



UNIVERSITY OF TECHNOLOGY
IN THE EUROPEAN CAPITAL OF CULTURE
CHEMNITZ

Neurocomputing

Vision Transformers

Julien Vitay

Professur für Künstliche Intelligenz - Fakultät für Informatik

1 - Vision transformers

Vision transformer (ViT)

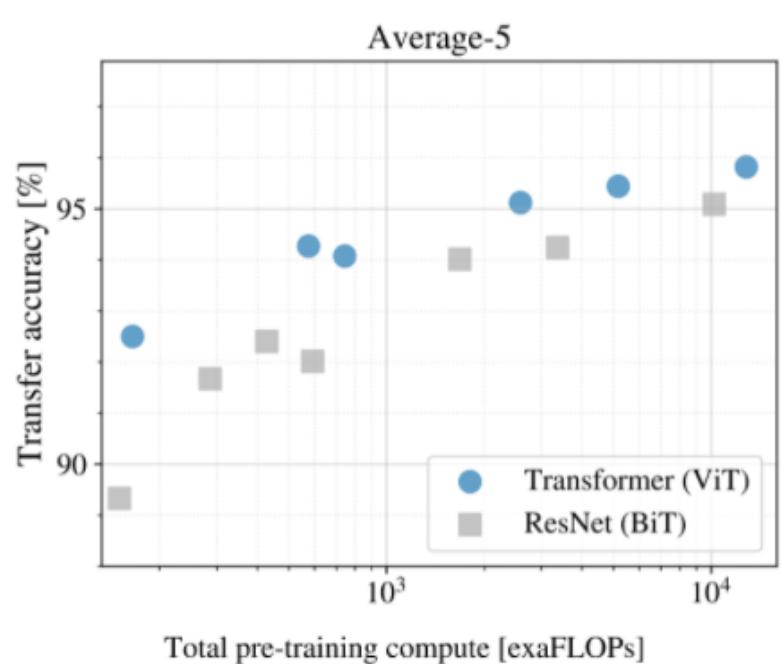
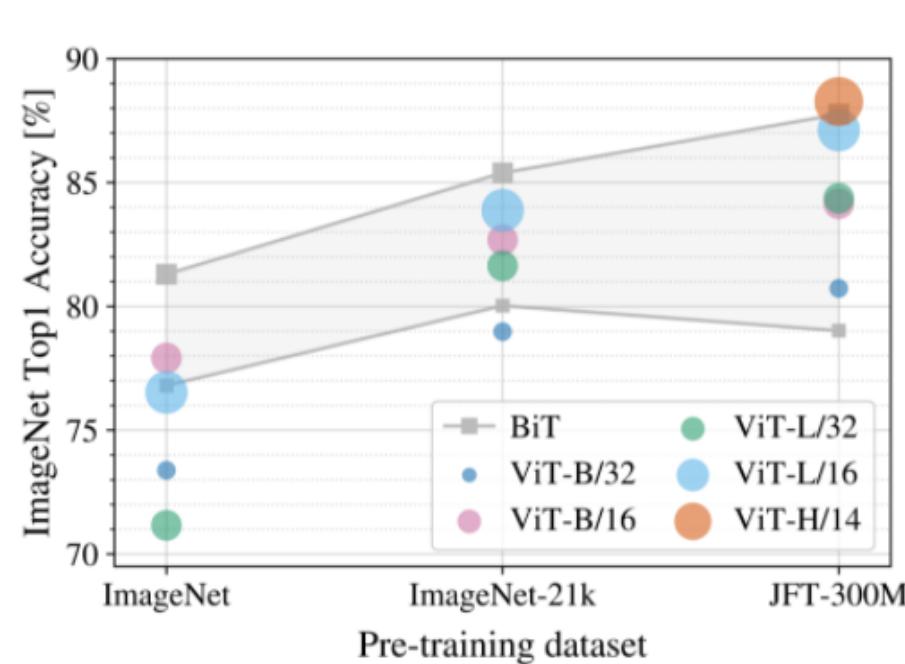
- The transformer architecture can also be applied to computer vision, by splitting images into a **sequence** of small patches (16x16).
- The sequence of patches can then be classified using the first output of the Transformer encoder (BERT) using supervised learning on Imagenet.



