- Fudong Lin
- Second-year Ph.D. student, advised by Dr. Xu Yuan
- DL safety & DL for Imbalanced classification & DL for climate change

UNIVERSITY of LOUISIANA LAFAYETTE

# An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (ViT)

ICLR 2021

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, *et al.*

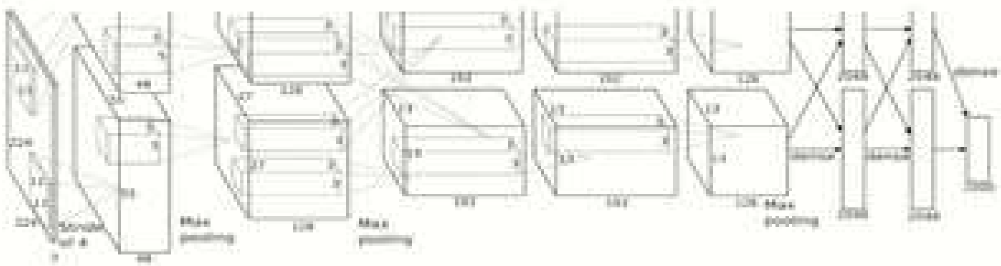
Google Research, Brain Team

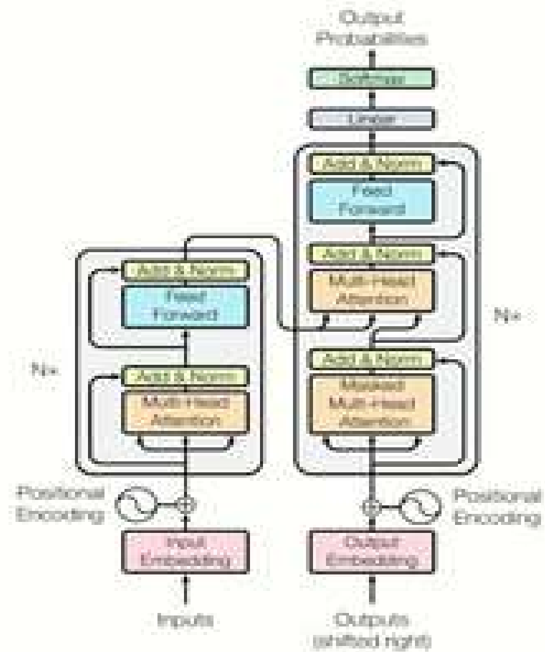
08/17/2022

UNIVERSITY of LOUISIANA LAFAYETTE

# Background

CNNs have been the de-facto architecture for vision for some time...



