Overview of papers

Table of contents

What is this?

Summaries of some of the papers I read and a serachable format.

Part I

Architectures

1 ViT

Tip

This is an adaptation of transformer models to work with an image data. Transformers lack the inductive biases of CNNs so they need way more data to perform well. Key things for ViT is scaling the dataset size and pre-training.

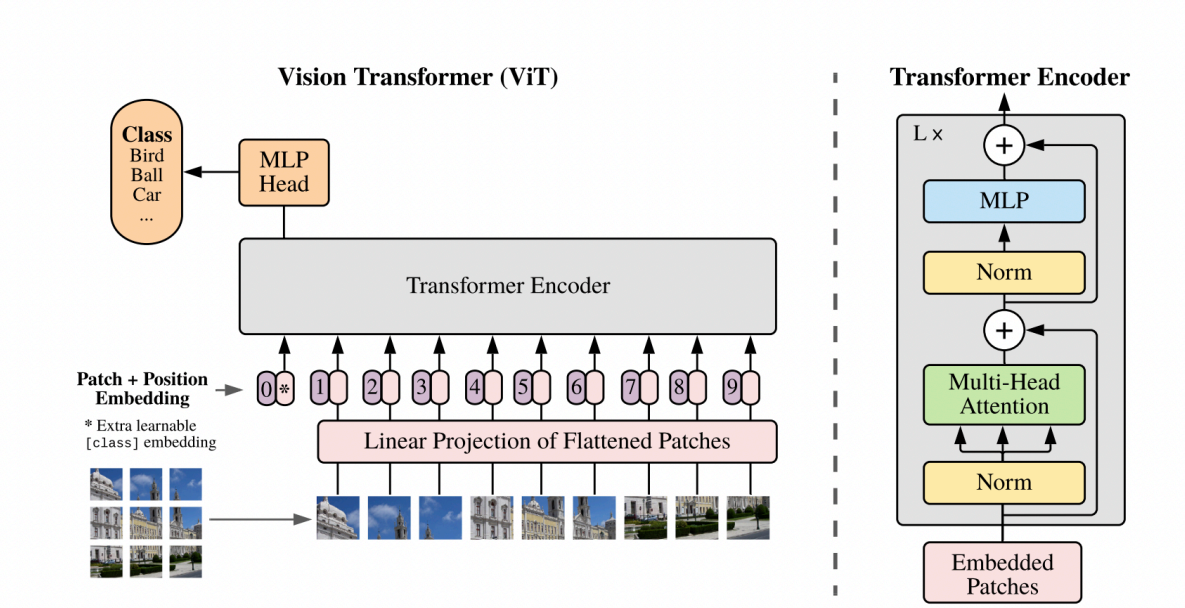


Figure 1.1: ViT Architecture

Using **huge datasets** allows re-learning some of the inductive biases CNNs have. I.e. positional encodings learn image locations even though they are 1D.

Key Points

Why is it a good idea to pre-train vs. use CNNs directly?

- Transformers allow easy multi modalities.
- XLA and other computational infrastructure exists to make transformers very efficient.

What is the difference in the way Transformers process images vs. CNNs?

Transformers are design to work on sets, so to make it work with image we need to somehow convert images to set. To turn sets into sequences we add positional encodings to our inputs so they will have the notion of order.

The way ViT adopt images to be used with transformers is by using 16x16 patches of raw pixels and transforms those with a linear transformation. The transformed vectors are the tokens.

CNNs have a local view of the image, that is their receptive field starts small and grows as we more layers. On the other hand, transformers can see the whole image. Locality in ViT is in the form of image patches, so each token represents some local area, but the attention layer can see all the patches.

Positional embeddings do improve performance, but using 1D is pretty much the same as using 2D. So it seems like the position is not really important, but more like the identity of the patch that matters. Patch distance seem to be encoded in the positional embeddings, also 2D structure seem to emerge from a 1D encoding, this is why hand-crafter 2D encoding doesn't seem to help there.

As we go deeper in attention layers, the attention distance grows. That implies that initial layers learn local features and deeper layers have more complex global representations. This is similar to the way CNNs work.

ViT features vs. CNNS features

ViT features contain semantic features with high spatial resolution. Each attention layers sees full image. CNN features are localized and coarse. Each feature contains global info (e.g. x32 for the last layer of ResNet)

Questions?

- Keys are the most stable representations of an image (**Why?**)

