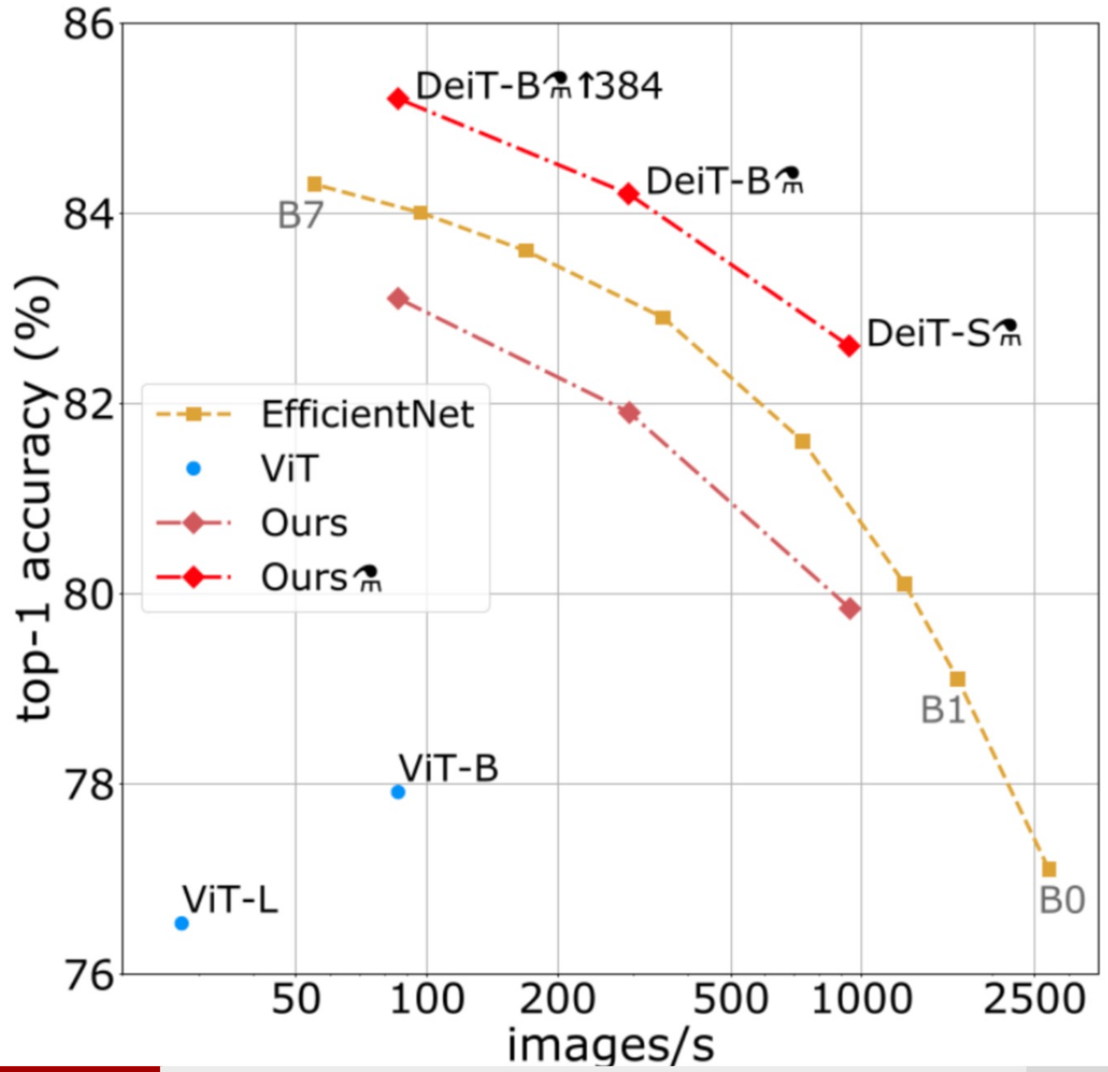
Fixing the positional encoding across resolutions

- Use a lower training resolution and fine-tune the network at the larger resolution speeds up the full training and improves the accuracy

- Interpolate the positional encoding when changing the resolution

Vision Transformers

[Touvron et al., 2021]



Vision Transformers

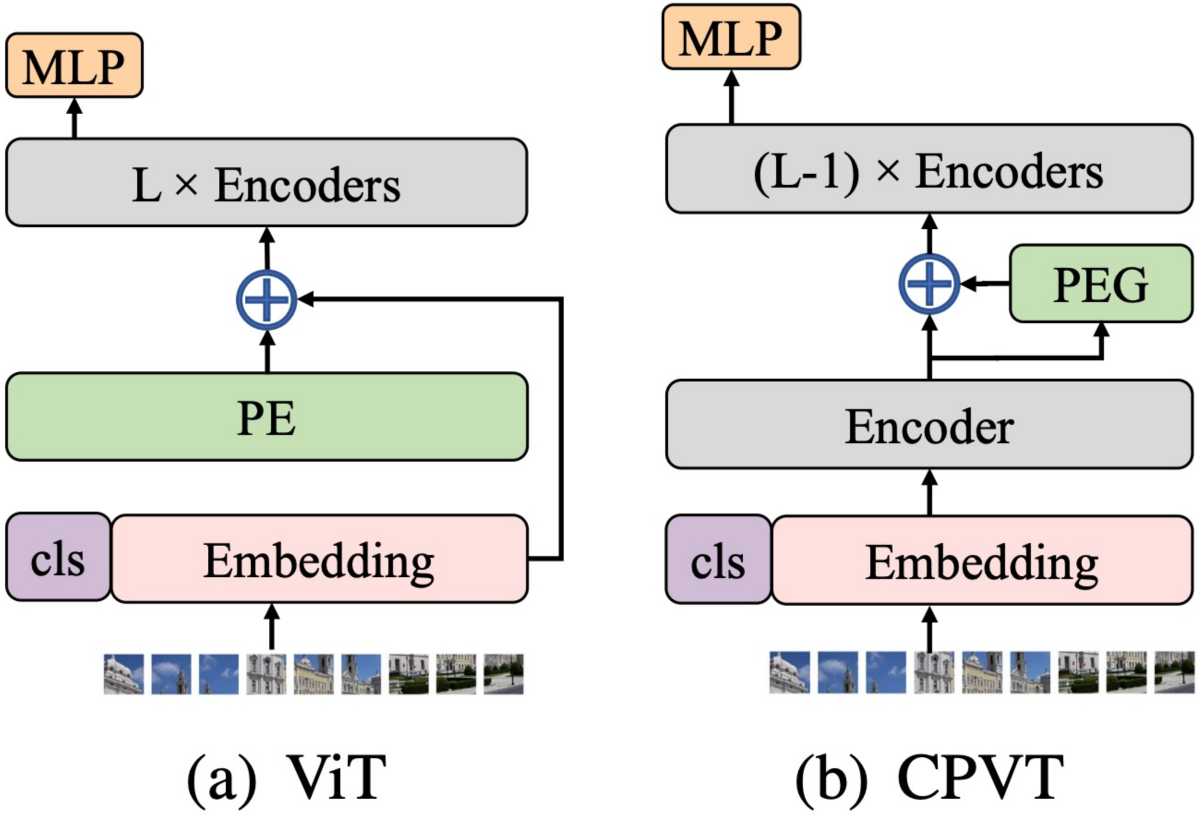
Positional Encodings' Limitations:

