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

Rishit Dagli
University of Toronto

ICLR International Conference On Learning Representations

UNIVERSITY OF TORONTO

Overview

We introduce a recipe for modifying hybrid transformer-convolution models to learn in the low-data regime while preserving translational equivariance.

Goals: New hybrid transformer-convolutional architecture comparable to the approach employed by CoAtNet for low-data regime while making the hybrid model retain translational equivariance.

Motivation

Dataset Sizes

- Evident that the presence of large-scale data has driven most of the advances
- Attention-based models scale very well with the amount of training data
- Collecting high-quality labeled data or human annotators is very expensive

Vision Transformers and Hybrid Models

- Vision Transformers (monolithic or non-monolithic) suffer heavily when trained from scratch on small datasets
- Hybrid transformer-convolution models help encode inductive biases but do not fully solve training in low-data regimes from scratch
- Transfer learning will not fully solve the problem either due to bias in pre-training datasets that do not reflect environments
- Self-supervised learning and semi-supervised learning are great, but for a lot of objectives, especially science objectives getting non-annotated data is costly

Previous work on merging convolutions and attention

Multiple approaches to unify convolutions and attention:

- Augment the ConvNet backbone with explicit self-attention or non-local modules like Block Attention:

$$F' = M_c(F) \otimes F$$
$$F'' = M_s(F') \otimes F'$$

  where F is an intermediate feature map, M_c represents a 1D channel attention map, and M_s represents the a 2D spatial attention map.
- Replace convolution layers with standard self-attention
- Start with a Transformer backbone and tries to incorporate explicit convolutions or their properties, like relative attention:

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

  where $x_i, y_i \in \mathbb{R}^d$ are the input and output at position i, w_{i-j} represents the depthwise convolution kernel as a scalar ($w \in \mathbb{R}^{\mathcal{O}(|\mathcal{G}|)}$) and $\mathcal{G}$ represents the global spatial space.

This work falls under the same category but specifically instantiates relative attention and designs the vertical layout design to be better suited toward low-data regime vision recognition tasks.

Method

Transformer Block

The Transformer block makes use of relative attention which efficiently combines depthwise convolutions and self-attention. A depthwise convolution uses a fixed kernel to extract features from a local region of the input data whereas self-attention allows the receptive field to be the global spatial space.

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

The attention weight $A_{i,j}$ is decided by both w_{i-j} and $x_i^\top x_j$. The update made to the attention weight is rather intuitive by simply summing a global static convolution kernel.

Network

- down-sampling the feature map via a multi-stage network with gradual pooling to reduce the spatial size
- employ global relative attention

5-stage network to do so:

- `S0` is a simple 2-layer convolutional Stem
- `S1`, `S2`, and `S3` employs Inverted Residual blocks
- `S4` employs a Transformer block

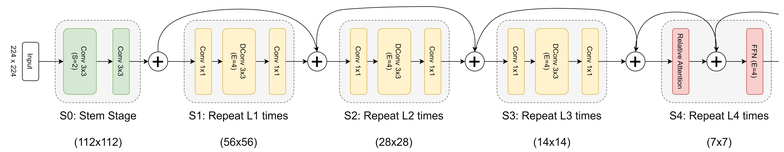

Figure 1. Overview of the proposed model, notice the stack-design design we employ.

Theoretical Improvements

- greater generalizability and does not face highly unstable training when training from scratch on small data unlike other models
- inherent translational equivariance makes training less prone to overfitting in the low-data regime

Proof Sketch (proof in the paper appendix A.2)

notice that the depthwise convolution kernel for any position pair (i, j) is only dependent on $i-j$, the relative positions rather than the values of i and j individually.

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

Results Overview

Full tabular results can be found in the paper Section 5 and Appendix A.5.

- 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%)

Scaling Curves

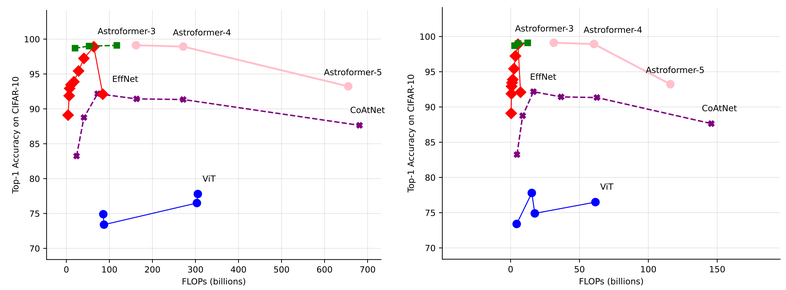

Figure 2. (a) The top-1 accuracy to parameter scaling curves for multiple models on the CIFAR-10 dataset. (b) The top-1 accuracy to FLOPs scaling curves for multiple models on the CIFAR-10 dataset. All these scaling curves are for the evaluation size of 224^2.

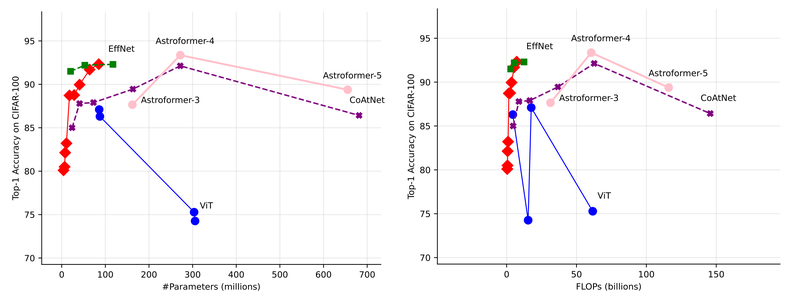

Figure 3. (a) The top-1 accuracy to parameter scaling curves for multiple models on the CIFAR-100 dataset. (b) The top-1 accuracy to FLOPs scaling curves for multiple models on the CIFAR-100 dataset. All these scaling curves are for the evaluation size of 224^2.

References

[1] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3965–3977. Curran Associates, Inc., 2021.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.