Source: <https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html>

Vision transformer (ViT)

- The Vision Transformer (ViT) outperforms state-of-the-art CNNs on Imagenet while requiring less computations (Flops), but only when pretrained on bigger datasets.
- The performance is acceptable when trained on ImageNet (1M images), great when pre-trained on ImageNet-21k (14M images), and state-of-the-art when pre-trained on Google's internal JFT-300M dataset (300M images).
- Transfer learning on smaller datasets is also SotA.



	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

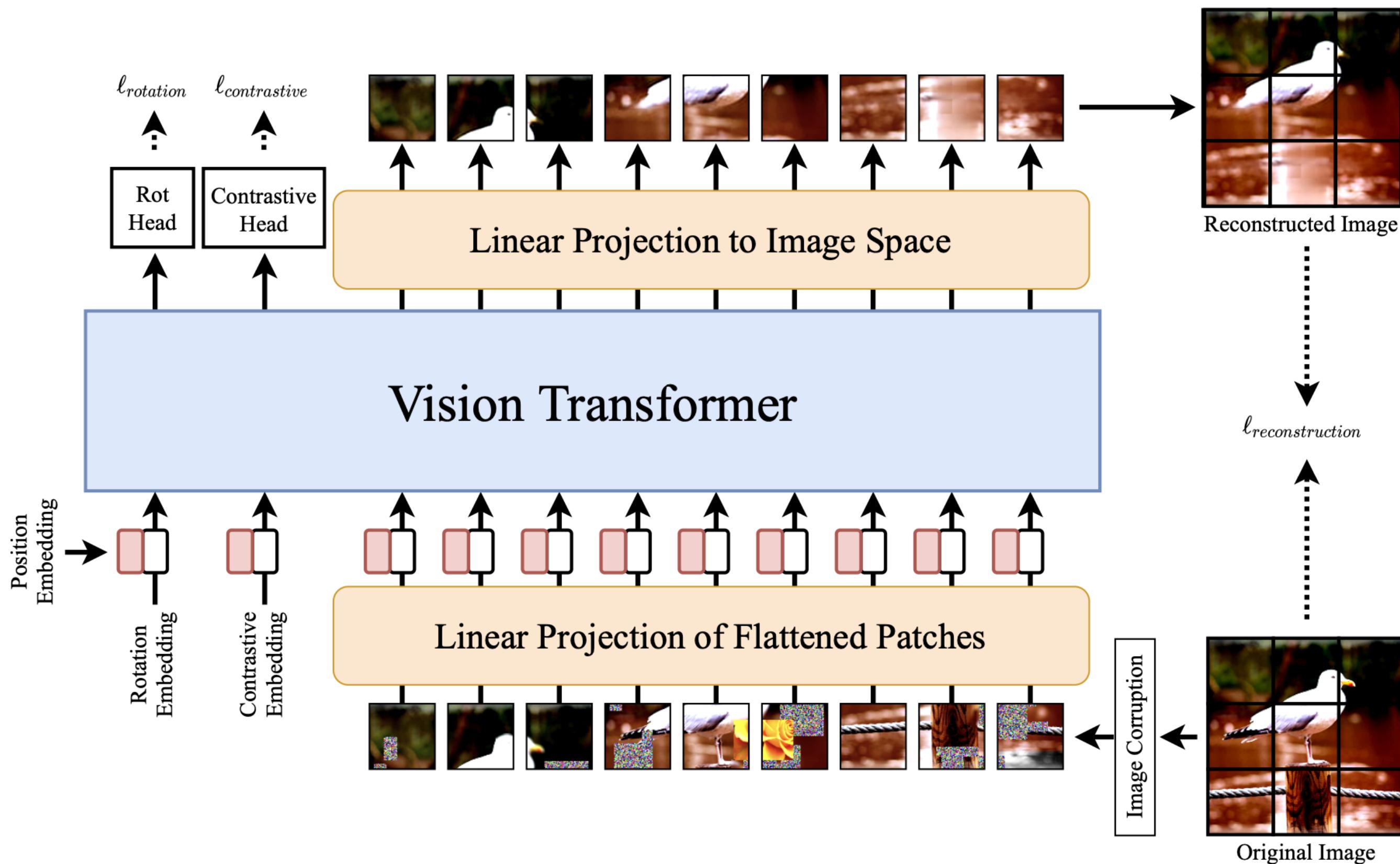
Model	Accuracy [%]
ViT-H/14	~78%
BiT-L (R152x4)	~76%
VIVI-Ex-100% (R50x3)	~70%
S4L (R50x1)	~67%

