Astroformer: More Data Might not be all you need for Classification

ICLR 2023

Rishit Dagli[1]

[1]Department of Computer Science,
University of Toronto

March 4, 2023

Table of Contents

UNIVERSITY OF TORONTO

Table of Contents

UNIVERSITY OF TORONTO

Data is at the heart of modern AI

- Evident that the presence of large-scale data has driven most of the advances in AI

Data is at the heart of modern AI

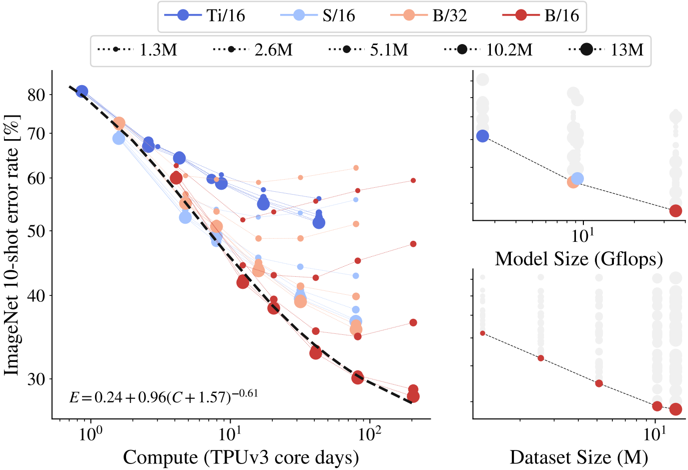
- Evident that the presence of large-scale data has driven most of the advances in AI
- Dataset Sizes for Image Classification:
    - MNIST - 60K
    - CIFAR 10,100 - 60K
    - Fashion-MNIST - 70K
    - ImageNet - 14M
    - JFT - 300M
    - JFT - 3B

UNIVERSITY OF TORONTO

Data is at the heart of modern AI

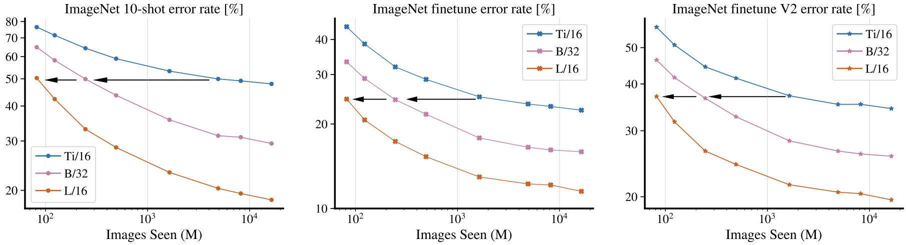
- Evident that the presence of large-scale data has driven most of the advances in AI
- Dataset Sizes for Image Classification:
  - MNIST - 60K
  - CIFAR 10,100 - 60K
  - Fashion-MNIST - 70K
  - ImageNet - 14M
  - JFT - 300M
  - JFT - 3B
- Rise of Transformers and Vision Transformers, well known since Bahdanau attention [Bahdanau et al.,2014] that Attention-based models scale very well with the amount of training data

Data is at the heart of modern AI



Scaling Vision Transformers, Zhai et al., 2022

Data is at the heart of modern AI



Scaling Vision Transformers, Zhai et al., 2022

Table of Contents

Vision Transformers are data-hungry

- Vision Transformers (monolithic or non-monolithic) suffer heavily when trained from scratch on small datasets

Vision Transformers are data-hungry

- Vision Transformers (monolithic or non-monolithic) suffer heavily when trained from scratch on small datasets

Vision Transformers are data-hungry

- Vision Transformers (monolithic or non-monolithic) suffer heavily when trained from scratch on small datasets

Model	top-1 accuracy (↑)
Efficient Adaptive Ensembling	99.612
ResNet56	74.44
ResNet110	76.18
EfficientNet B0	61.64
ResNet18	64.49
ViT (scratch)	73.81
ViT-Drloc	58.29
SL-ViT	76.92

Vision Transformers are data-hungry

- Vision Transformers (monolithic or non-monolithic) suffer heavily when trained from scratch on small datasets
- Inherent inductive biases allow training CNNs on small-scale datasets from scratch [D'Ascoli et al., 2021]

UNIVERSITY OF TORONTO

Vision Transformers are data-hungry

- Vision Transformers (monolithic or non-monolithic) suffer heavily when trained from scratch on small datasets
- Inherent inductive biases allow training CNNs on small-scale datasets from scratch [D'Ascoli et al., 2021]
- Vision Transformers often have a lack of locality, inductive biases, and hierarchical structure of the representations

How to Train your ViT?

