An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale

Presenter: Gunjan Bhattarai

September 14, 2021

Problem Setting

- Self-attention based architectures (ex: Transformers) have become SOTA in NLP and ASR and can scale up to 100+ billion parameters
- Unfortunately, neither ResNet-style CNNs (the previous Vision SOTA) nor CNNs combined with self-attention scale as well
- The authors experiment with applying a standard Transformer directly to images with the fewest possible modifications

Motivation: Why Do We Care?

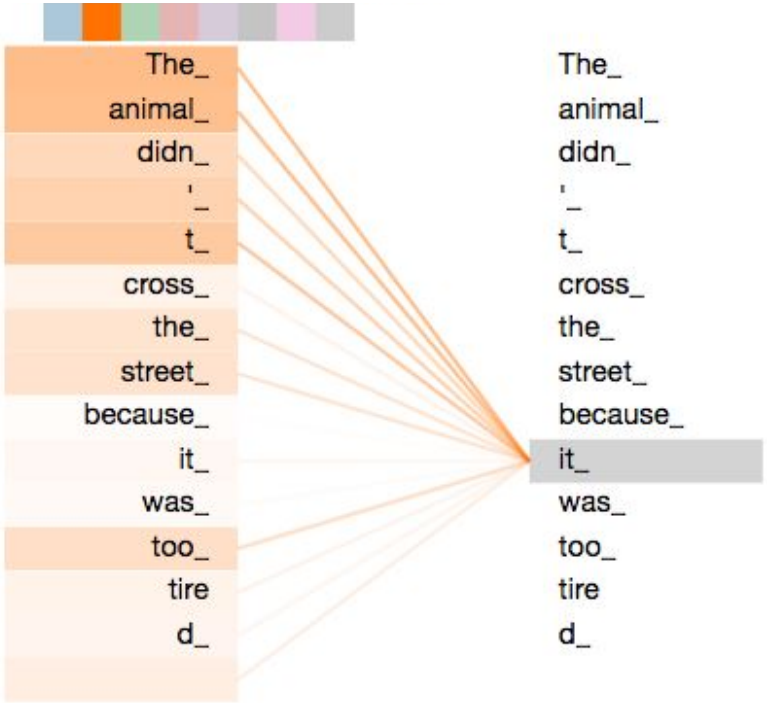
- Robots need to be able to perceive their surroundings; ex: What objects are around them?
- To do this, they will need human-level perception capabilities at the lowest possible computational cost
- In NLP and ASR, Transformers have a lower computational cost than CNNs for the same level of performance and have thus replaced them. Can the same be true for vision tasks like object detection?

Related Work: Attention

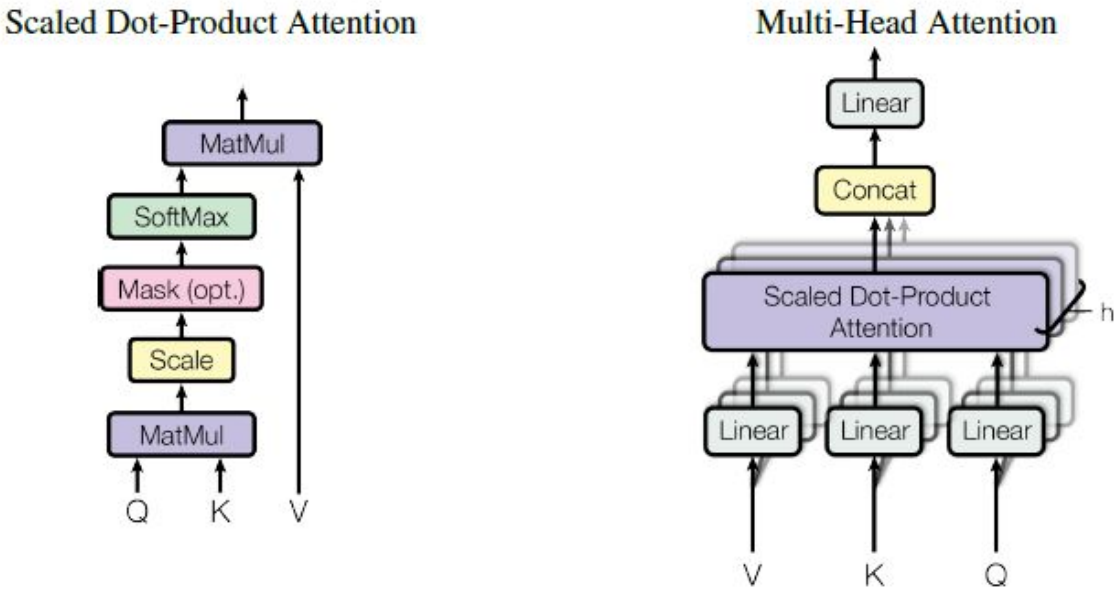
- English: “I have a dog”, Spanish: “Yo tengo un perro”
- Keys: Our inputs -> “I have a dog”
- Queries: Our outputs -> “Yo tengo un perro”
- Values: The combination of our inputs that equal the outputs
- Trained with a feedforward network using backpropagation

