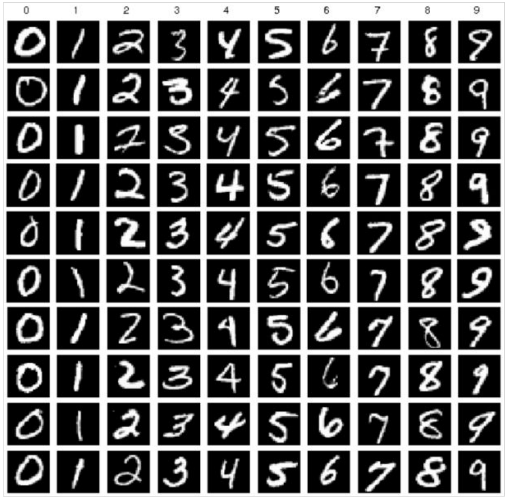
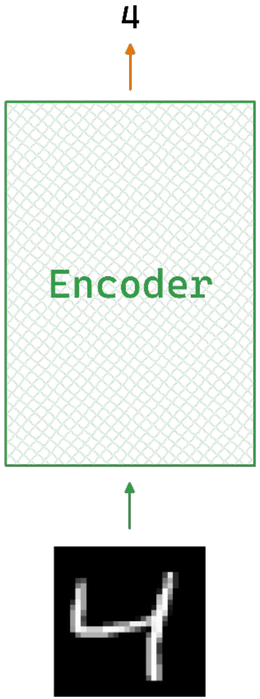
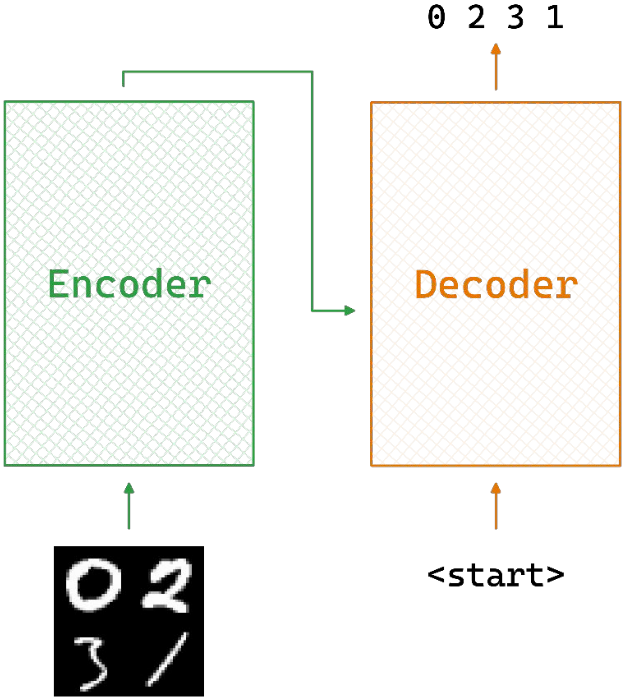
A picture is worth a thousand words

Can you write them?