- ViTs require large-scale pre-training to learn the properties they lack from large-scale data

How to Train your ViT?

- ViTs require large-scale pre-training to learn the properties they lack from large-scale data
- All SOTA Vision transformers: BASIC-L, CoCa, CoAtNet, and ViT-G are trained on JFT datasets

UNIVERSITY OF TORONTO

How to Train your ViT?

- ViTs require large-scale pre-training to learn the properties they lack from large-scale data
- All SOTA Vision transformers: BASIC-L, CoCa, CoAtNet, and ViT-G are trained on JFT datasets
- Transformers are often pre-trained and then fine-tuned for acceptable performance

UNIVERSITY OF TORONTO

Table of Contents

Intuitively Merging Convolutions and Attention

Multiple attempts in the past.

$$A_{i,j} = \sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k\right) \quad \text{(standard self-attention)}$$

$$A_{i,j} = \sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k + w_{i-k}\right)$$

x_i, $y_i \in \mathbb{R}^d$ are the input and output at position i, w_{i-j} represents the depthwise convolution kernel and $\mathcal{G}$ represents the global spatial space

Intuitively Merging Convolutions and Attention

Multiple attempts in the past.

$$A_{i,j} = \sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k\right) \quad \text{(standard self-attention)}$$

$$A_{i,j} = \sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k + w_{i-k}\right)$$

x_i, $y_i \in \mathbb{R}^d$ are the input and output at position i, w_{i-j} represents the depthwise convolution kernel and $\mathcal{G}$ represents the global spatial space

Intuitively Merging Convolutions and Attention

Multiple attempts in the past.

$$A_{i,j} = \sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k\right) \quad \text{(standard self-attention)}$$

$$A_{i,j} = \sum_{k \in \mathcal{G}} \exp\Big(x_i^\top x_k + \underbrace{\boxed{w_{i-k}}}_{\text{depthwise-convolution}}\Big)$$

x_i, $y_i \in \mathbb{R}^d$ are the input and output at position i, w_{i-j} represents the depthwise convolution kernel and $\mathcal{G}$ represents the global spatial space

Intuitively Merging Convolutions and Attention

$$y_i = \sum_{j \in \mathcal{G}} \frac{\exp\left(x_i^\top x_j + w_{i-j}\right)}{\sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k + w_{i-k}\right)} x_j$$

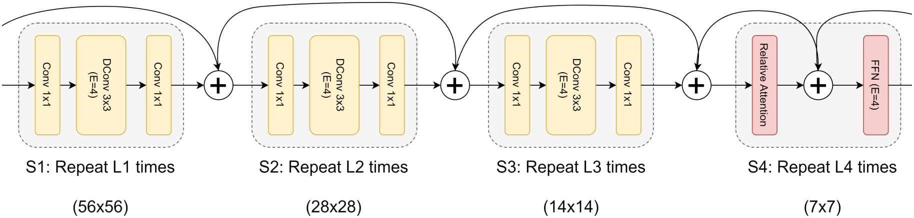
x_i, $y_i \in \mathbb{R}^d$ are the input and output at position i, w_{i-j} represents the depthwise convolution kernel and $\mathcal{G}$ represents the global spatial space

Intuitively Merging Convolutions and Attention

$$y_i = \sum_{j \in \mathcal{G}} \frac{\exp\left(x_i^\top x_j + w_{i-j}\right)}{\sum_{k \in \mathcal{G}} \exp\left(x_i^\top x_k + w_{i-k}\right)} x_j$$

x_i, $y_i \in \mathbb{R}^d$ are the input and output at position i, w_{i-j} represents the depthwise convolution kernel and $\mathcal{G}$ represents the global spatial space

Model



- Core includes C-C-C-T which has been earlier thought to not work well [Dai et al., 2021; Tu et al., 2022]
- Comes with theoretical improvements
- Multiple other components

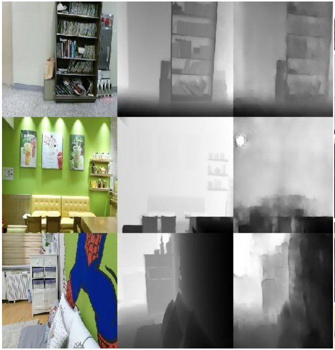
UNIVERSITY OF TORONTO

Table of Contents

Results

- We set a new SOTA on Tiny ImageNet (by 0.98%)
- We set a new SOTA on CIFAR-100 w/o extra training data (by 3.46%)
- We set a new SOTA on Galaxy10 DECals (by 4.62%)
- Competitive performance on CIFAR-10 (99.12%)

As a backbone



UNIVERSITY OF TORONTO

Thank You