Bahdanau et al., 2014

Related Work: Attention



Related Work: Scaled-Dot Product Attention



Vaswani et al., 2017

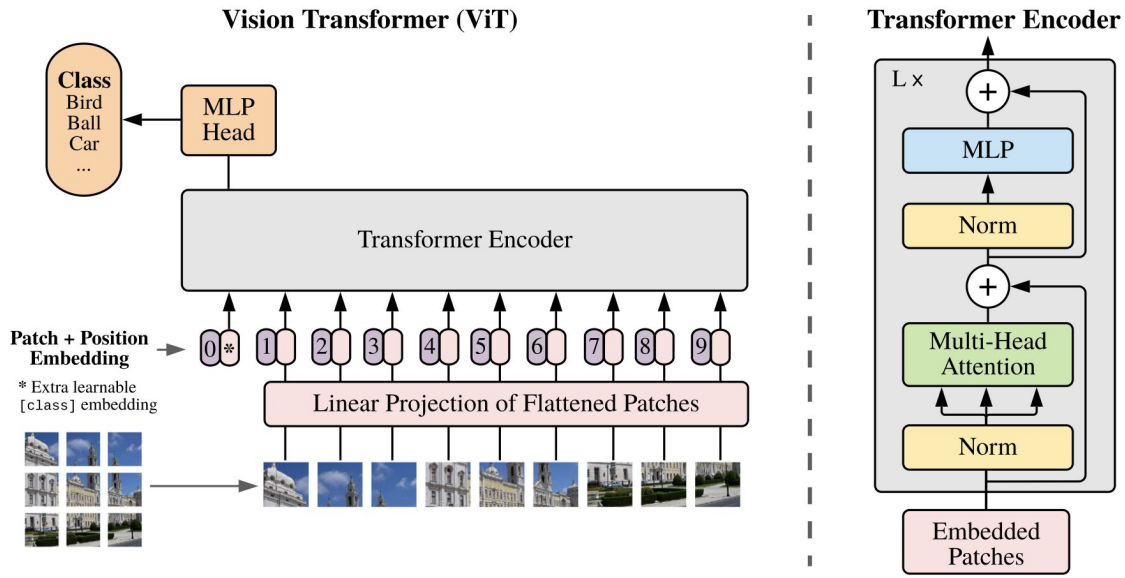
Related Work: Multi-Head Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- Q = Queries, K = Keys, V = Values
- d_k is the dimension of the key matrix. This is used for normalization to minimize the chance of vanishing/exploding gradients.
- Multiple heads are used to reduce peaking and to capture additional distributions (e.g. local vs global self-attention)