- Positional encodings have a negative impact on the flexibility of the Transformers
- Absolute positional encoding scheme breaks the translation-invariance
- Relative positional encodings do not work equally well as the absolute ones

Vision Transformers

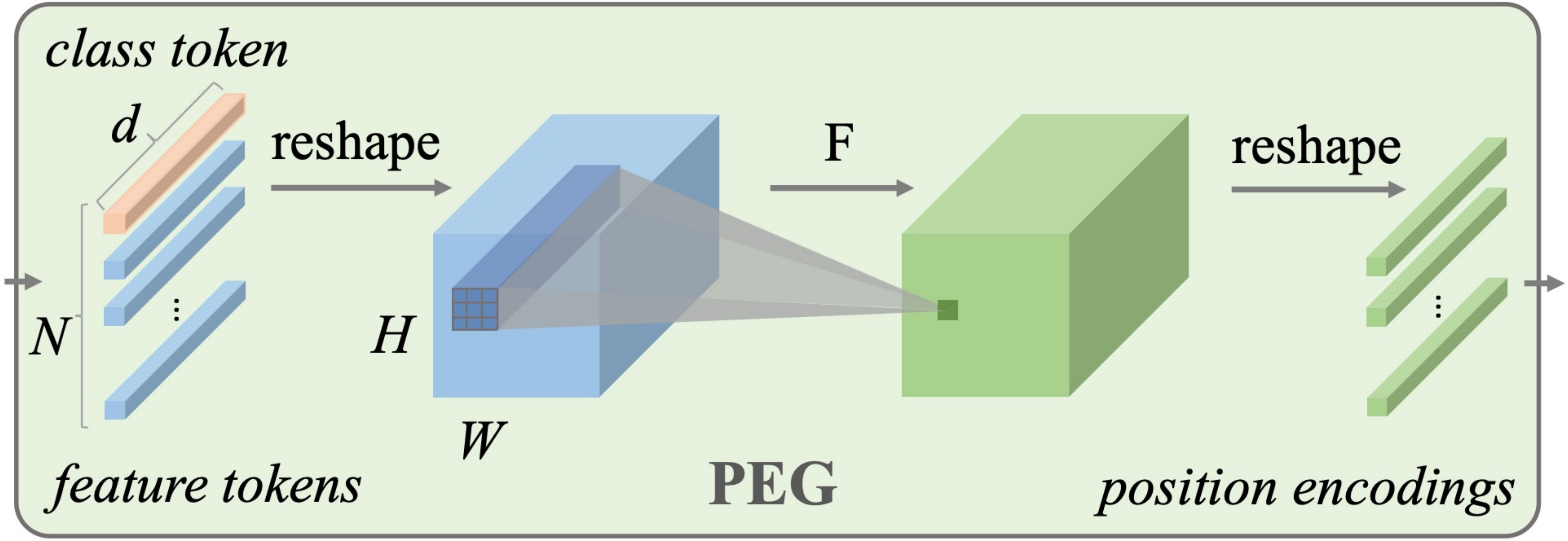
Conditional Positional Encodings for Vision Transformers (CPVT)

[Chu et al., 2021]



Vision Transformers

[Chu et al., 2021]



Vision Transformers

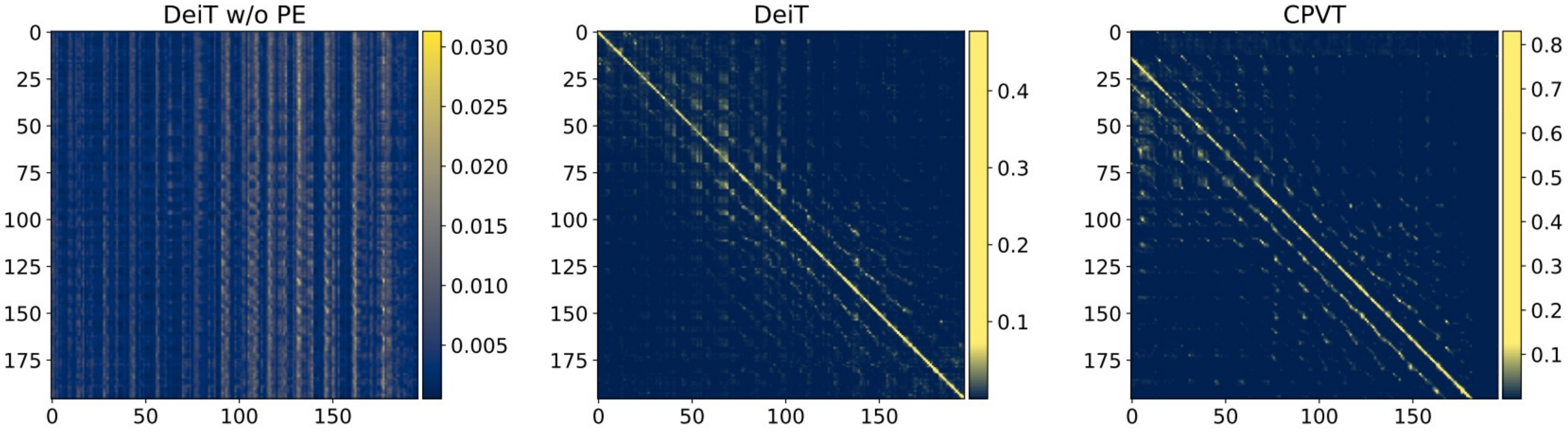
[Chu et al., 2021]