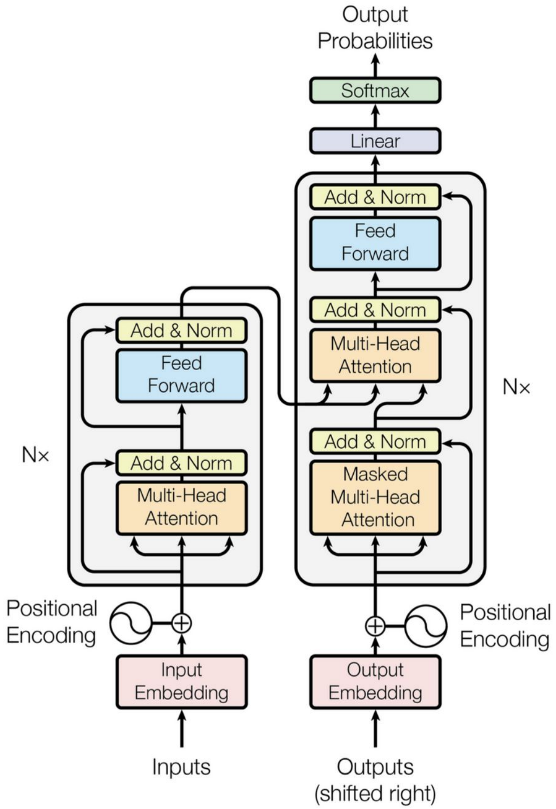
Task



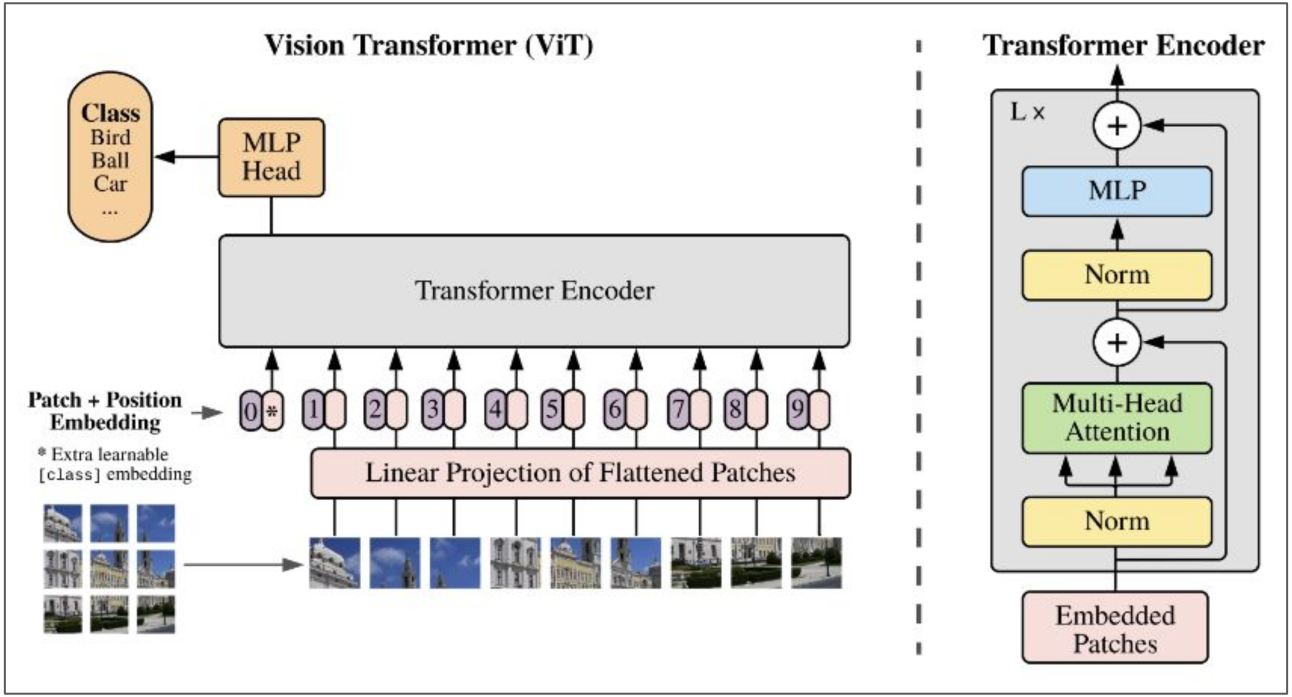




