RESEARCH

Data-efficient image Transformers: A promising new technique for image classification

December 23, 2020

What the research is:

We've developed a new method to train computer vision models that leverage Transformers, the breakthrough deep neural network architecture that has recently unlocked dramatic advances across many areas of AI. Transformer models have produced state-of-the-art results in natural language processing and machine translation, and Facebook AI has used the architecture to break new ground with other tasks, such as speech recognition, symbolic mathematics, and translation between programming languages. But the AI research community is only beginning to bring Transformers to the field of computer vision, with projects such as Facebook AI's DETR object detection architecture, released earlier this year.

Our new technique — Data-efficient image Transformers (DeiT) — requires far less data and far less computing resources to produce a high-performance image classification model. Training a DeiT model with just a single 8-GPU server over 3 days, we achieved 84.2 top-1 accuracy on the widely used ImageNet benchmark without using any external data for training. This result is competitive with the performance of cutting-edge

convolutional neural networks (CNNs), which have been the dominant approach to image classification for many years.

By showing that Transformers can be trained efficiently for image classification, using only regular academic datasets, we hope to advance the field of computer vision, extend Transformers to new use cases, and help make this work more accessible to researchers and engineers who don't have access to large-scale systems for training massive AI models.

DeiT was developed in collaboration with Professor Matthieu Cord from Sorbonne University. We are now open-sourcing the code and publishing our research so that others can reproduce our results and build upon our work.

This graph shows the performance curve comparing our approach (DeiT and DeiT with distillation) with that of previous visual Transformer models and modern state-of-the-art CNNs. The models shown here were trained on ImageNet.

How it works:

Image classification — the task of understanding the main content of an image — is easy for humans but hard for machines. In particular, it is challenging for convolution-free Transformers like DeiT because these systems don't have many statistical priors about images: They typically have to "see" a lot of example images in order to learn to classify different objects. DeiT, however, can be trained effectively with 1.2 million images, rather than requiring hundreds of millions of images.

The first important ingredient of DeiT is its training strategy. We built upon and adapted existing research initially developed for convolutional neural networks. In particular, we used data augmentation, optimization, and regularization in order to simulate training on a much larger dataset.

Equally important, we modified the Transformer architecture to enable native distillation. Distillation is the process by which one neural network (the student) learns from the output of another network (the teacher). We used a CNN as a teacher model for our Transformer. Since the CNN's architecture has more priors about images, it can be trained with a comparatively smaller number of images.

Using distillation can hamper the performance of neural networks. The student model pursues two different objectives that may be diverging: learning from a labeled dataset (strong supervision) and learning from the teacher. To alleviate this, we introduced a distillation token, which is a learned vector that flows through the network along with the

transformed image data. The distillation token cues the model for its distillation output, which can differ from its class output. This new distillation method is specific to Transformers and further improves the image classification performance.

We add a distillation token to the Transformer. It interacts with the classification vector and image component tokens through the attention layers. The objective of this distillation token is to learn from the teacher model (a CNN).

Why it matters:

DeiT is an important step forward in using Transformers to advance computer vision. Its performance is already competitive with that of CNNs, even though the latter have been the dominant approach for computer vision tasks for the last eight years and have benefited from many improvements and adjustments. We hope this indicates that additional research will produce significant additional gains.

This work will also help democratize AI research. DeiT shows that it is possible for developers with limited access to data and computing resources to train or use these new models. We hope that it will help foster advances by a larger community of researchers.

Read the paper and get the code:

Code available at: https://github.com/facebookresearch/deit

Paper available at:
Training data-efficient image transformers and distillation through attention

Written By

Hugo Touvron
Research Assistant

Matthijs Douze
Research Scientist

Francisco Massa
Research Engineer

Alex Sablayrolles

Herve Jegou

Director, Research Scientist

Our Work

RESEARCH

Facebook AI's Joelle Pineau receives Governor General's Innovation Award

The award recognizes Canadian leaders for their groundbreaking innovations and positive impact on the quality of life in the country.

May 22, 2019

RESEARCH

Facebook Research at ICLR 2020

Facebook researchers will be participating in several activities at this year's virtual ICLR 2020.

April 27, 2020