Model	Params	Top-1@224(%)	Top-1@384(%)
DeiT-tiny [30]	6M	72.2	71.2
DeiT-tiny (sine)	6M	72.3	70.8
CPVT-Ti	6M	72.4	73.2

Vision Transformers

PEG vs original positional encodings

[Chu et al., 2021]



Vision Transformers

CvT: Introducing Convolutions to Vision Transformers

Vision Transformers performances are still below similarly sized CNN counterparts when trained on smaller amounts of data

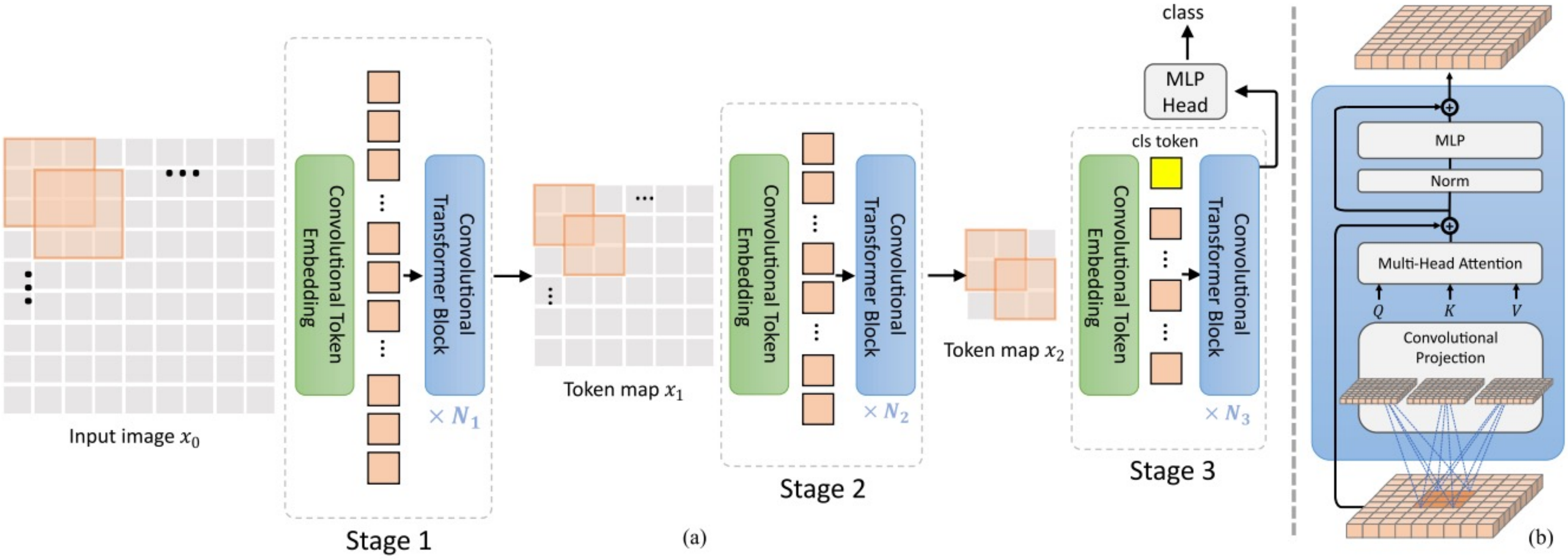
CNN architecture:

- capture local structure
- achieves shift, scale, and distortion invariance

Introduction of two convolution-based operations into the Vision Transformer architecture: Convolutional Token Embedding and Convolutional Projection [Wu et al., 2021]