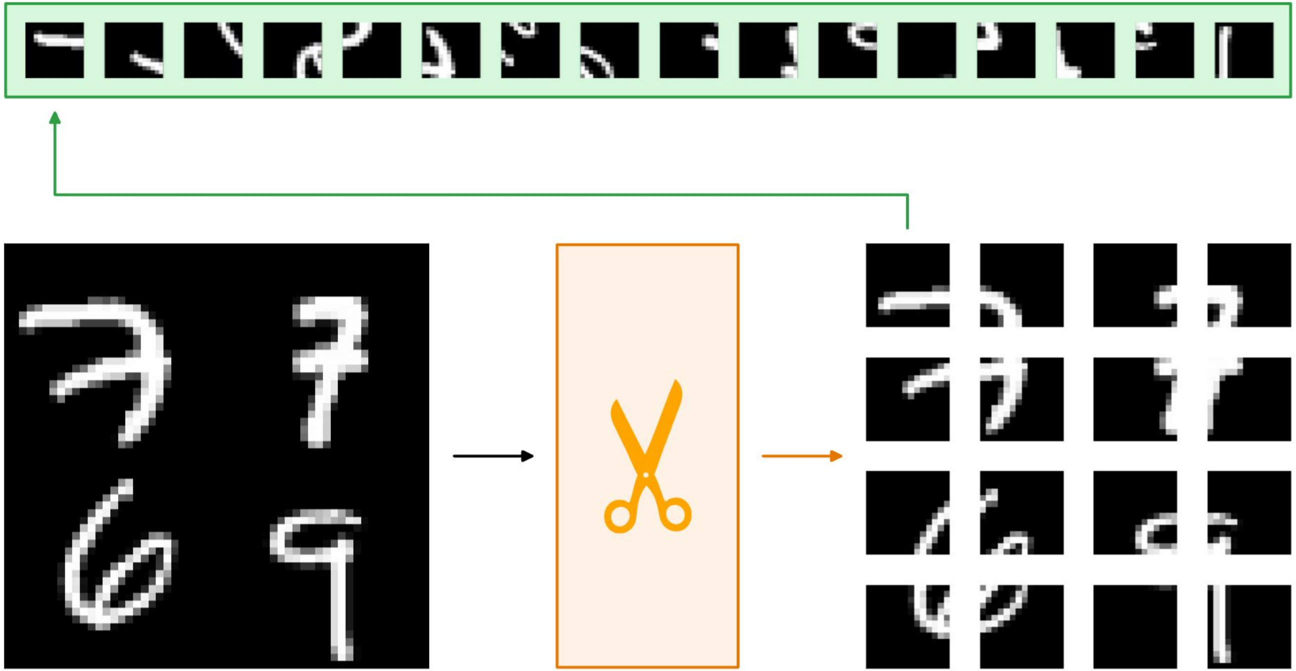
Attention Is All You Need



