Title: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
arxiv.org/abs/2010.11929
Authors:Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby
Field: Computer Vision
Submitted: Oct 22

Model: Transformer encoder >> MLP(s) head (Classification)
Split image into sequence of fixed size image patches (14^2 or 16^2). Add positional embedding + Linearly embed patches.
Linear embedding of patches: Unroll patch, multiply by an embedding matrix.

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

Method:
Can pretrain a-la BERT. Authors pretrain on lower resolution image, fine tune on higher resolution image.

Input sequence $[x_{numclass}, Ex_1, Ex_2, , Ex_n]$ $+E_{pos}$,

$x_{numclass}$ is a token that represents the number of classes. (A token, not an integer)
$E \in M(p^2, D)$ maps to the latent (embedding) dimension D.

Pretraining and fine-tuning:
During fine tuning, the prediction head is replaced by a $D/timesK$ feed forward layer. When fine tuning, patch sizes are kept the same, input images are possibly of higher resolution; the positional embedding needs to be interpolated.

The classification head is a 2 layer MLP during pre-training and single layer perceptron during fine-tuning.

Results:
Table 2 in paper;
The ViT were pre-trained on JFT-300 (and 121K) datasets
The tasks are ImageNet,ImageNet Real, CIFAR10-100, Oxford IIIT pet,Oxford Flowers-102 ,VTAB.

ViT-L (16) matches BiT(BigTransfer-https://arxiv.org/abs/1912.11370) ResNet154x4's performance for
Around 1/15th the TPU core days.

ViT-H(14) surpassed BiT ResNet154x4 by around 1% across all tasks for 1/4th the TPU core days.

ViT-H(14) matches Noisy Student(EfficientNetL2-https://arxiv.org/abs/1911.04252) for around 1/5th the TPU core days.