Vision Transformers: A New Computer Vision Paradigm

Explore and Understand Visual Transformers to perform Image Recognition

* ***Strength and weaknesses of the current Convolutional Neural Network***
* ***How Visual Transformers resolve the weakness in CNN by understanding its working***
* ***Different variations of Visual Transformers.***

Strength and Weaknesses with Convolutional Neural Network(CNN)

**Strength**

* **CNN has inductive biases like translation invariance and locally restricted receptive field**. Translational invariance recognizes an object even when its appearance varies, like changing orientation or zoom in or zoom out.

**Weakness**

* **Specifically designed for images.**
* It has a **domain-specific design and is not scalable for it to be domain agnostic.**CNN uses pixel arrays where each pixel represents varying importance and is domain-specific.
* **CNN lacks a global understanding of the images.**Itonly looks for the presence of the image's features and does not understand the structural dependency between its features.
* **Computationally expensive**: each pixel bears varying importance for the target task, which causes redundancy in both computation and representations.

Visual Transformer

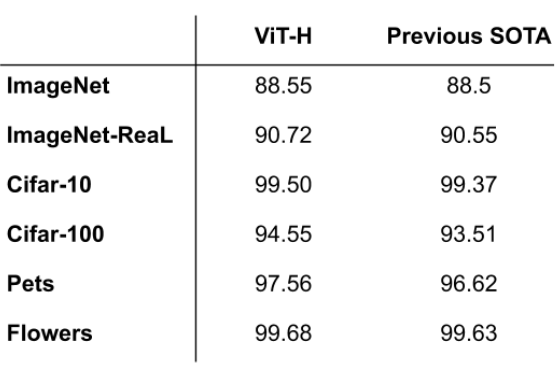
Transformers have great success with NLP and are now applied to images. CNN uses pixel arrays, whereas Visual Transformer(ViT) divides the image into visual tokens.

***ViT splits an image into fixed-size patches, linearly embed each of them, add positional embedding as an input to Transformer Encoder.***

ViT outperforms the state of the art CNN in terms of accuracy and also computational efficiency.

***When ViT is trained on sufficient data it outperforms the state-of-the-art CNN by about four times fewer computational resources***



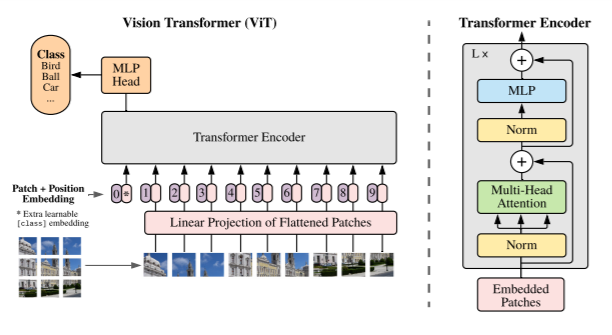


**Accuracy performance between ViT and SOTA CNN on different datasets**. Source: [Transformers for Image Recognition at Scale](http://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html)-by Neil Houlsby and Dirk Weissenborn, Research Scientists, Google Research

***Self-attention layer in ViT allows it to integrate information globally across the entire image. ViT learns to encode the relative location of the patches to reconstruct the image structure from the training data***

Working of Visual Transformer





Source: [AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE](https://arxiv.org/pdf/2010.11929.pdf)

* **Image is first split into fixed-size patches.**

The 2D image of size H \*W is split into N patches where N=H\*W/P²

If the image is of size 48 by 48 and the patch size is 16 by 16, then there will be 9 patches for the image.

The cost of self-attention is quadratic. If we pass each pixel of the image as input, then self-attention would require each pixel to attend to every other pixel. The quadratic cost of the self-attention will be very costly and not scale to realistic input size; hence, the image is divided into patches.

* **Flatten the 2D patches to 1D patch embedding and linearly embed them**

Each patch is flattened into a 1D patch embedding by concatenating all pixel channels in a patch and then linearly projecting it to the desired input dimension.