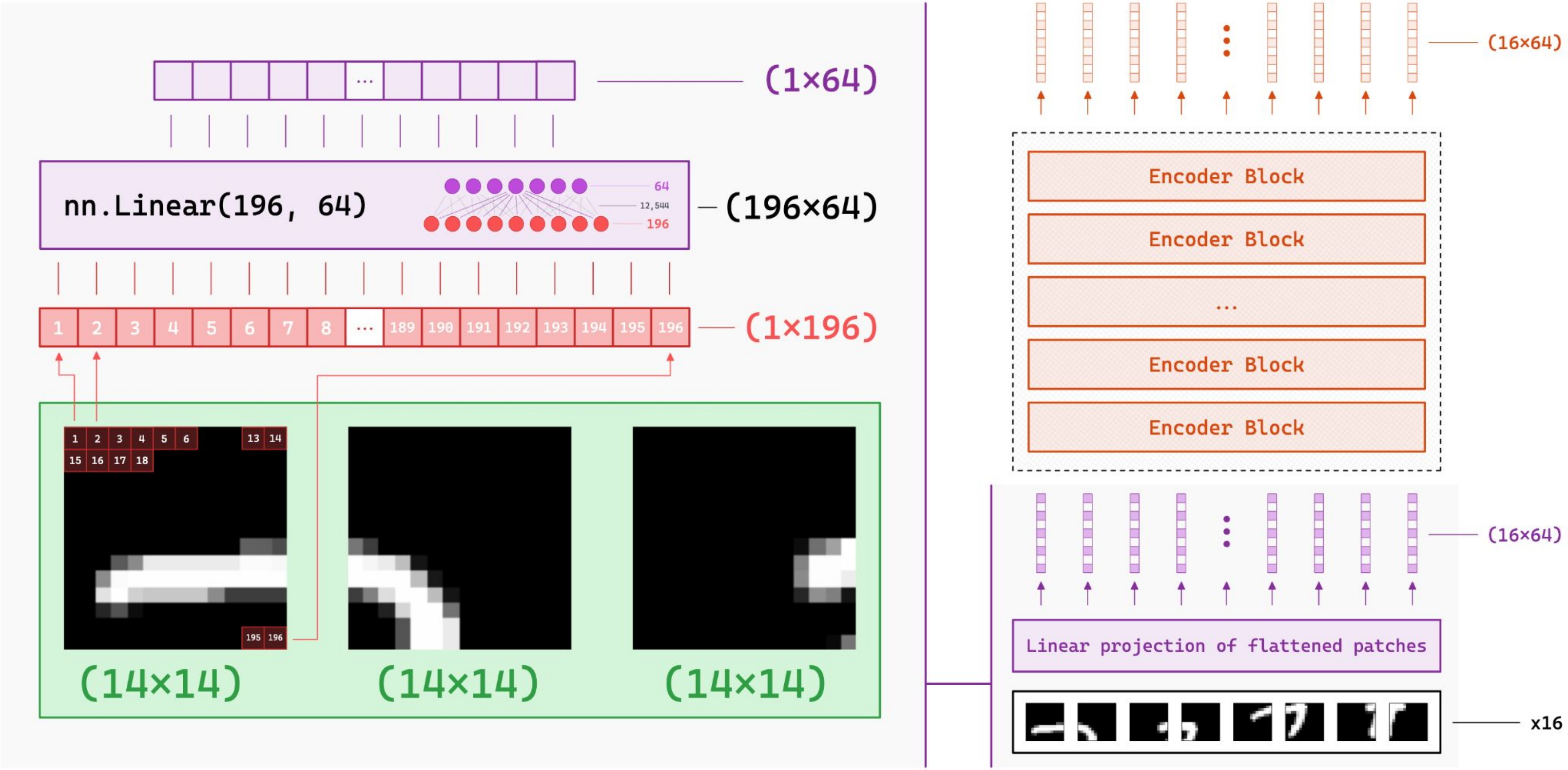
Vision Transformer (ViT)