Vaswani et al., 2017

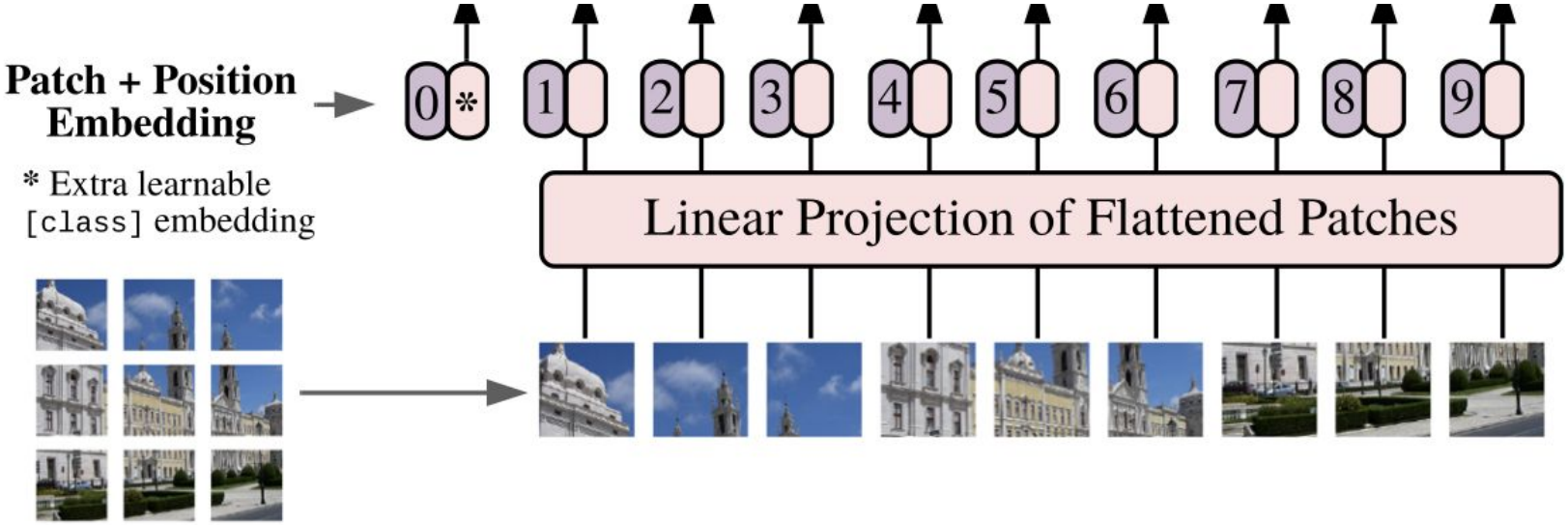
Context: A Self-Attention Network (Transformer)



Proposed Approach: Vision Transformer

- How is this different from a standard Transformer?
  - First, we convert H x W x C dimensional images into N x (P^2 x C)
    - H x W is resolution of the original image, C is the number of channels, P x P is the resolution of the patch, N is the resultant number of patches
  - We have N patches because self-attention is quadratic -> previous papers **struggled to make images work** with this **due to computational cost**