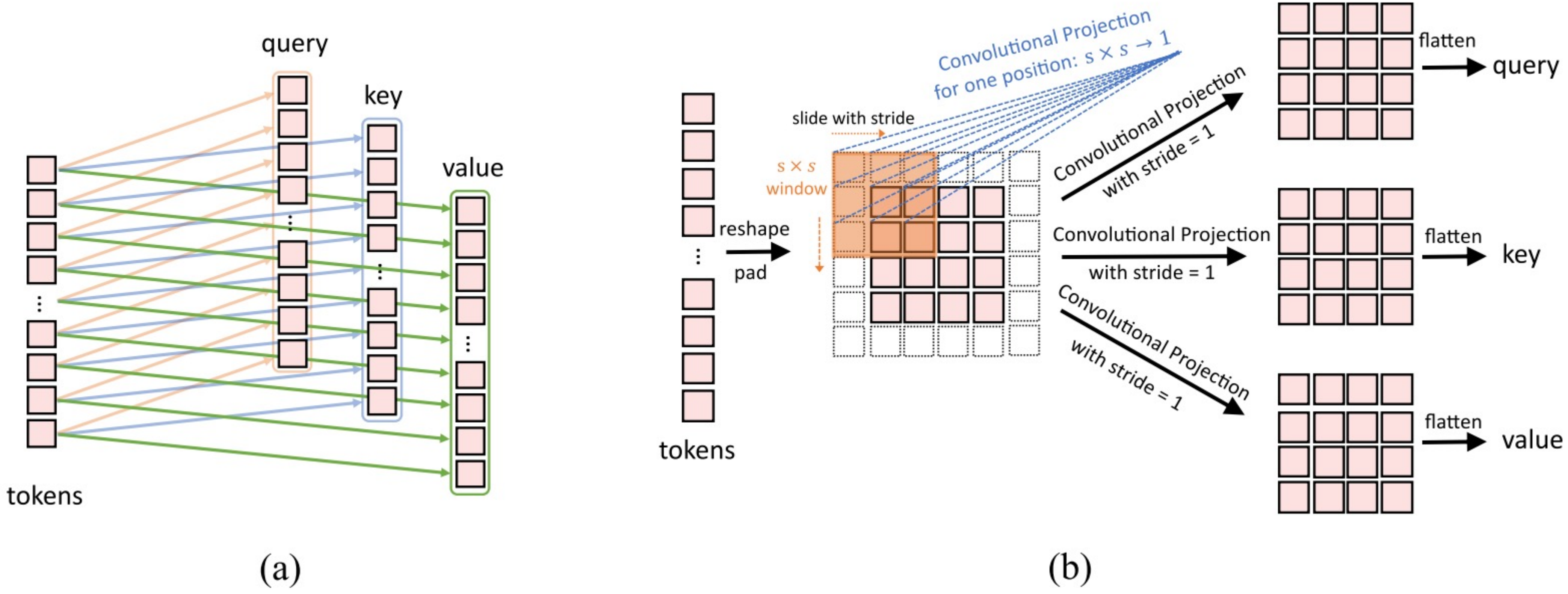
Vision Transformers

[Wu et al., 2021]



Vision Transformers

[Wu et al., 2021]



Vision Transformers

[Wu et al., 2021]

Method Type	Network	#Param. (M)	image size	FLOPs (G)	ImageNet top-1 (%)	Real top-1 (%)	V2 top-1 (%)
Convolutional Networks	ResNet-50 [15]	25	224^2	4.1	76.2	82.5	63.3
	ResNet-101 [15]	45	224^2	7.9	77.4	83.7	65.7
	ResNet-152 [15]	60	224^2	11	78.3	84.1	67.0
Transformers	ViT-B/16 [11]	86	384^2	55.5	77.9	83.6	–
	ViT-L/16 [11]	307	384^2	191.1	76.5	82.2	–
	DeiT-S [30][arxiv 2020]	22	224^2	4.6	79.8	85.7	68.5
	DeiT-B [30][arxiv 2020]	86	224^2	17.6	81.8	86.7	71.5
Convolutional Transformers	**Ours:** CvT-13	20	224^2	4.5	81.6	86.7	70.4
	Ours: CvT-21	32	224^2	7.1	82.5	87.2	71.3
	Ours: CvT-$13_{\uparrow 384}$	20	384^2	16.3	83.0	87.9	71.9
	Ours: CvT-$21_{\uparrow 384}$	32	384^2	24.9	**83.3**	**87.7**	**71.9**
	Ours: CvT-13-NAS	18	224^2	4.1	82.2	87.5	71.3

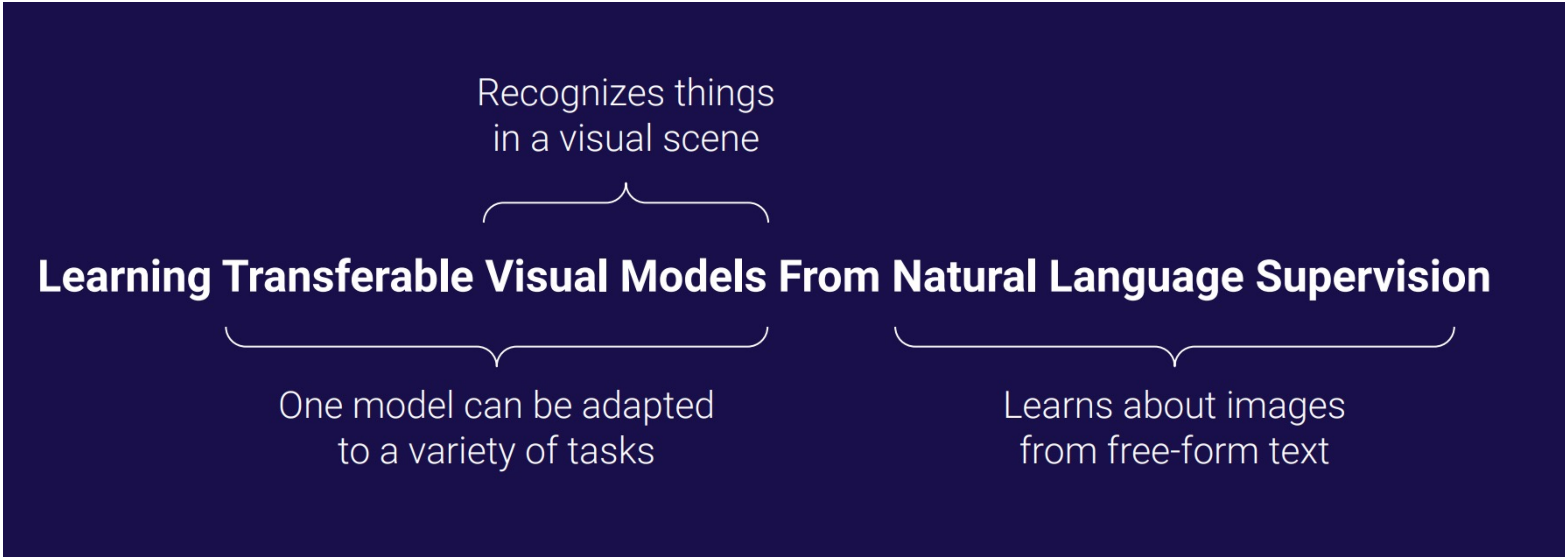
Multi-modal Transformers

Several modalities in Transformers:

