Tutorial Title

Transformers for Visual Recognition

Abstract

Vision Transformers (ViTs) have emerged as a competitive alternative to Convolutional Neural Networks (CNNs) for image recognition. Since ViT, the use of Transformers for visual representation learning has sparked interest in the computer vision community.

CNNs are well-suited for extracting visual representations because of their inherent local inductive bias. On top of that, CNNs also exploit the scope of parallelization on modern GPUs, making them efficient to train. However, CNNs lack a global understanding of the image, essential for visual recognition tasks.

To track long-range dependencies within an image, CNNs need large receptive fields, which involve large kernels or long sequences of convolutional layers, significantly hurting the efficiency of the training process. Transformers can overcome these shortcomings by leveraging self-attention.

The visual representations generated by self-attention do not contain the spatial constraints of convolutions; instead, they can learn the most suitable inductive biases from the input data depending on the task and the position of the self-attention operation within the pipeline.

This tutorial aims to equip the attendees with a holistic understanding of how ViTs work with a key focus on their internals, training methodologies, investigating their learned representations. Additionally, we'll also show where ViTs stand across different visual recognition tasks with respect to similar CNN models. By the end of this tutorial, we hope that the attendees will have a fresh perspective when designing efficient SoTA solutions for computer vision problems in the future.

Description and Outline

- **A birds-eye view of convolutional architectures in vision**
    - Common architecture patterns:
      We will discuss the common architecture patterns for CNNs while focusing on the evolution of the architectures. This section forms a vital prerequisite for our audience.
    - Performance :
      Parallel discussion on the performance of the CNN models helps build an intuition of the overall evolution of the architectures.
- **Motivating the need for Transformers and introduction**
    - Introduction to the attention mechanism:
      We will begin by addressing the shortcomings of CNNs in modeling long-range dependencies and why it might be crucial to look at visual recognition tasks through the lens of performance and explainability. We will cover the in-depth

formulation of the attention mechanism (scaled-dot product attention) taking references from Attention is All You Need.

- Introduction to Transformers:
  - Main components:
    We will discuss the main components of the Transformer architecture with our main references being the original Transformer work and the original ViT work. The following components will be covered: patchification of images and linear projections, positional embeddings, stack of Transformer blocks and their constituent blocks (self-attention module, residual connections, feed-forward module), and representation pooling with CLS token. We will also draw their parallels to the vanilla Transformer architecture.
  - Its success in NLP (GPT-3, T5, Copilot, etc.)
  - Transformers as a general computational primitive to learn strong inductive biases from data (modality unification)
  - Its relevance in computer vision (allows for fewer inductive biases, ability to shine under more extensive data regimes, improved robustness, etc.)

- **Review of popular Transformer-based architectures in computer vision**
  - End-to-End Object Detection with Transformers (one of the first architectures to have explored object detection with Transformers)
  - An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (ViT for short, the architecture that boomed the popularity of Transformers in vision)
  - Training data-efficient image transformers & distillation through attention
  - Swin Transformer: Hierarchical Vision Transformer using Shifted Windows
  - Focal Self-attention for Local-Global Interactions in Vision Transformers

  This section will include the main ideas presented in the mentioned works along with their pros and cons and what might have motivated their developments. We will discuss how these architectures perform on the standard image recognition benchmarks. Besides, we will cover some interpretability techniques introduced in Do Vision Transformers See Like Convolutional Neural Networks? so that the participants understand how vision transformers process images and can characterize what makes their workings different from that of convolutional architectures.
- **Training methodologies for Transformer-based architectures**
  - Scaling properties (as discussed in Scaling Vision Transformers)
  - Techniques to counter the larger data regimes
    - How to train your ViT? Data, Augmentation, and Regularization in Vision Transformer
    - When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations
    - Training data-efficient image transformers & distillation through attention
    - Vision Transformer for Small-Size Datasets

- **Emergence of hybrid models in computer vision**
    - Self-attention based
        - Attention Augmented Convolutional Networks
        - Stand-Alone Self-Attention in Vision Models
        - Bottleneck Transformers for Visual Recognition
        - Augmenting Convolutional networks with attention-based aggregation
    - Transformer-based
        - ResNet-ViT (introduced in ViT)
        - CoAtNet: Marrying Convolution and Attention for All Data Sizes
        - ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases
        - MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer
- **Other promising areas of Transformers in Vision and conclusion**
    - Learning Transferable Visual Models From Natural Language Supervision (CLIP)
    - Zero-Shot Text-to-Image Generation (DALL-E) and Taming Transformers for High-Resolution Image Synthesis (VQGAN)
    - Emerging Properties in Self-Supervised Vision Transformers (DINO)

Presenters (unordered)

- Aritra Roy Gosthipaty, Deep Learning Associate at PyImageSearch, <aritra.born2fly@gmail.com>, 81-Milan Park, Garia, Kolkata-700084, India, Ph: +918420922326.