<https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html>

2 - Self-supervised Vision Transformer

Self-supervised Vision Transformer (SiT)

- ViT only works on big supervised datasets (ImageNet). Can we benefit from self-supervised learning as in BERT or GPT?
- The Self-supervised Vision Transformer (SiT) has an denoising autoencoder-like structure, reconstructing corrupted patches autoregressively.



Self-supervised Vision Transformer (SiT)

- Self-supervised learning is possible through from **data augmentation**, where various corruptions (masking, replacing, color distortion, blurring) are applied to the input image, but SiT must reconstruct the original image (denoising autoencoder, **reconstruction loss**).

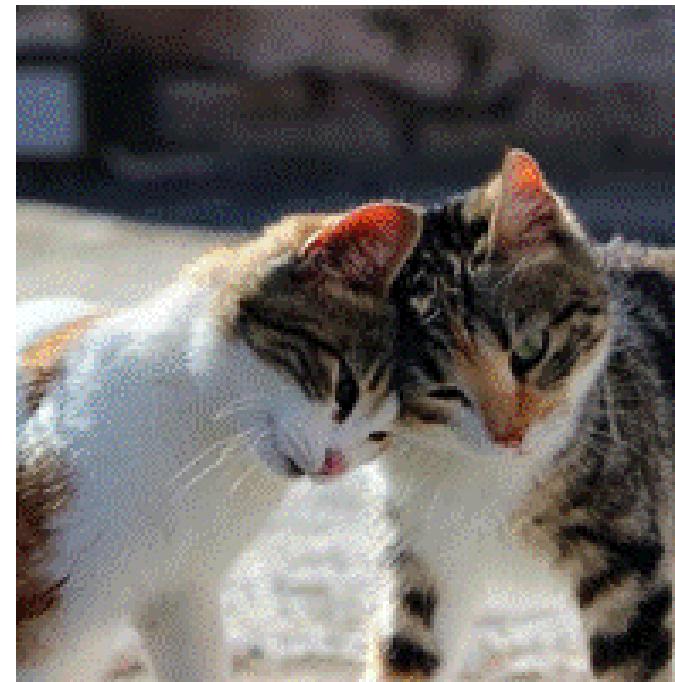


- An auxiliary **rotation loss** forces SiT to predict the orientation of the image (e.g. 30°).
- An auxiliary **contrastive loss** ensures that high-level representations are different for different images.