- Image + Speech → AV Align [Sterpu et al. 2020]
- Text + Image → CLIP [Radford et al., 2021]
- Text + Video → [Gabeur et al., 2020]

Learning Transferable Visual Models From Natural Language Supervision

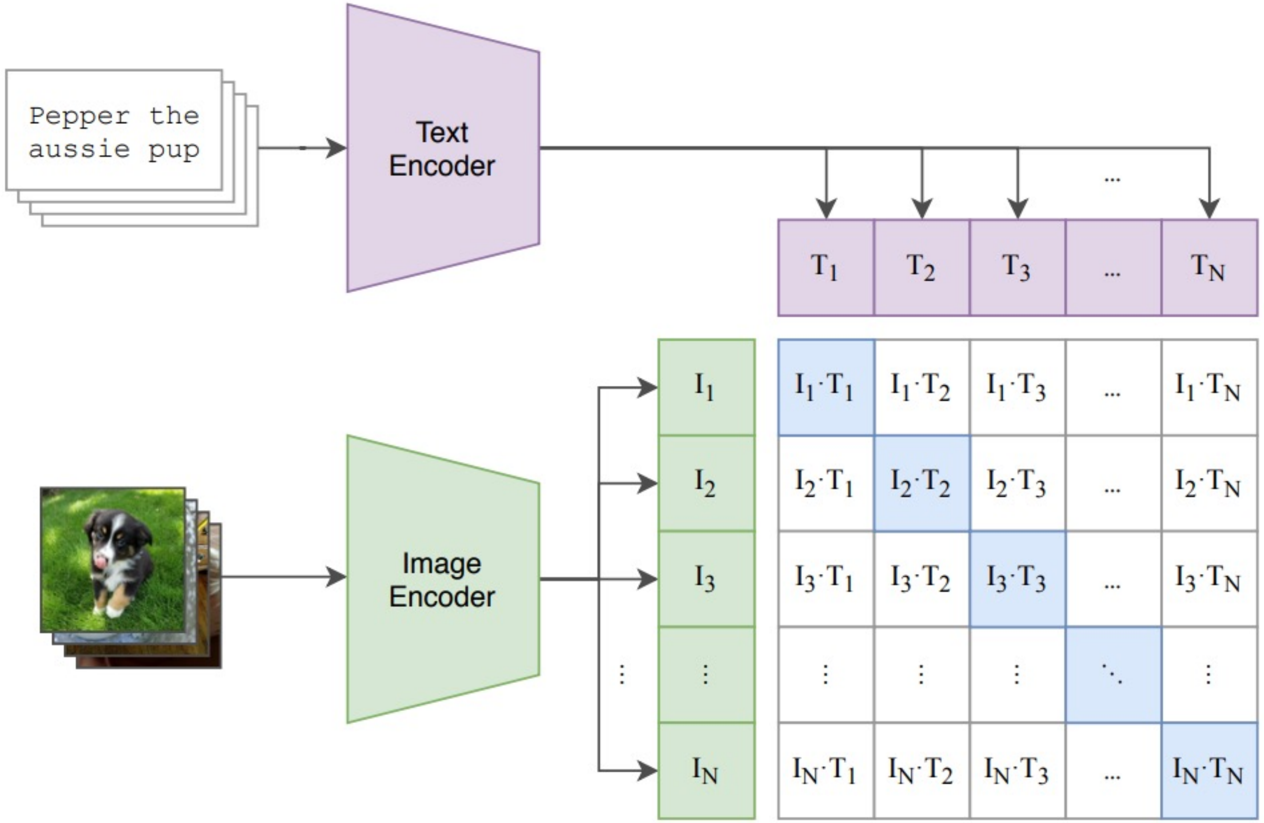
[Radford et al., 2021]



Learning Transferable Visual Models From Natural Language Supervision

[Radford et al., 2021]

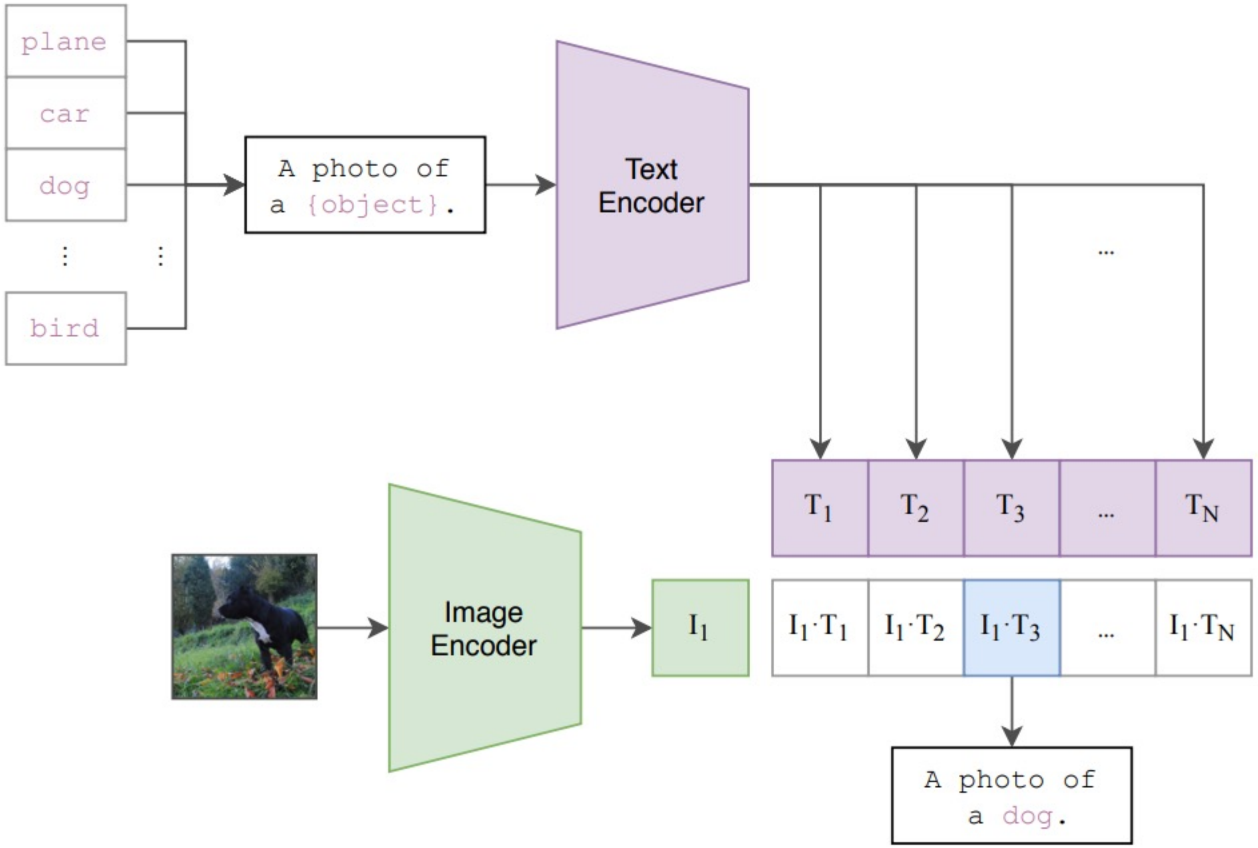
CLIP: Contrastive Language-Image Pre-training



