HTWG Konstanz

MSI Seminar Advanced Topics in Data Analysis and Deep Learning

Vision Transformer (ViT)

Spotlight Talk - Alexander Haab

Sommersemester 2025

Spotlight Talk

- Paper Introduction
- Model Overview
- Key Contributions
- Coming Up

Paper Introduction

- NLP: Transformer
- Vision: CNN, Hybrid (CNN + Attention)

- Foundation paper
- Introducing ViT architecture
- Google Research, 2020

- Pure transformer
- Image classification tasks
- Pre-trained on large amounts
- Fine-tuned for task
- Trained in supervised fashion

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

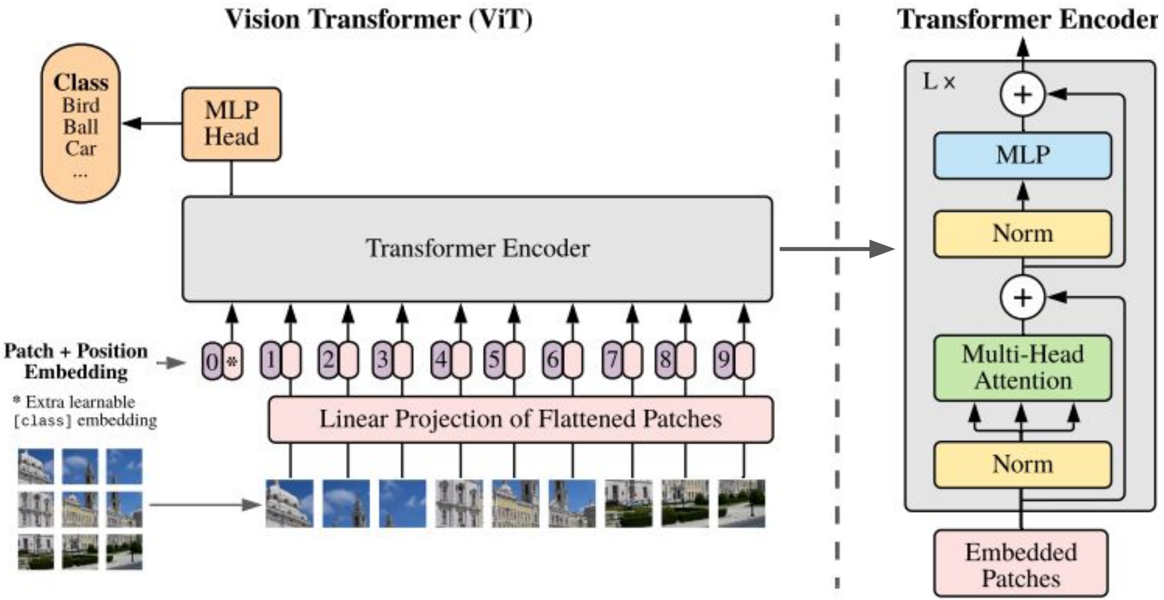
{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

(Dosovitskiy et al., 2021)

Model Overview



(Dosovitskiy et al., 2021)