Method	Backbone	Linear Evaluation			Domain Transfer	
		CIFAR10	CIFAR100	Tiny-ImageNet	C100 → C10	C10 → C100
DeepCluster [19]	ResNet-32	43.31% ± 0.62	20.44% ± 0.80	11.64% ± 0.21	43.39% ± 1.84	18.37% ± 0.41
RotationNet [23]	ResNet-32	62.00% ± 0.79	29.02% ± 0.18	14.73% ± 0.48	52.22% ± 0.70	27.02% ± 0.20
Deep InfoMax [20]	ResNet-32	47.13% ± 0.45	24.07% ± 0.05	17.51% ± 0.15	45.05% ± 0.24	23.73% ± 0.04
SimCLR [8]	ResNet-32	77.02% ± 0.64	42.13% ± 0.35	25.79% ± 0.4	65.59% ± 0.76	36.21% ± 0.16
SimCLR [8]	ResNet-56	78.75% ± 0.24	44.33% ± 0.48	n/a	66.19% ± 0.80	36.79% ± 0.45
Relational Reasoning [21]	ResNet-32	74.99% ± 0.07	46.17% ± 0.16	30.54% ± 0.42	67.81% ± 0.42	41.50% ± 0.35
Relational Reasoning [21]	ResNet-56	77.51% ± 0.00	47.90% ± 0.27	n/a	68.66% ± 0.21	42.19% ± 0.28
SiT (ours) - Linear projection	Transformer	81.98% ± 0.24	54.31% ± 0.13	40.35% ± 0.27	73.79% ± 0.15	55.72% ± 0.13
SiT (ours) - Non-Linear projection	Transformer	83.50% ± 0.11	57.75% ± 0.21	43.06% ± 0.14	75.52% ± 0.11	57.89% ± 0.14

Self-distillation with no labels (DINO)

- Another approach for self-supervised learning has been proposed by Facebook AI using **self-distillation**.
- The images are split into **global** and **local patches** at different scales.
- Global patches contain label-related information (whole objects) while local patches contain finer details.

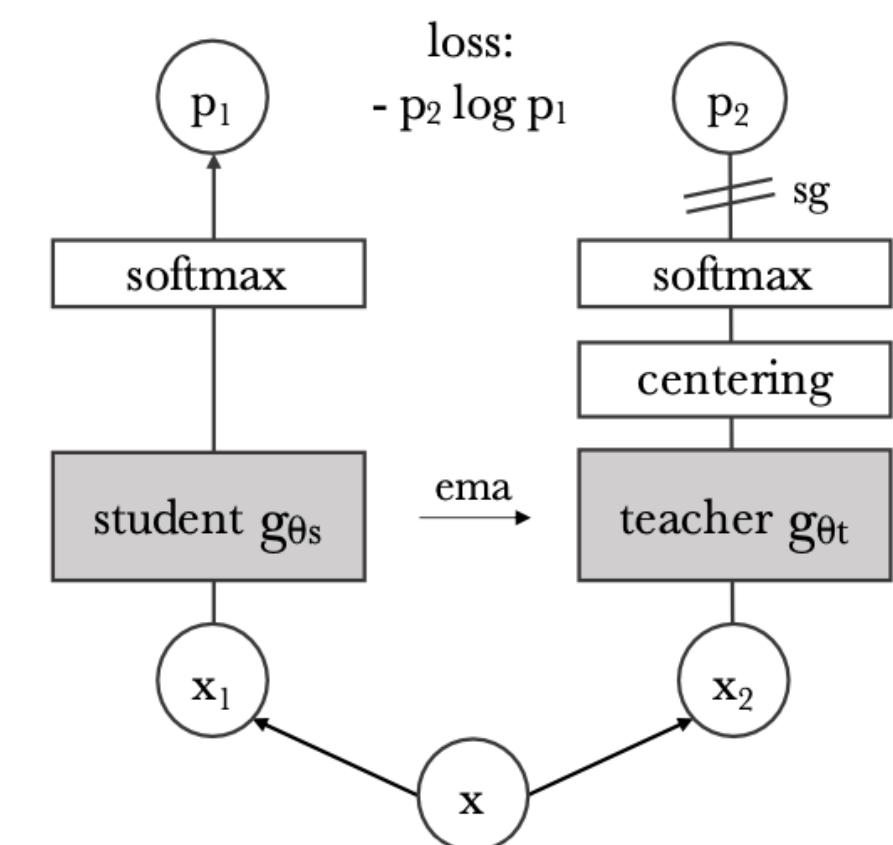


Davide Cacominini | 2021

Source: <https://towardsdatascience.com/on-dino-self-distillation-with-no-labels-c29e9365e382>