Learning Transferable Visual Models From Natural Language Supervision

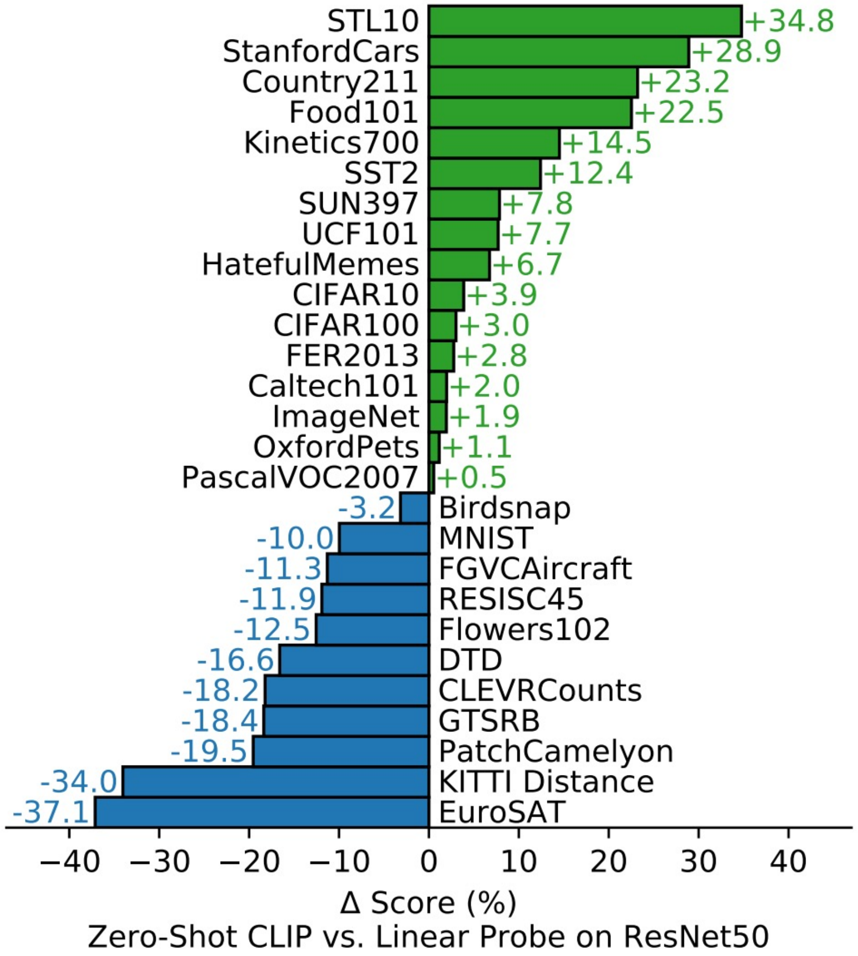
[Radford et al., 2021]