Proposed Approach: Vision Transformer

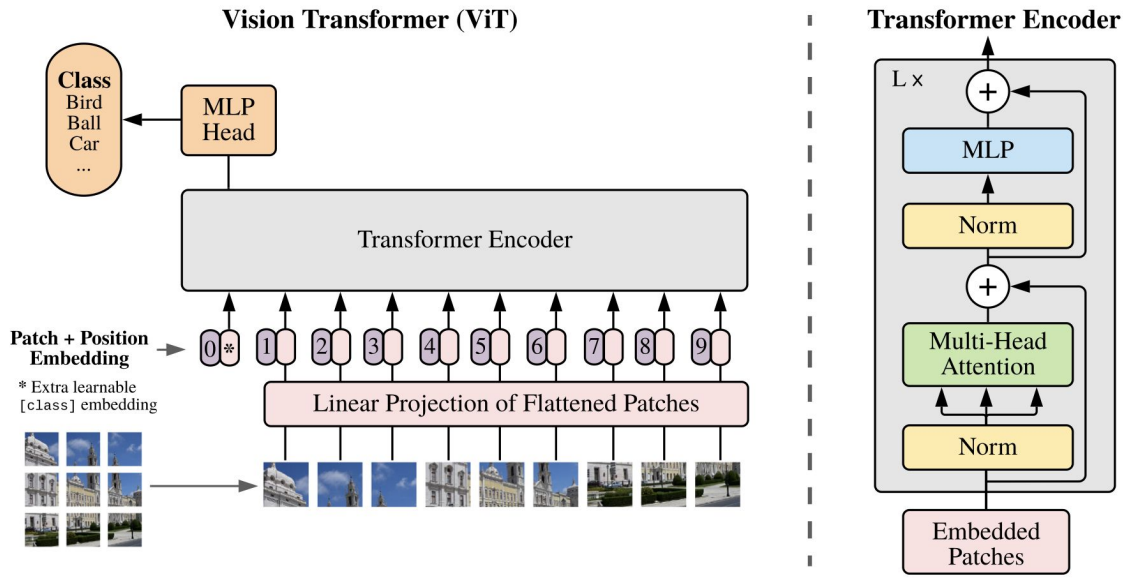


Proposed Approach: Vision Transformer

- We flatten the patches into D dimensions to get our embeddings
  - Alternatively, can use a ResNet (CNN + skip connections) to generate features (hybrid setup)
- We use **supervised pre-training** (classification) over semi-supervised training. Unlike in NLP and speech, supervised pre-training performs better (~84% vs ~80% in ViT-B/16 trained on ImageNet)
- During fine-tuning, we replace the classification head with a zero-initialized linear layer. We ultimately optimize for Top1 ImageNet accuracy.

He et al., 2015

Proposed Approach: A Self-Attention Network (Transformer)



Experimental Setup: ViT Training Details

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

- Pre-training: Adam optimizer, batch size 4096, weight decay 0.1
- Fine-tuning: SGD with momentum, batch size 512
- Baselines: ResNet 152x4, EfficientNet-L2
- Metrics: ImageNet Top-1 Accuracy

Experimental Results

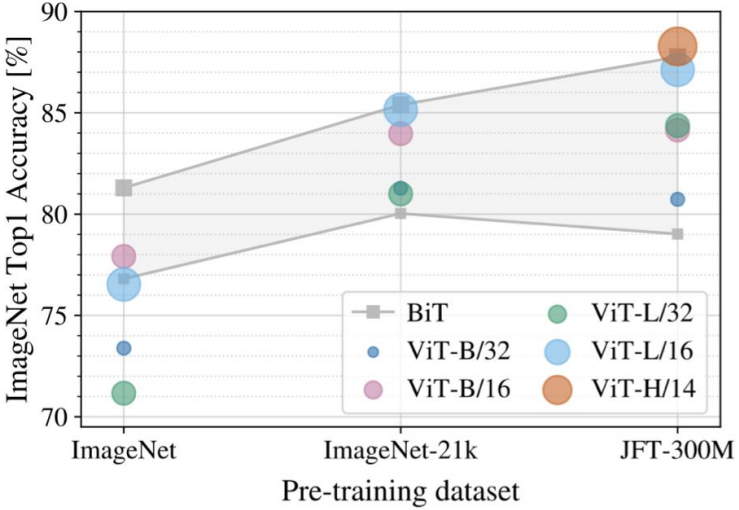
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