  Bio: Aritra is currently working at PyImageSearch breaking down complicated Deep Learning research papers and providing code walkthroughs for the same. He is interested in a wide spectrum of representation learning. Aritra has contributed (1, 2, 3, 4, 5) technical posts to Keras pertaining to Transformers, mainly Vision Transformers. He has also contributed the TensorFlow versions of some Vision Transformer-based models to the Hugging Face Transformers library. His speaking engagements can be found at this URL.
- Sayak Paul, Machine Learning Engineer at Carted, <spsayakpaul@gmail.com>, Sukanta Pally, Baruipur, Kolkata: 700144, India, Ph: +918981257929.

  Bio: Sayak is currently working at Carted building an NLP-powered application that can extract attributes from E-Commerce product webpages. He is interested in representation learning, in particular topics like self-supervision, semi-supervision, model robustness fascinate him. Sayak has contributed (1, 2, 3, 4, 5) technical posts to Keras pertaining to Transformers and has also contributed a series of Vision Transformer-based models to TensorFlow Hub as well as Hugging Face Transformers. Most recently his work Vision Transformers are Robust Learners got accepted to AAAI 2022. Additionally, Sayak presented a tutorial at CVPR 2021 and also served as a Program Committee member for the

UDL workshop at ICML 2021. His speaking engagements can be found at this URL.

- Kranthi Kiran GV, graduate student at Courant Institute of Mathematical Sciences at the New York University, <kranthi.gv@nyu.edu>, Phone: +1 646 359 1360.
  Bio: Kranthi is currently a computer science graduate student at NYU. He is interested in computer vision and machine learning. He is currently working with Prof. Krzysztof Geras on detection of breast cancer using vision transformers, and extraction of structured information from pathology reports using transformers. Kranthi is also studying the social biases in large pre-trained transformers used for language modeling in NLP.

It is to be noted that we all will be contributing to the development of the materials for the tutorial. We will divide presentation time equally amongst ourselves.

Similar tutorial(s)

Self-Attention for Computer Vision (ICML 2021, July 19). Differences include - coverage of training mechanisms for Vision Transformers, emerging properties of Vision Transformers, and more common and simpler hybrid models. Similarities include - coverage of Self-attention and Transformers for Computer Vision.

On the importance of the tutorial at NeurIPS 2022

The computer vision community continues to develop and investigate the use of Transformers for modeling perception. Therefore we believe there hasn't been a better time than this to take a step back and revisit what we have learned about them over these past few years. Transformers are becoming general computation primitives. They help us to induce a form of universality in model architectures to apply deep learning to different data modalities without requiring hand-engineered components. So, it now could be made possible to apply deep learning to rare and important modalities of data (protein sequencing for example) with Transformers that were not possible otherwise. This, in turn, could be helpful for subjects concerning physical sciences. Therefore we hope to benefit the community by providing a holistic understanding of Transformers across many different aspects.

Furthermore, with the help of the panel, we aim to gather thoughts around the challenges around ViTs for dense prediction tasks as well as generative tasks and approaches to solve these challenges. We believe this will complement the overall theme of our tutorial and will provide the attendees with notable research directions.

Diversity

The speakers and panelists come from varied backgrounds portraying different ethnicities, professions, and races. They also work in geographical locations across the world including North America, Europe, and Asia. The speakers are both from industry and academia.

Notes

- We have embedded the references as links at their respective places and hence we are not providing a separate reference section.
- We plan to engage with the audience by including short QA segments in between our presentation to keep it interactive. Additionally, we will always have a moderator (amongst the speaker) rotating who will be responsible for attending the QA otherwise.

Panelists (unordered)

Ishan Misra

Ishan is a Research Scientist at Facebook AI Research (FAIR) where he works on Computer Vision and Machine Learning. His research interest is in reducing the need for supervision in visual learning. Ishan has co-authored a number of seminal works involving the use of Transformers for Computer Vision. Some examples include DINO, Omnivore, Siamese Masked Networks. These works show different aspects of using Transformer-based backbones for visual recognition tasks such as self-supervision, modality unification, and low-shot transfer. His Google Scholar profile can be found here.

Lucas Beyer

Lucas is a Senior Research Engineer on the Brain Team at Google Research. His research interests include Representation Learning, Reinforcement Learning, Computer Vision, and Robotics. Lucas is one of the primary authors of the seminal paper An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Since the release of that paper, he has worked on investigating how to scale Vision Transformers, how to train Vision Transformers under various data regimes, and a better Vision Transformer baseline for image classification (to name a few). His Google Scholar profile can be found here.

Irwan Bello

Irwan is a Research Scientist at Google Brain where he works on artificial intelligence, large-scale language models, and computer vision. His research currently focuses on making large-scale language models cheaper to work with - via sparsity, adaptive computation, and better-distributed computing infrastructure. In computer vision, he authored some of the pioneering work on Attention for Vision (1, 2, 3), proposed LambdaNetworks as a faster alternative, and worked on simple vision baselines. His Google Scholar profile can be found here.