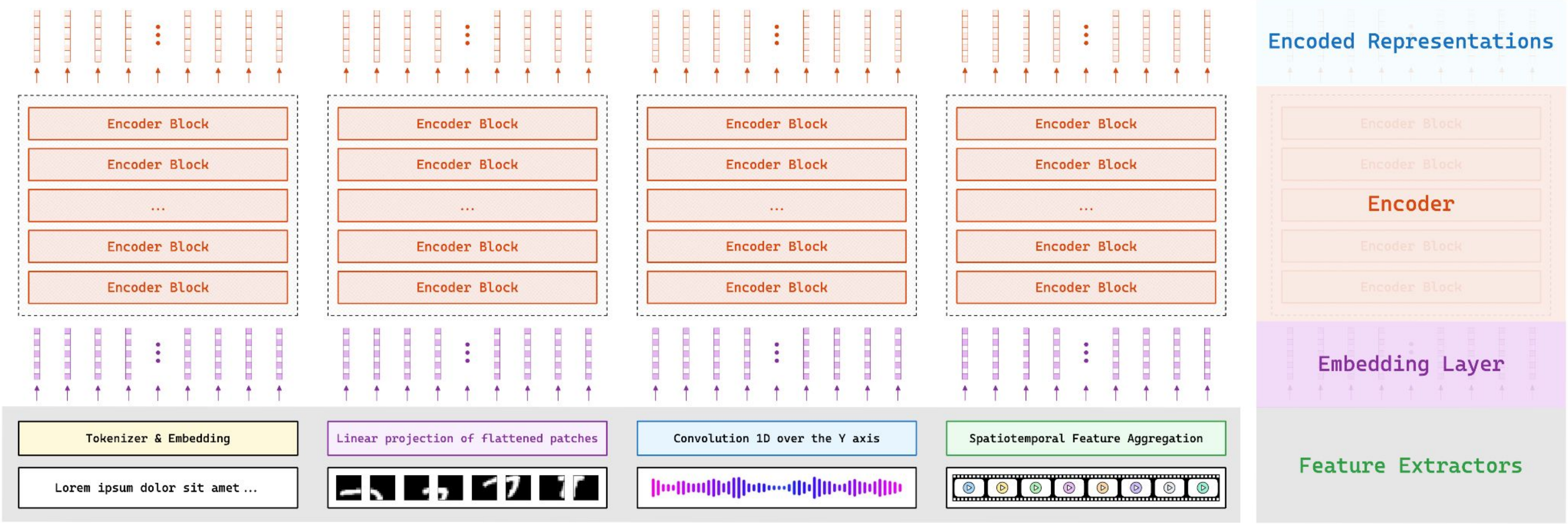
Prep



Patch Projection

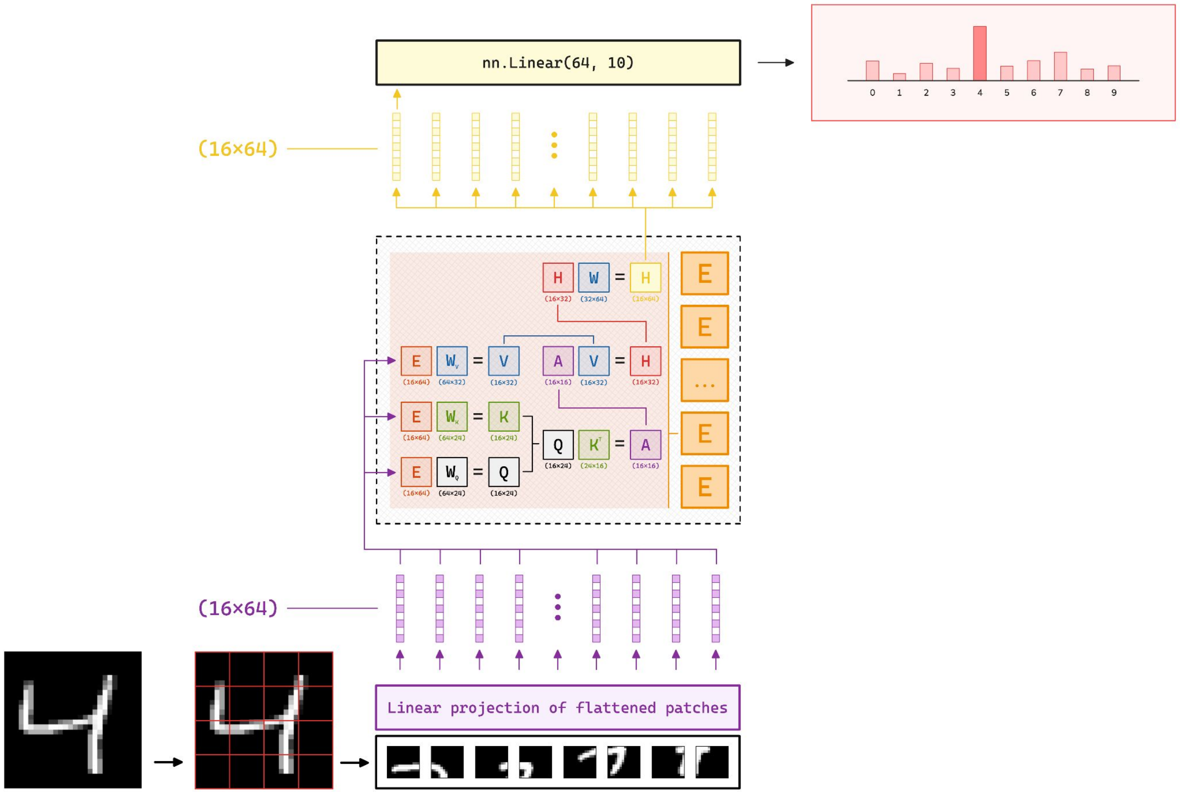


Encoders



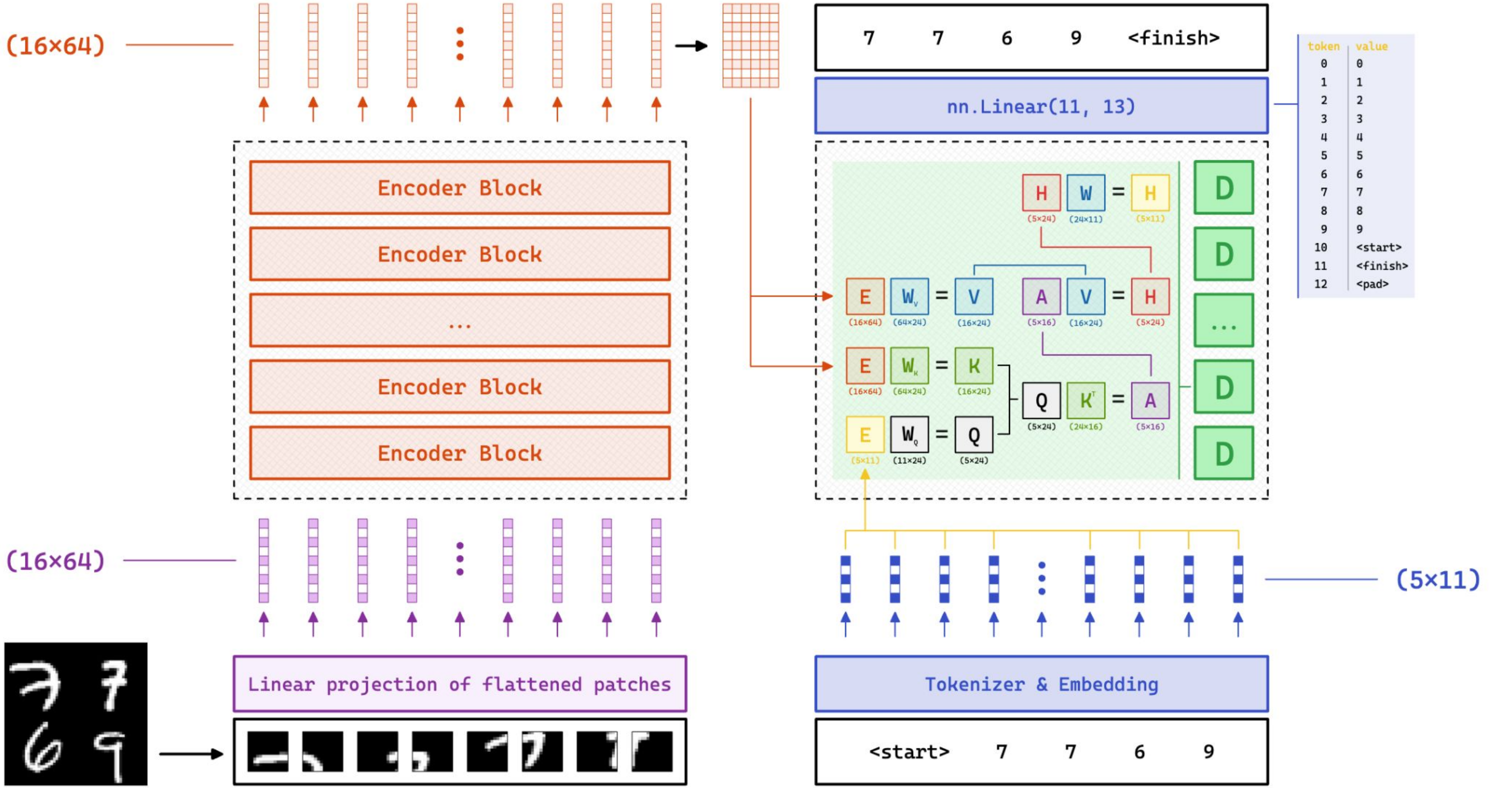
Encoded Representations
Encoder Block
Encoder Block
...
Encoder Block
Encoder Block
Encoder
Embedding Layer
Tokenizer & Embedding
Lorem ipsum dolor sit amet ...
Linear projection of flattened patches
Convolution 1D over the Y axis
Spatiotemporal Feature Aggregation
Feature Extractors