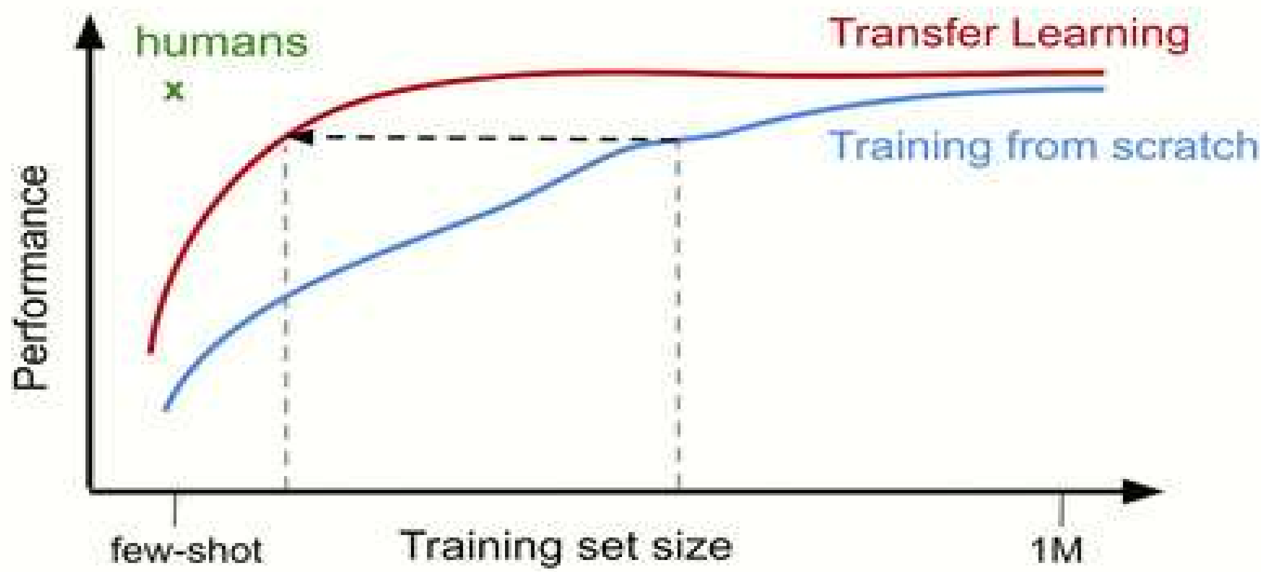



... but Transformers are popular in language, and scale very well, can we use them for vision?