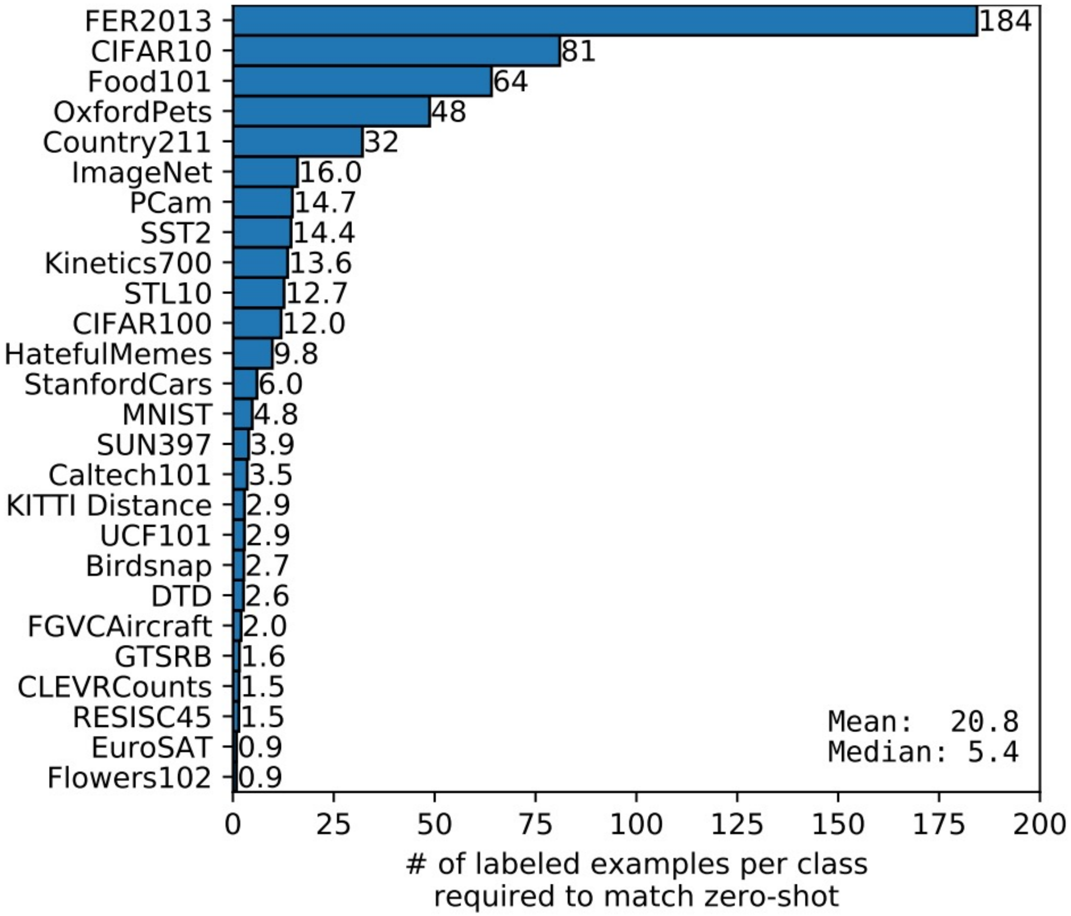
Zero-shot image classification



Learning Transferable Visual Models From Natural Language Supervision

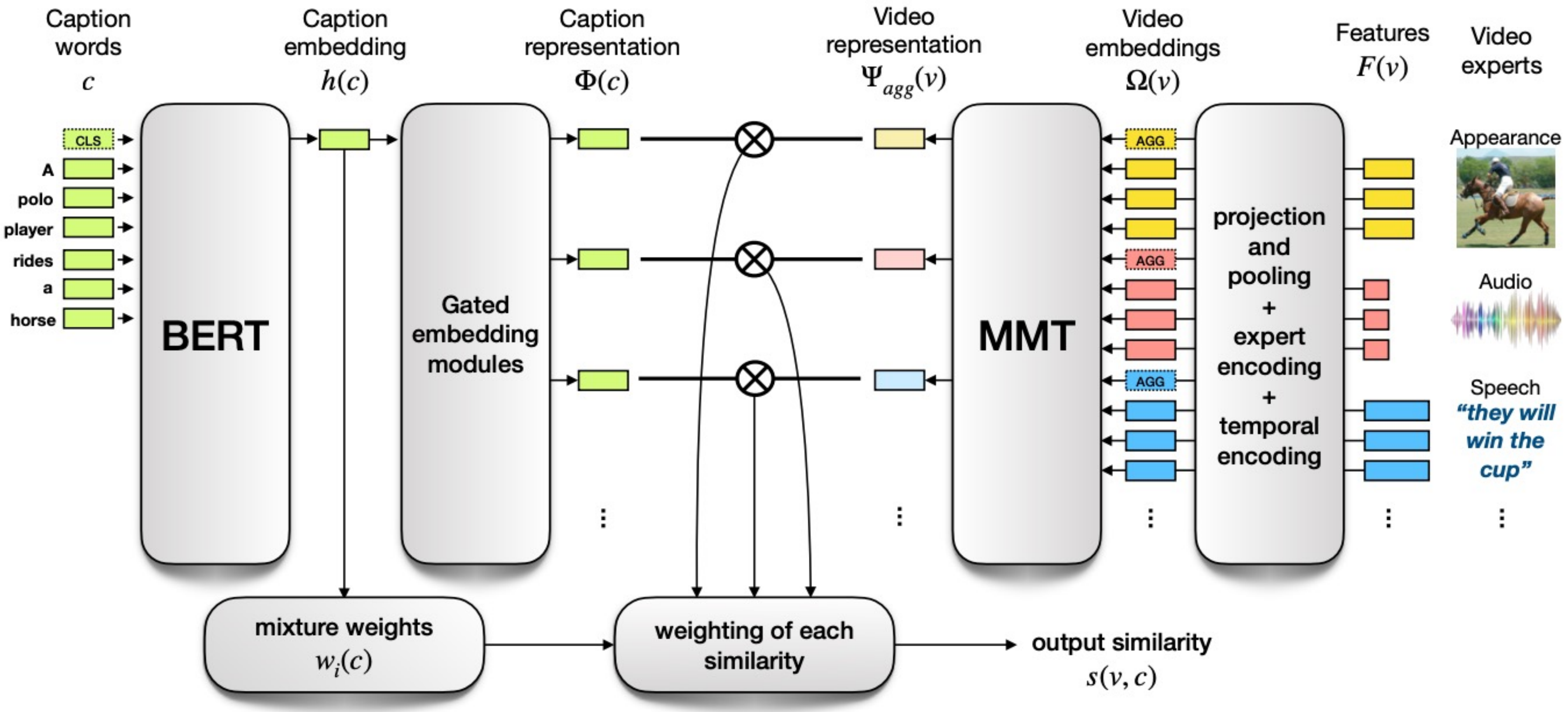
[Radford et al., 2021]





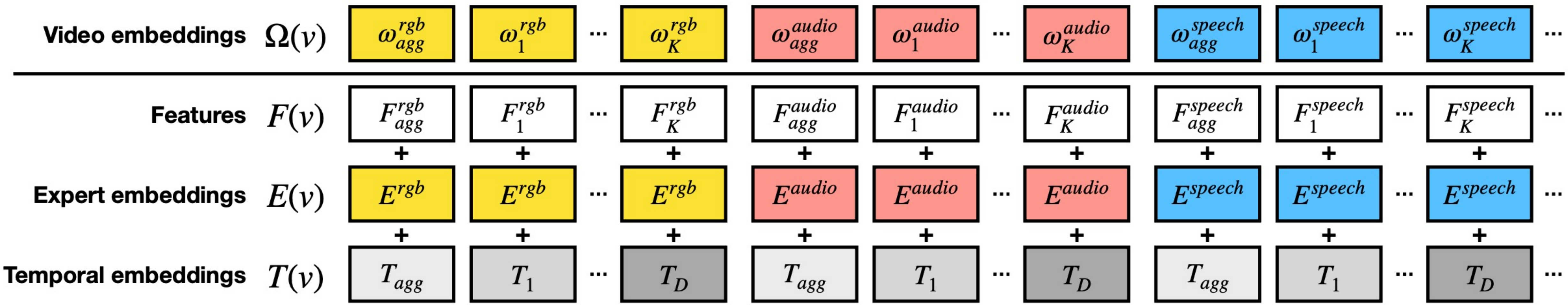
Multi-modal transformer for video retrieval