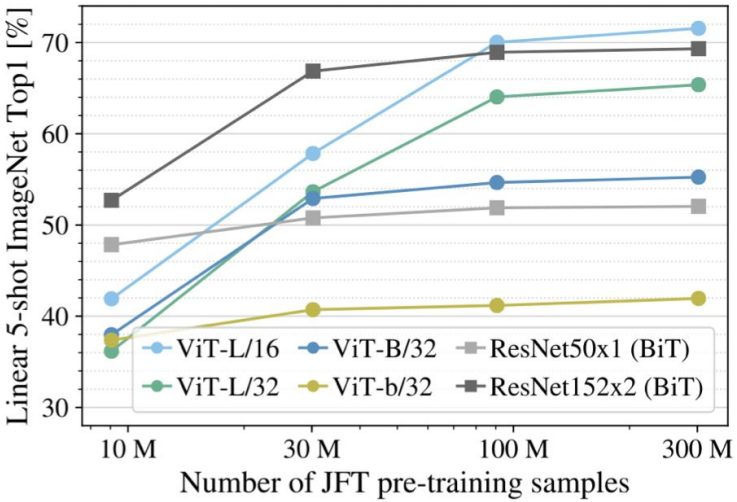
Discussion of Results

	Ours-JFT (ViT-H/14)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)
ImageNet	**88.55** ± 0.04	85.30 ± 0.02	87.54 ± 0.02
ImageNet ReaL	**90.72** ± 0.05	88.62 ± 0.05	90.54
CIFAR-10	**99.50** ± 0.06	99.15 ± 0.03	99.37 ± 0.06
CIFAR-100	**94.55** ± 0.04	93.25 ± 0.05	93.51 ± 0.08
Oxford-IIIT Pets	**97.56** ± 0.03	94.67 ± 0.15	96.62 ± 0.23
Oxford Flowers-102	99.68 ± 0.02	99.61 ± 0.02	99.63 ± 0.03
VTAB (19 tasks)	**77.63** ± 0.23	72.72 ± 0.21	76.29 ± 1.70
TPUv3-core-days	2.5k	0.23k	9.9k

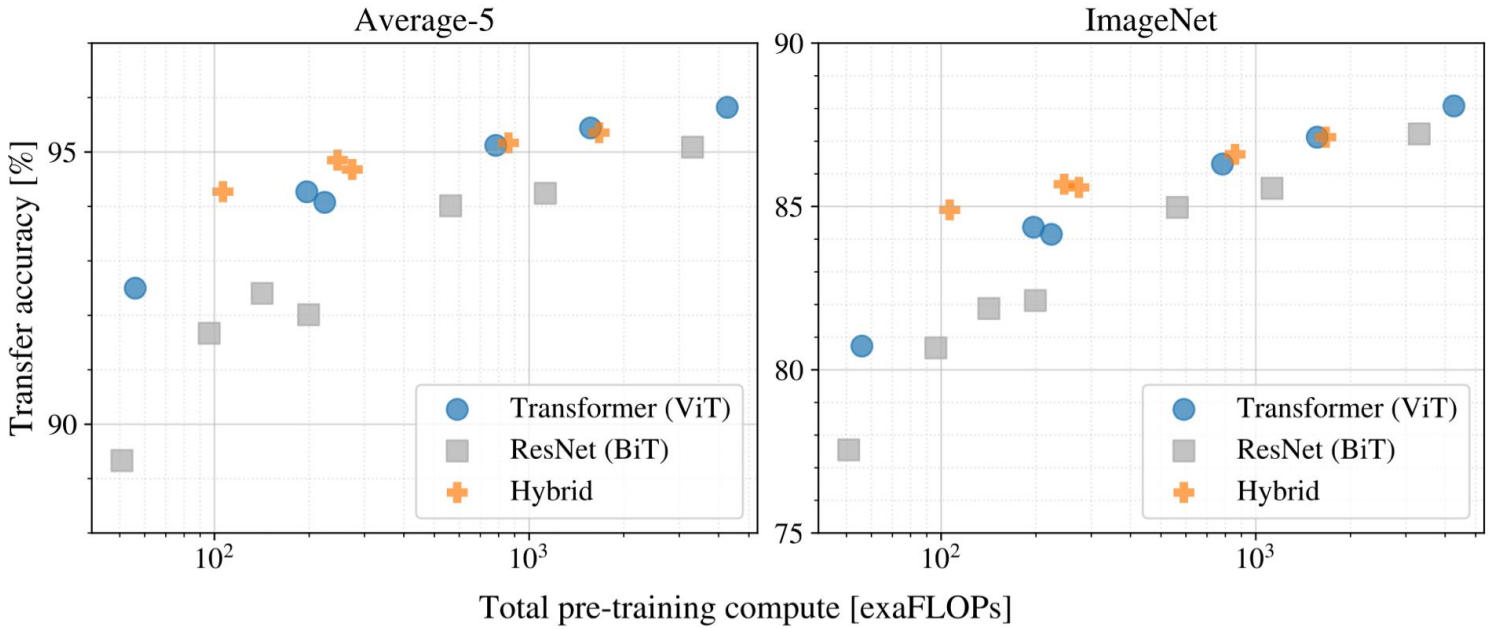
- ViT is far more computationally efficient than ResNet and EfficientNet (in terms of FLOPs and memory), but you still shouldn't be training this at home
- When pre-trained on large datasets (e.g. JFT-300M), ViT is the new state-of-the-art. This is due to the **higher receptive field of self-attention** vs convolutions.
- On the other hand, it trails CNN-based models when pre-trained on smaller datasets due to lack of **inductive bias**