Google Research

# Background

## Transfer works well for small-data tasks

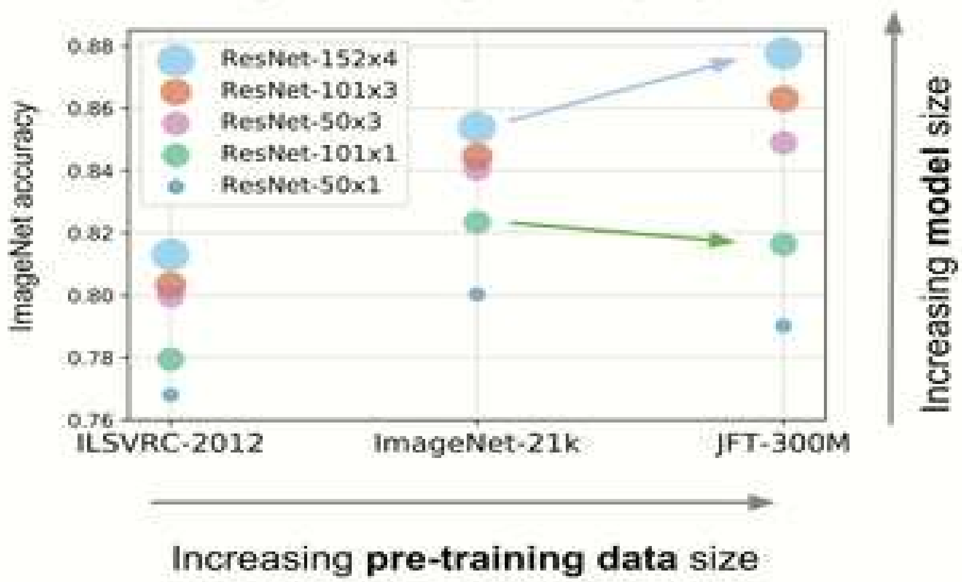



Google Research

# Background

## Transfer learning benefits from scale

Figure from Big Transfer (BiT)*



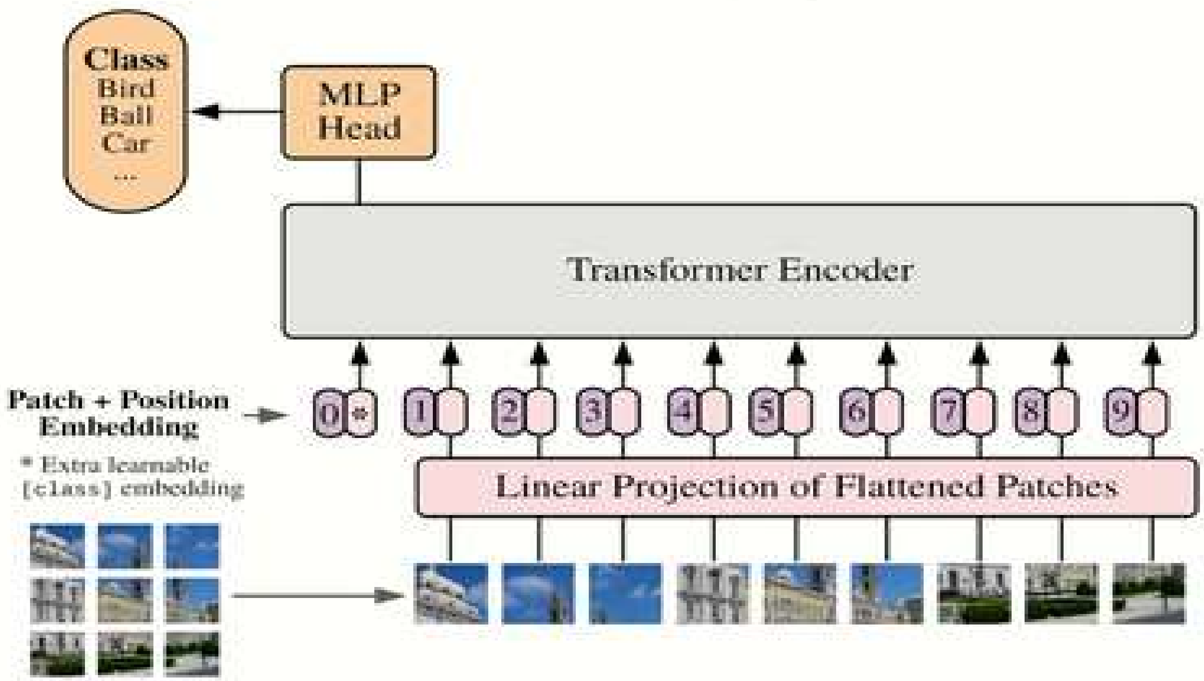

*Note*: pre-training may be expensive, cost is amortized by **cheap transfer** --- BiT models can be fine-tuned with 500-10k steps

Google Research

Big Transfer: General Visual Representation Learning, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby, ECCV 2020

# ViT

## Vision Transformer




Google Research

# ViT

## Architectures

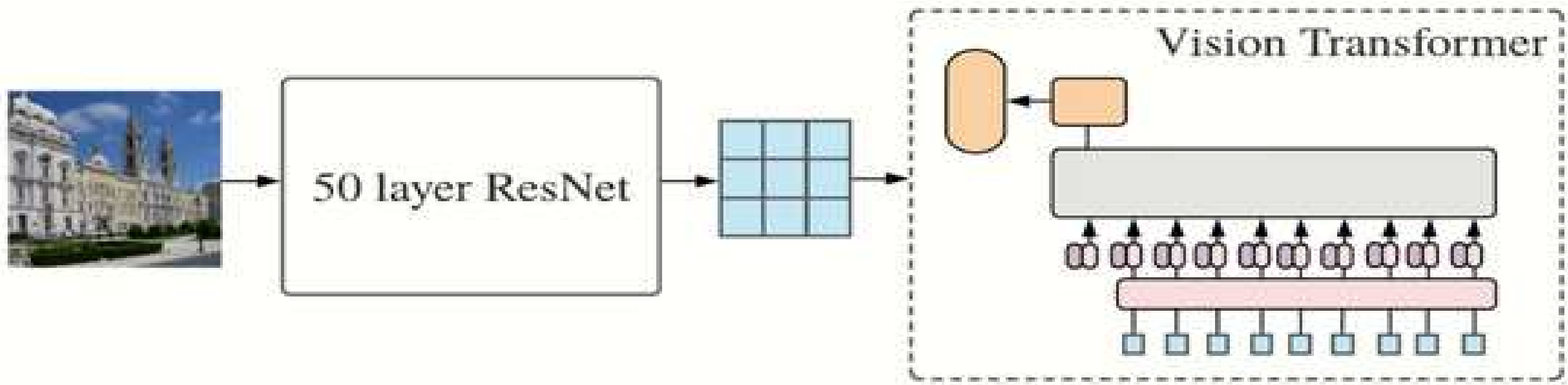
| Model | Layers | Hidden size $D$ | MLP size | Heads | Params |
|---|---|---|---|---|---|
| ViT-Base | 12 | 768 | 3072 | 12 | 86M |
| ViT-Large | 24 | 1024 | 4096 | 16 | 307M |
| ViT-Huge | 32 | 1280 | 5120 | 16 | 632M |