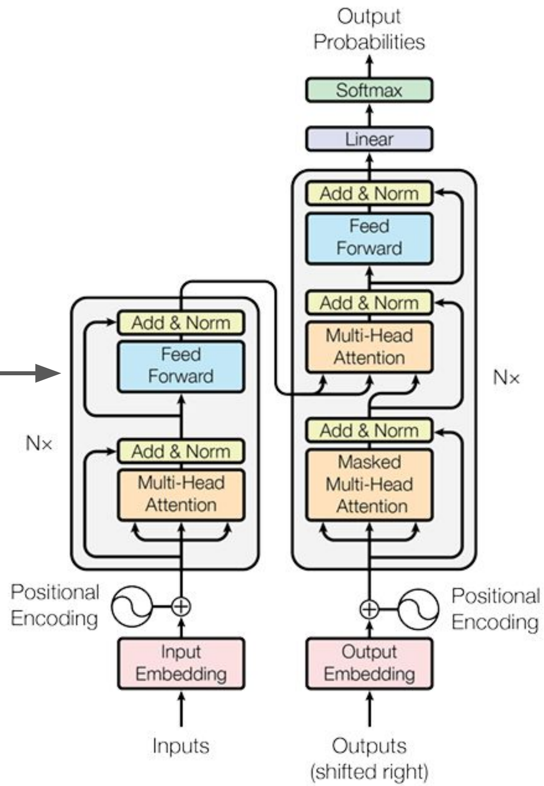


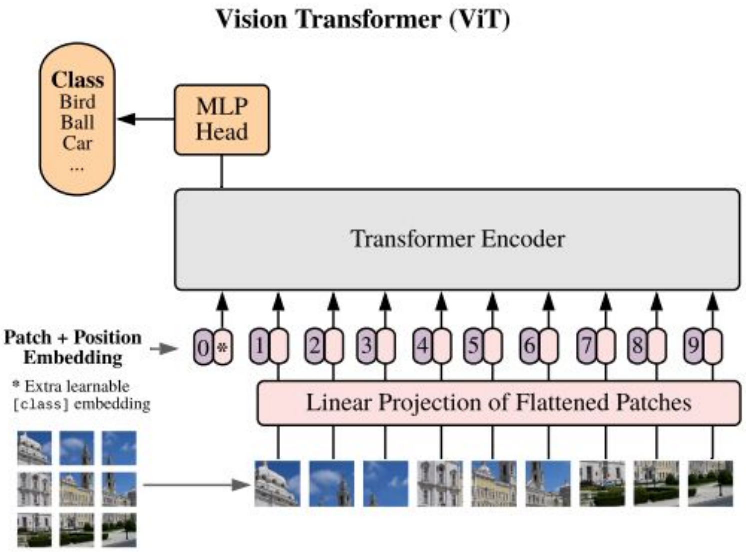
Figure 1: The Transformer - model architecture.
(Vaswani et al., 2017)

Model Overview

Transformer input: sequence of tokens

Goal: image = sequence of tokens

1. Split image into patches: 3x3, 16x16 = 256 tokens
2. Flatten patches: 2D input to 1D vector
3. Lineare projection: Map to token dimension
4. Add classification token
5. Position embedding: position + patch
6. Feed to standard transformer encoder
7. Input classification token into small MLP
8. Class probability distribution



(Dosovitskiy et al., 2021)

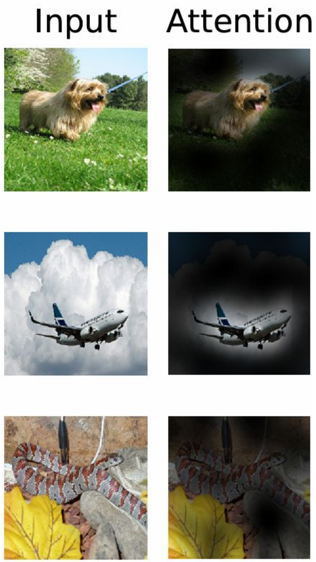
Key Contributions

- Trained on mid-sized datasets: Modest accuracy
- Trained on large datasets (14M - 300M images) :
    - Beats state-of-the-art convolutional networks,
    - on multiple image recognition benchmarks,
    - while requiring substantially fewer computational resource to train

- Benefits of pure transformers to computer vision:
    - computational efficiency
    - scalability (possibility to train large models)

Coming Up

- Inspecting Vision Transformer
- Model Variants, Training & Fine-tuning
- Comparison to State of the Art
- Related Work
  - Cites: Attention Is All You Need. (Vaswani et al., 2017), …
  - Cited by: ?
- Fine-tuning Code and pre-trained models available at github:
  - https://github.com/google-research/vision_transformer



(Dosovitskiy et al., 2021)

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021).
An image is worth 16x16 words: Transformers for image recognition at scale.
arXiv. https://arxiv.org/abs/2010.11929
Github. https://github.com/google-research/vision_transformer

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017).
Attention Is All You Need.
arXiv. https://arxiv.org/abs/1706.03762