Encoder



nn.Linear(64, 10)
0 1 2 3 4 5 6 7 8 9
(16×64)
H (16×32)
W (32×64)
H (16×64)
E (16×64)
W_V (64×32)
V (16×32)
A (16×16)
V (16×32)
H (16×32)
E (16×64)
W_K (64×24)
K (16×24)
Q (16×24)
K^T (24×16)
A (16×16)
E (16×64)
W_Q (64×24)
Q (16×24)
E
E
...
E
E
(16×64)
Linear projection of flattened patches

Transformer



Papers

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

Attention Is All You Need

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

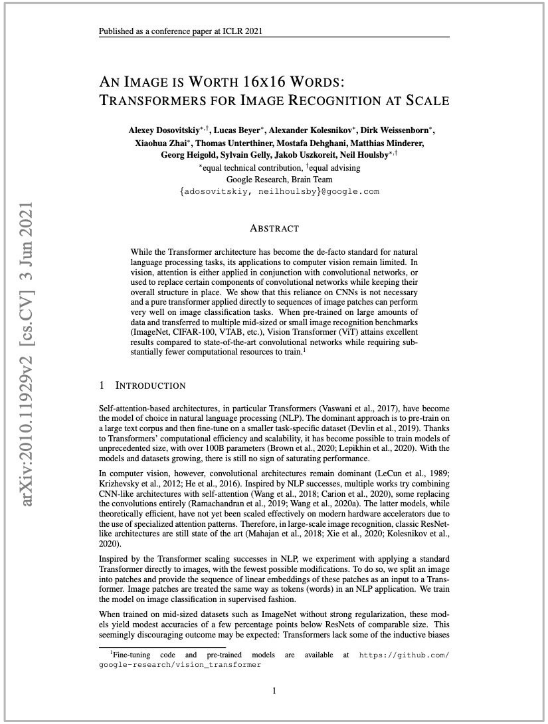
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.
†Work performed while at Google Brain.
‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Vaswani et al. (NeurIPS 2017)



Published as a conference paper at ICLR 2021

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

arXiv:2010.11929v2 [cs.CV] 3 Jun 2021

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]

1 Introduction

Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

[1]Fine-tuning code and pre-trained models are available at https://github.com/google-research/vision_transformer

1

Dosovitskiy et al. (ICLR 2021)

Suggestions

- do not just copy-paste from ChatGPT

- understand and practice PyTorch ops
- use google colab for little snippets
- pair programming, swap pairs within the team
- watch tutorials but make sure to talk
- no need for GPUs yet, do everything local

FOUNDERS AND CODERS

Good luck!

Machine Learning Institute

Recurrent Rebels

Adam Beedell
Aparna Pillai
Helen Zhou
Esperanza Shi

Gradient Gigglers

Clement Ha
Andrew
Umut Sagir
Jacob Jenner

Overfitting Overlords

Nikolas Kuhn
David Edev
Peter O'Keeffe
Miguel Parracho

Hyperparameter Hippies

Tyrone Nicholas
Dan Goss
Anton Dergunov
Ben Liong

Perceptron Party

Tao Zamorano
Tomas Krajcoviech
James Carter
Charles Cai

Backprop Bunch

Ethan Edwards
Melanie Wong
Kadriye Turkcan
Ben Williams

Dropout Disco

James Yan
Andrei Zhirnov
Prima Gouse
Jingyan Chen

Kernel Kittens

Joao Esteves
Hikaru Tsujimura
Rosh Beed
Felipe Lavratti

Bayesian Buccaneers

Rasched Haidari
Ewan Beattie
Marcin Tolysz
Maria Sharif

Feature Fiestas

Ben Bethell
Arjuna James
Yali Pan
Ayman Abbas