[Gabeur et al., 2020]



Multi-modal transformer for video retrieval

[Gabeur et al., 2020]

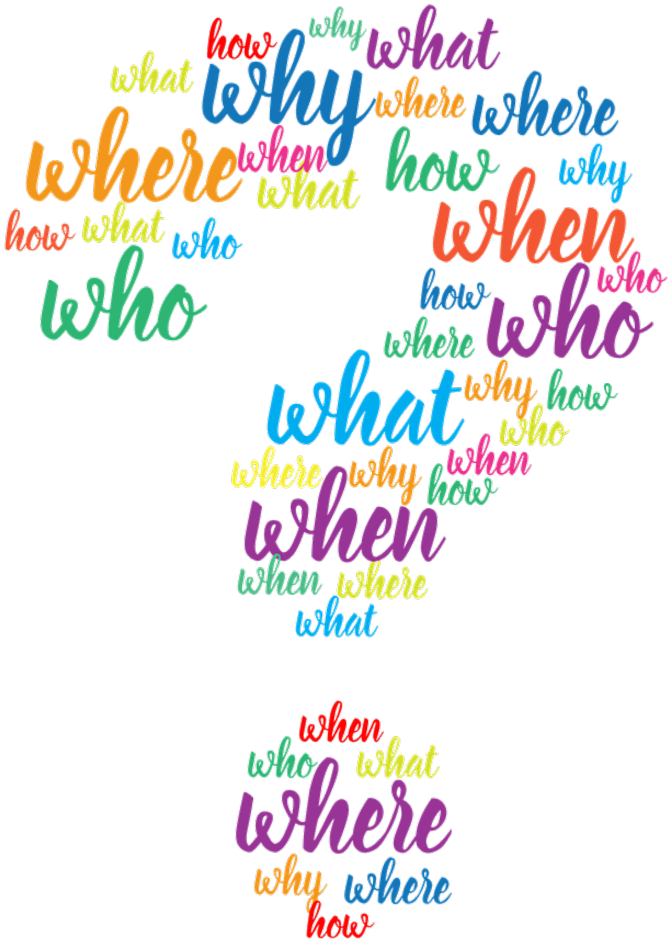


Multi-modal transformer for video retrieval

[Gabeur et al., 2020]

		Text ⟶ Video		
Encoder	Input	R@5↑	MdR↓	MnR↓
COLL	max pool	$51.3_{\pm 0.8}$	$5.0_{\pm 0.0}$	$29.5_{\pm 1.8}$
MMT	max pool	$52.5_{\pm 0.7}$	$5.0_{\pm 0.0}$	$27.2_{\pm 0.7}$
MMT	shuffled feats	$53.3_{\pm 0.2}$	$5.0_{\pm 0.0}$	$27.4_{\pm 0.7}$
MMT	ordered feats	$\mathbf{54.0}_{\pm 0.2}$	$\mathbf{4.0}_{\pm 0.0}$	$\mathbf{26.7}_{\pm 0.9}$

Thank you


how
why
what
what
why
where
where
where
when
what
how
why
how
what
who
when
who
how
who
where
what
why
how
who
where
why
when
how
when
when
where
what
when
who
what
where
why
where
how

References

[Ramachandran et al., 2019] Ramachandran, Prajit, et al. "Stand-Alone Self-Attention in Vision Models." *Advances in Neural Information Processing Systems* 32, 2019.

[Vaswani et al., 2017] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[Dosovitskiy et al., 2020] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *International Conference on Learning Representations*, 2020.

[Touvron et al., 2021] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." *International Conference on Machine Learning*, 2021.

[Chu et al., 2021] Chu, Xiangxiang, et al. "Conditional positional encodings for vision transformers." *arXiv preprint arXiv:2102.10882*, 2021.

References

[Chu et al., 2021] Chu, Xiangxiang, et al. "Do we really need explicit position encodings for vision transformers?." *arXiv preprint arXiv:2102.10882,* 2021.

[Wu et al., 2021] Wu, Haiping, et al. "Cvt: Introducing convolutions to vision transformers." *arXiv preprint arXiv:2103.15808,* 2021.

[Gabeur et al., 2020] Gabeur, Valentin, et al. "Multi-modal transformer for video retrieval." *European Conference,* 2020.

[Radford et al., 2021] Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision" *Image*, vol. 2, 2021.

[Sterpu et al. 2020] Sterpu, George, Christian Saam, and Naomi Harte. "Should we hard-code the recurrence concept or learn it instead? Exploring the Transformer architecture for Audio-Visual Speech Recognition." *arXiv preprint arXiv:2005.09297,* 2020.

References

[Xu et al., 2015] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. PMLR, 2015.

[Wang et al., 2018] Wang, Xiaolong, et al. "Non-local neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[Wang et al., 2020] Wang, Huiyu, et al. "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation." *European Conference on Computer Vision*. Springer, Cham, 2020.

[Bello et al., 2019] Bello, Irwan, et al. "Attention augmented convolutional networks." *International conference on computer vision*. 2019.

[Chen et al., 2018] Chen, Yunpeng, et al. "A^2-Nets: Double Attention Networks." *Advances in Neural Information Processing Systems* 31 (2018): 352-361.