Self-distillation with no labels (DINO)

- The idea of **self-distillation** in DINO is to use two similar ViT networks to classify the patches.
- The **teacher** network gets the global views as an input, while the **student** network get both the local and global ones.
- Both have a MLP head to predict the softmax probabilities, but do **not** use any labels.



- The student tries to imitate the output of the teacher, by minimizing the **cross-entropy** (or KL divergence) between the two probability distributions.
- The teacher slowly integrates the weights of the student (momentum or exponentially moving average ema):

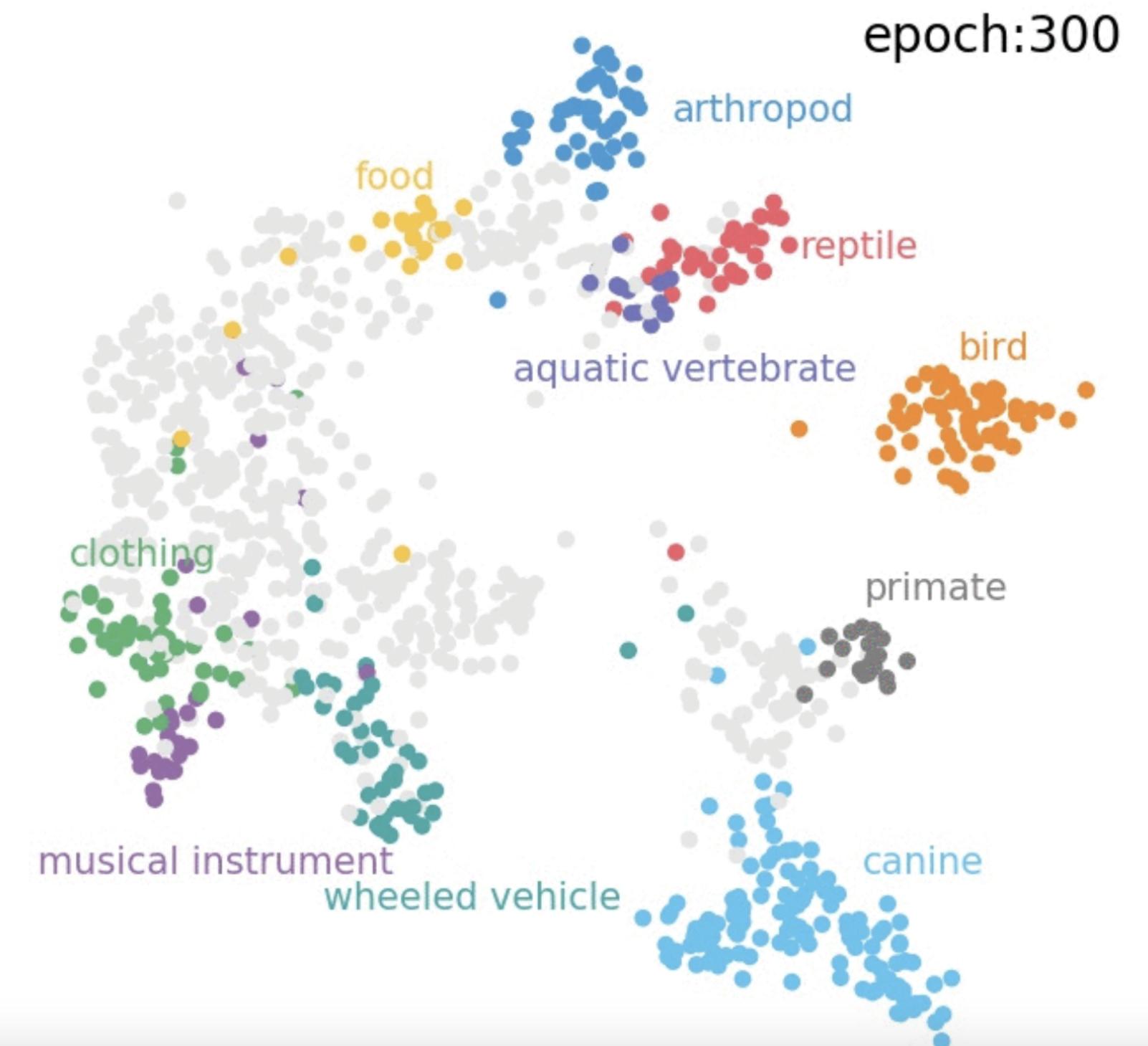
$$\theta_{\text{teacher}} \leftarrow \beta \theta_{\text{teacher}} + (1 - \beta) \theta_{\text{student}}$$