notation example: ViT-L/16

Google Research

# ViT-CNN Hybrid

## ViT-CNN Hybrid

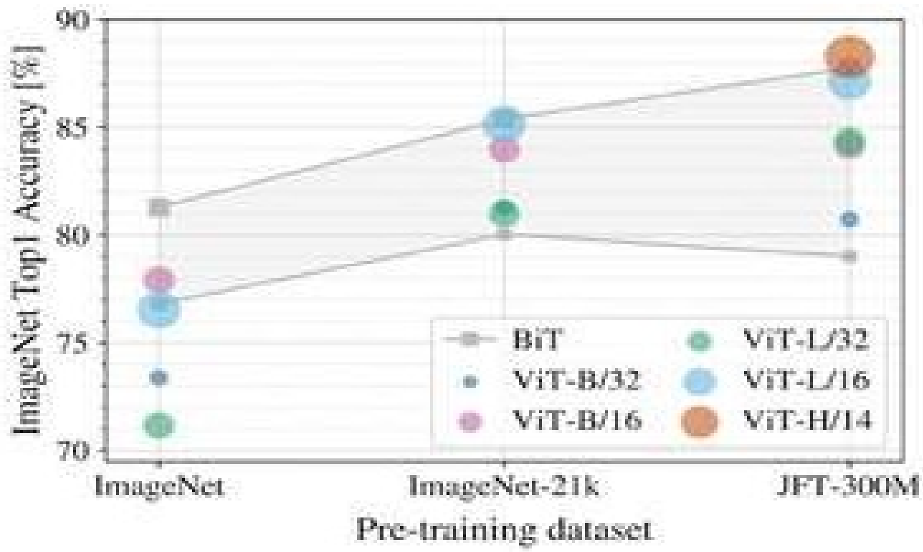



Google Research

# Experiment

## Pre-training Dataset Size

Key
**ViT** = Vision Transformer (this work)
**BiT** = Big Transfer (~ResNet)