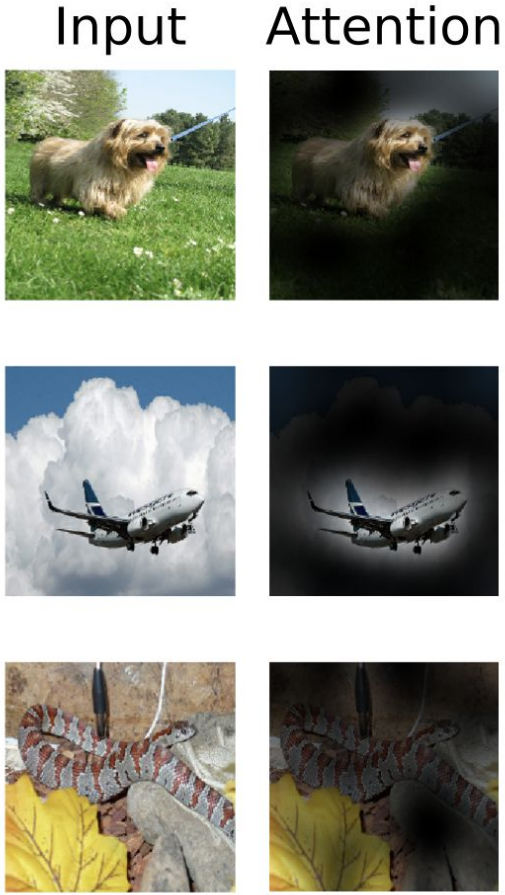
More Experimental Results





More Experimental Results





More Discussion of Results

- Position embeddings are fairly good at learning what part of the image a patch is in despite not being given this information beforehand
- Attention becomes effective very early on -> some heads attend to most of the image already in the lowest layers, although the highly localized nature decreases in hybrid models that apply a ResNet first
- Model ultimately attends to image regions semantically relevant for classification

Limitations

- Hard to reproduce from scratch: JFT-300 is a private Google dataset and ImageNet is too small to train ViT from a randomly initialized checkpoint
- O(n^2) self-attention layer will punish models inputting a large amount of patches (major problem for object detection and segmentation)
  - Paper showed this started to be a problem at size 384 x 384 for larger models (also when ViT inference times began to diverge from ResNet)
  - Solutions exist to this in NLP, ex: Longformer (https://arxiv.org/pdf/2004.05150.pdf) and FNet (https://arxiv.org/pdf/2105.03824.pdf), but we'll have to wait for Brain/FAIR/MSR/etc. to train us a model with these benefits

Future Work/Challenges

- At the time of writing, ViT was only applied to classification and not other computer vision tasks like detection and segmentation
- Are there other ways to make self-supervised pre-training work? No real reason for vision to be unique here.
- Further scaling of the model to billions of parameters like its NLP brethren GPT-3, T5, and the Switch Transformer

Extended Readings

- ConViT: Improving Vision Transformers With Transformer With Soft Convolutional Inductive Biases: https://arxiv.org/pdf/2103.10697.pdf
- Vision Transformers for Dense Prediction: https://arxiv.org/pdf/2103.13413.pdf
- Towards Transformer-Based Object Detection: https://arxiv.org/pdf/2012.09958.pdf
- Transformers in Vision: A Survey: https://arxiv.org/pdf/2101.01169.pdf

Summary

- Problem: Can Transformers be directly applied to image recognition? If so, how do they perform compared to traditional CNNs?
- Previous Limitations: Neither vanilla CNNs nor CNNs with self-attention scale as well as Transformers do in NLP and ASR
- Importance: If this experiment succeeds, we can get the same (or better) performance at lower computational cost - instrumental for mobile robots
- Key Insights: ViT is more computationally efficient for the same results than ResNet, but require large datasets to become state-of-the-art
- Demonstrated a new SOTA in image classification