Self-distillation with no labels (DINO)

Source: <https://ai.facebook.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training/>

Self-distillation with no labels (DINO)

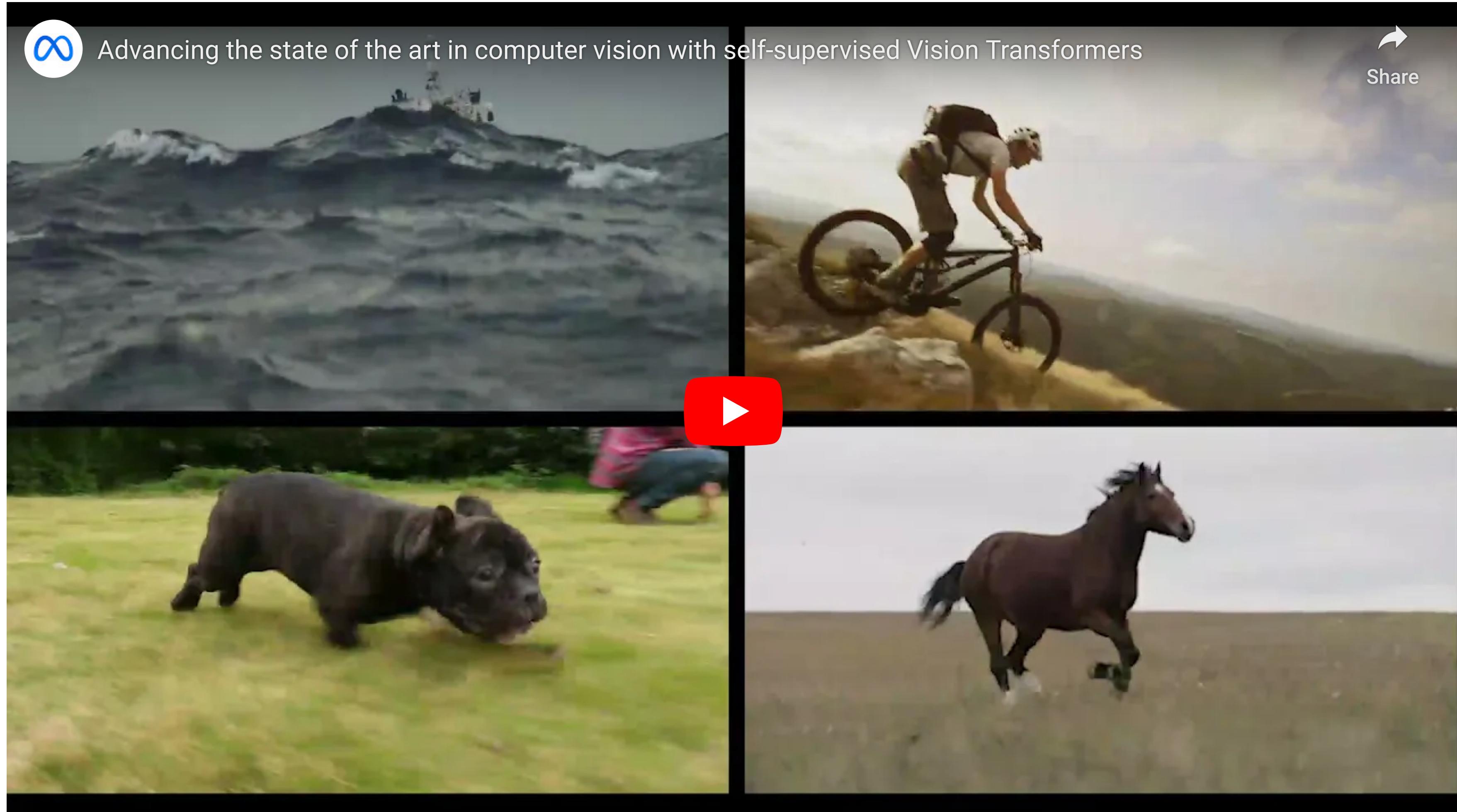
- The predicted classes do not matter when pre-training, as there is no ground truth.
- The only thing that matters is the **high-level representation** of an image before the softmax output, which can be used for transfer learning.
- Self-distillation forces the representations to be meaningful at both the global and local scales, as the teacher gets global views.
- ImageNet classes are already separated in the high-level representations: a simple kNN (k-nearest neighbour) classifier achieves 74.5% accuracy (vs. 79.3% for a supervised ResNet50).