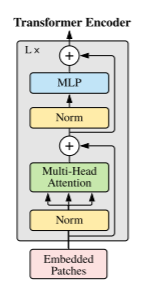
* **Position embeddings are added to the patch embeddings to retain positional information.**

Transformers are agnostic to the structure of the input elements. Adding the learnable position embeddings to each patch will allow the model to learn about the structure of the image.

**We add an extra learnable “classification token” to the patch embedding at the start of the sequence.**

**This sequence of patch embedding vectors will be used as an input sequence length for the Transformer Encoder.**

Transformer Encoder



The Transformer Encoder consists of

* **Multi-Head Self Attention Layer(MSP)** to concatenate the multiple attention outputs linearly to expected dimensions. The **multiple attention heads help learn local and global dependencies in the image.**
* **Multi-Layer Perceptrons(MLP) contains two-layer with Gaussian Error Linear Unit(GELU)**
* **Layer Norm(LN) is applied before every block** as it does not introduce any new dependencies between the training images. **Help improve the training time and generalization performance.**
* **Residual connections are applied after every block as they allow the gradients to flow through the network directly without passing through non-linear activations.**

For image classification**, a classification head is implemented using MLP with one hidden layer at pre-training time and a single linear layer for fine-tuning.**

The higher layers of ViT learn the global features, whereas the lower layers learn both global and local features. This allows ViT to learn more generic patterns.

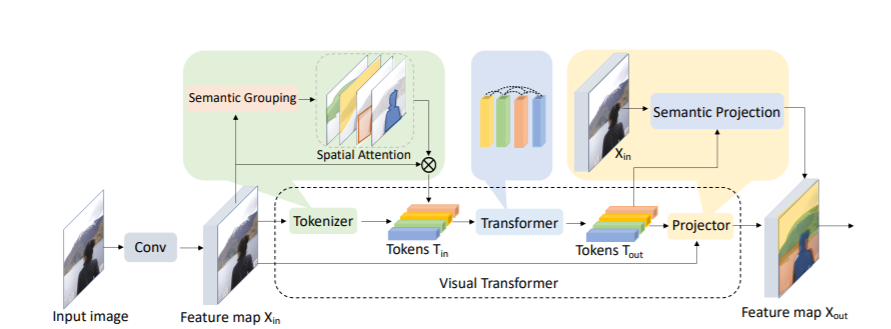
ViT Pre-training and fine-tuning

ViT is pre-trained on large datasets and finetuned to a smaller dataset.

* When fine-tuning, the last pre-trained prediction head is removed, and we attach a zero-initialized feed-forward layer to predict the classes based on the smaller dataset.
* Fine-tuning can be applied to a higher resolution image than what the model was pre-trained on, but the patch size should remain the same.

***Transformers has no prior knowledge about the image structure and hence have longer training times, and require large datasets for training the model.***

A combination of CNN and Visual Transformer



Source: [Visual Transformers: Token-based Image Representation and Processing for Computer Vision](https://arxiv.org/pdf/2006.03677.pdf)

In this model, **the CNN is used to extract the low-level features in an image, and ViT is used for relating high-level concepts.**

The low-level features from CNN are fed to the ViT. Visual Transformer pays attention to only important regions, encodes the semantic concepts into few visual tokens by relating the spatially distant concepts using self-attention. These **visual tokens can be used for image classification or projected back to the feature map for semantic segmentation.**

Conclusion:

ViT uses multi-head self-attention in computer vision removing image-specific inductive biases. ViT divides the images into a sequence of positional embedding patches processed by a Transformer Encoder to understand the local and global features present in the image. ViT has higher accuracy on a sustainably large dataset with reduced training time.

References:

[AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE](https://arxiv.org/pdf/2010.11929.pdf).

[Visual Transformers: Token-based Image Representation and Processing for Computer Vision](https://arxiv.org/pdf/2006.03677.pdf)

[Transformers for Image Recognition at Scale](http://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html)