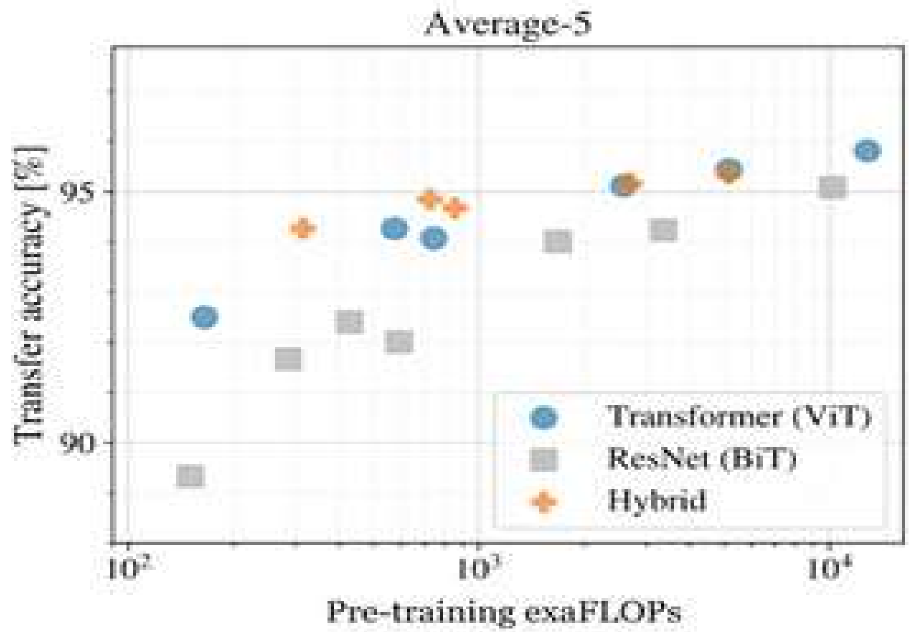



Google Research

**Conclusion**: ViT tends to overfit on ImageNet, but is much better on larger datasets.

# Experiment

## *Pre-training* Compute




**Conclusion 1:** (given sufficient data) ViT gives good performance/FLOP *at all scales*.

Google Research

**Conclusion 2:** ViT-CNN hybrids offer a great deal at small scale, but benefits diminish at large scale.

# Experiment

## Vision Transformer Surpasses Massive CNNs

| | Noisy Student (EfficientNet-L2) | BiT-L (ResNet152x4) |
|---|---|---|
| ImageNet | 88.5 | 87.54 |
| ImageNet ReaL | 90.55 | 90.54 |
| CIFAR-10 | - | 99.37 |
| CIFAR-100 | - | 93.51 |
| Oxford-IIIT Pets | - | 96.62 |
| Oxford Flowers - 102 | - | 99.63 |
| VTAB (19 tasks) | - | 76.29 |
| TPUv3-core-days | 12.3k | 9.9k |

Google Research

# Experiment

## Vision Transformer Surpasses Massive CNNs

| | Noisy Student (EfficientNet-L2) | BiT-L (ResNet152x4) | ViT-Huge/14 |
|---|---|---|---|
| ImageNet | 88.5 | 87.54 | **88.55** |
| ImageNet ReaL | 90.55 | 90.54 | **90.72** |
| CIFAR-10 | - | 99.37 | **99.50** |
| CIFAR-100 | - | 93.51 | **94.55** |
| Oxford-IIIT Pets | - | 96.62 | **97.56** |
| Oxford Flowers - 102 | - | 99.63 | 99.68 |
| VTAB (19 tasks) | - | 76.29 | **77.63** |
| TPUv3-core-days | 12.3k | 9.9k | 2.5k |

ViT-Huge beats SOTA while being ~4x cheaper to pre-train

Google Research

# Experiment

## Vision Transformer Surpasses Massive CNNs

| | Noisy Student (EfficientNet-L2) | BiT-L (ResNet152x4) | ViT-Huge/14 | ViT-Large/16 |
|---|---|---|---|---|
| ImageNet | 88.5 | 87.54 | **88.55** | 87.76 |
| ImageNet ReaL | 90.55 | 90.54 | **90.72** | 90.54 |
| CIFAR-10 | - | 99.37 | **99.50** | 99.37 |
| CIFAR-100 | - | 93.51 | **94.55** | 93.90 |
| Oxford-IIIT Pets | - | 96.62 | **97.56** | 97.32 |
| Oxford Flowers - 102 | - | 99.63 | 99.68 | **99.74** |
| VTAB (19 tasks) | - | 76.29 | **77.63** | 76.28 |
| TPUv3-core-days | 12.3k | 9.9k | 2.5k | 0.68k |

ViT-Large ~matches SOTA* while being >14x cheaper to pre-train

*except for ImageNet

Google Research

# Conclusion

- The first pure transformer architecture in CV.

- Vanilla transformers are surprisingly good at image classification.

- More computation-efficient and easier to scale up

# Limitation

- Data-hungry, e.g., 14M ~ 300M images

- Computation and memory complexity, quadratic

- Poor on self-supervised pre-training

# Reference

- DeiT

  - Training data-efficient image transformers & distillation through attention, ICML 2021

- Swin Transformer

  - Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, ICCV 2021 (Best Paper)

- MAE

  - Masked Autoencoders Are Scalable Vision Learners, CVPR 2022 (Best Paper Nominee)

# The End

Thank you!