<https://ai.facebook.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training>

Self-distillation with no labels (DINO)

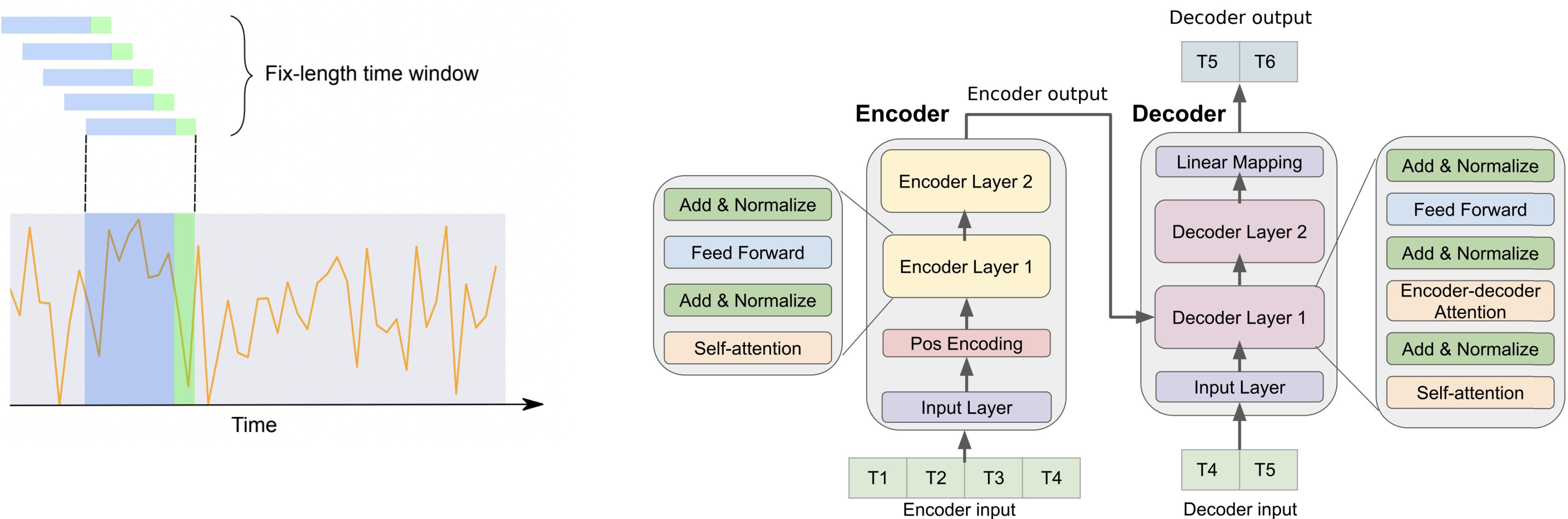
- More interestingly, by looking at the self-attention layers, one can obtain saliency maps that perform **object segmentation** without ever having been trained to!