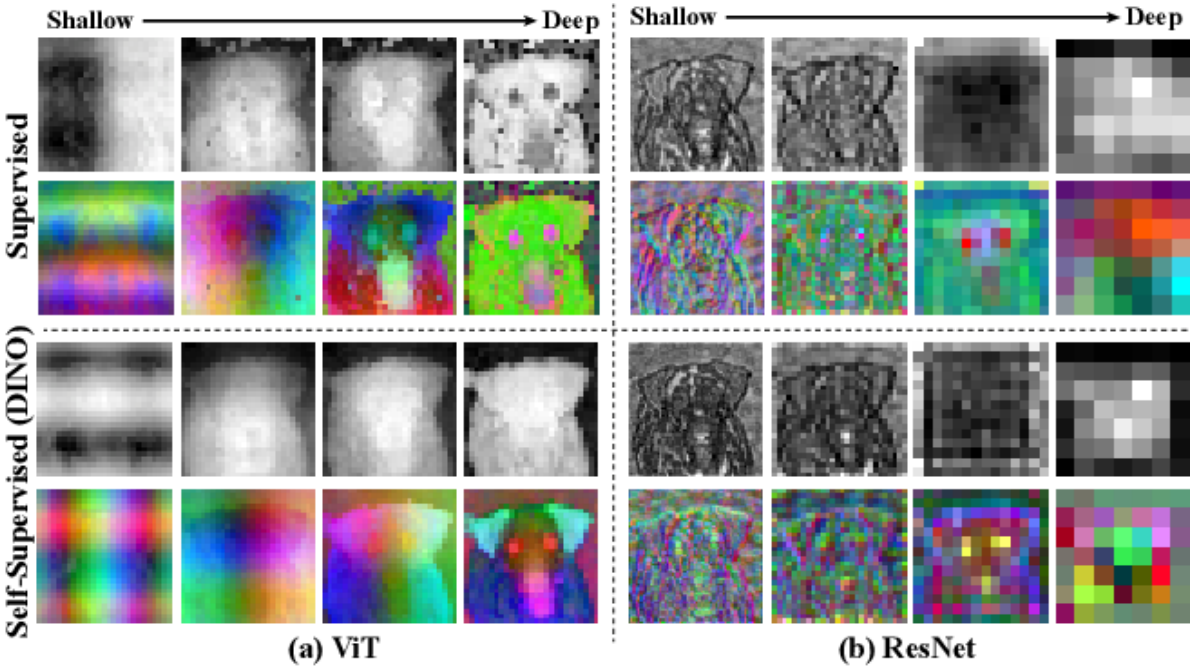


Figure 3: **Deep features visualization via PCA.** Applied

Figure 1.2: ViT vs. CNN Features

Sources

- Dosovitskiy et al. (2020)
- Amir et al. (2021)

Part II

SSL

2 SSL Video Pre-training

TL;DR

- Random crops for video datasets need to be smaller because frame variability is larger than imagenet.
- Replace MoCLR's average-pooling with small mask prediction network and do the pooling with it.

Video should be great for pertaining representation for images, you can see objects change shape and oriantation. For some reason representation pertained on video are not as good as representation pertained on ImageNet. This paper tries to explain why this is the case and proposes ways to close this gap.

Method

`MoCLR` is the baseline contrastive learning method they compare against. The method produces multiple views of the same image, average-pools the feature map of all the views, and passes the averaged pooled feature through an MLP. The loss forces all views have the same reduced feature (trained with contrastive loss). The training is done with two copies of the same network, one being trained by gradient decent and the other by exponential moving average of the other.

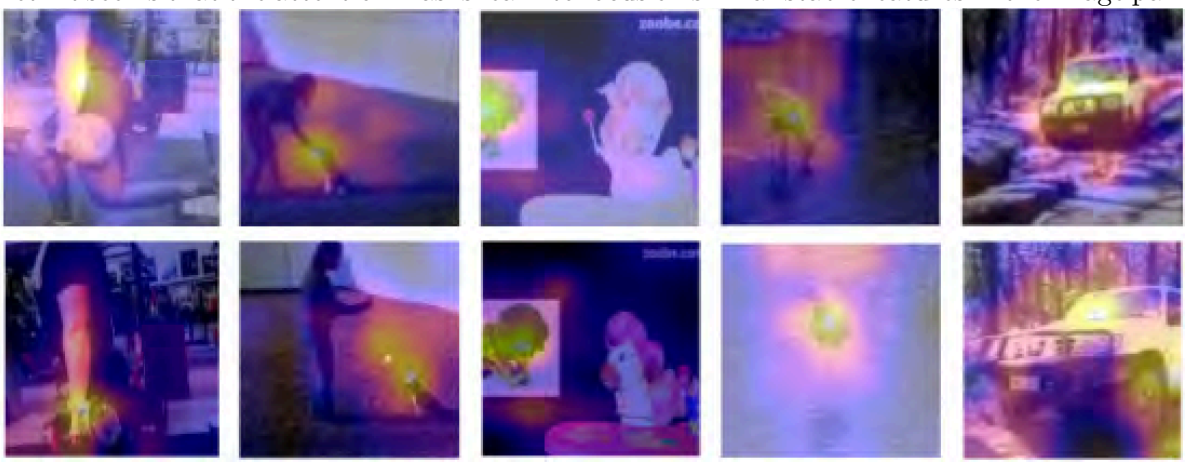
What's new here?

Larger Random Sized Crops

Contrastive learning methods are designed to fit ImageNet like datasets. ImageNet means single object images and lower variability in image content. Lowe variability allows us to get away with aggressive cropping (8% of the original image). In videos (or natural datasets) aggressive cropping can get you a totally different semantic meaning.

Attention Pooling

In videos we can do temporal augmentation (look at nearby frames as being similar). Here average pooling might not be as good. Instead of doing object tracking or something similar, they predict an attention mask and use it to do the pooling. The masking is done on multiple scales of the network and all pooled vectors get concatenated before going through the final MLP net. It seems that the attention masks learn to focus on similar stable features in the image pair.



Curate dataset to match imagenet

Tasks that we use to measure the quality of a representation are imagenet biased so they do some curation to match distribution of videos to that of imagenet images (by running inference on frames and looking for imagenet categories).

Sources

- Parthasarathy et al. (2022)

References

Amir, Shir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. 2021. "Deep ViT Features as Dense Visual Descriptors." *ArXiv* abs/2112.05814.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2020. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv. https://doi.org/10.48550/ARXIV.2010.11929.

Parthasarathy, Nikhil, S. M. Ali Eslami, João Carreira, and Olivier J. H'enaff. 2022. "Self-Supervised Video Pretraining Yields Strong Image Representations." In. https://www.semanticscholar.org/paper/Self-supervised-video-pretraining-yields-strong-Parthasarathy-Eslami/c95527eed7758904823eb43300044fbd0cb1881c.