3 - Other domains

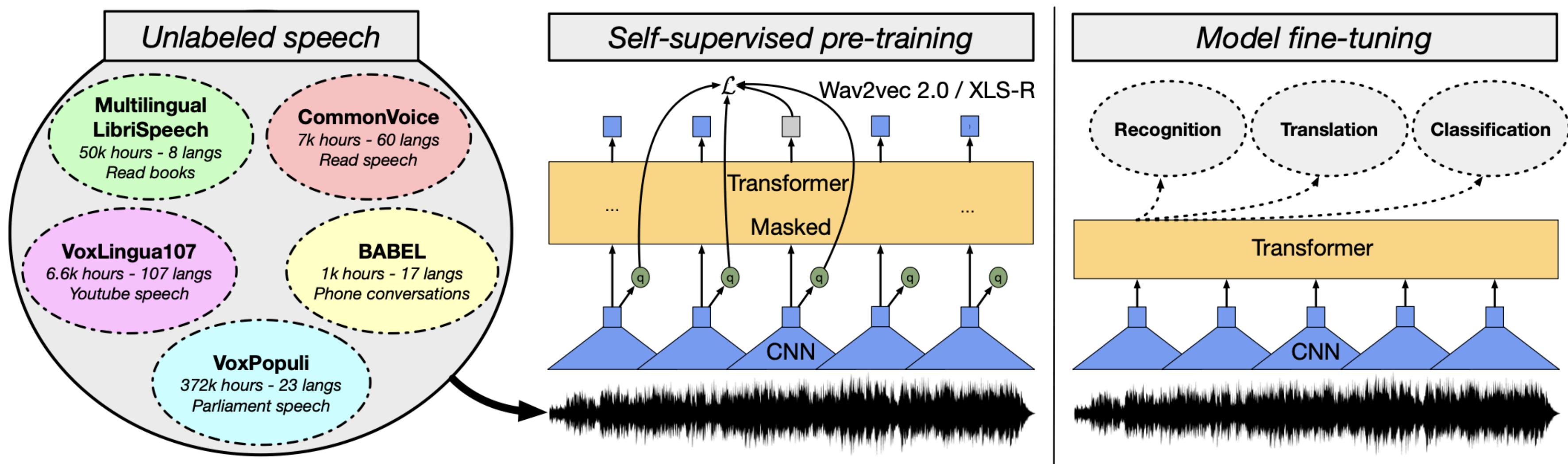
Transformer for time series

- Transformers can also be used for time-series classification or forecasting instead of RNNs.
- Example: weather forecasting, market prices, etc.



Speech processing

- XLS-R from Facebook is a transformer-based architecture trained on 436,000 hours of publicly available speech recordings, from 128 languages.
- Self-supervised: contrastive learning and masked language modelling.
- Other models: UniSpeech, HuBERT, BigSSL...



Source: <https://ai.facebook.com/blog/xls-r-self-supervised-speech-processing-for-128-languages/>

Additional resources

<https://theaisummer.com/vision-transformer/>

<https://theaisummer.com/transformers-computer-vision/>

<https://iaml-it.github.io/posts/2021-04-28-transformers-in-vision/>

https://d2l.ai/chapter_attention-mechanisms-and-transformers/vision-transformer.html