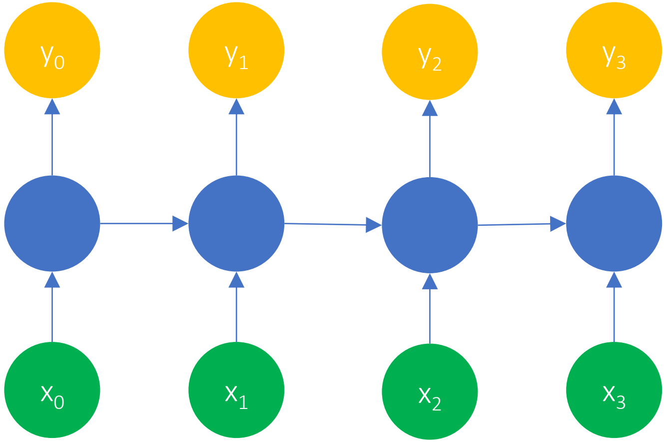
Transformers

Why another Architecture?

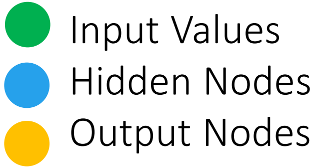
About the need for improvement in Recurrent Neural Networks

- RNNs work sequential
- Example: Natural Language Processing
  - Words are processed one by one – hard to parallelize
  - Order of words important, e.g.

    „Alice loves Bob“ vs. „Bob loves Alice“
  - Problems with larger sequences (forgetting of past information)
  - hard to train (vanishing or exploding gradients)



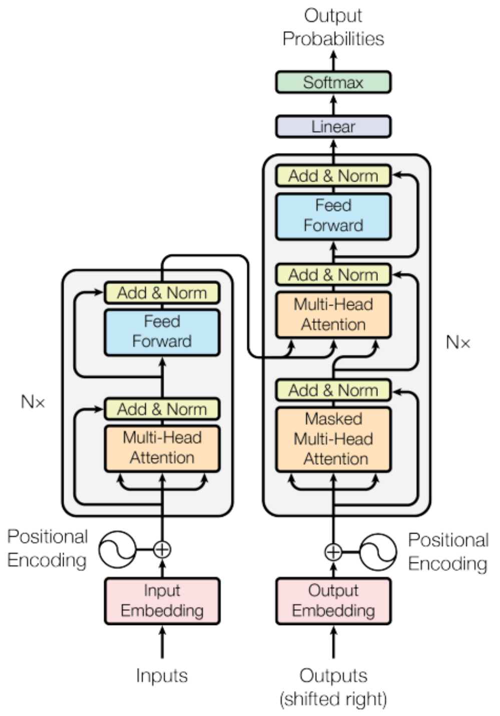
Unrolled RNN



Transformers

Introduction

- Developed in 2017 at Google with focus on translation
- Main benefits:
  - Keeps track of word order
  - No vanishing or exploding gradients
  - Training can be parallelized
  - Allows for huge models

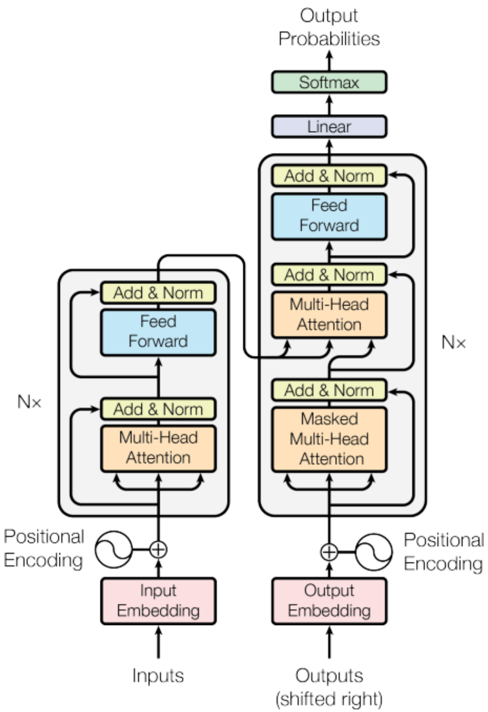


Model Architecture based on paper:
Vaswani et. Al. „Attention is all You Need“

Transformers

How does it work?

- Three main Features:
  - Positional encoding
  - Attention
  - Self-Attention



Model Architecture based on paper:
Vaswani et. Al. „Attention is all You Need“

Transformers

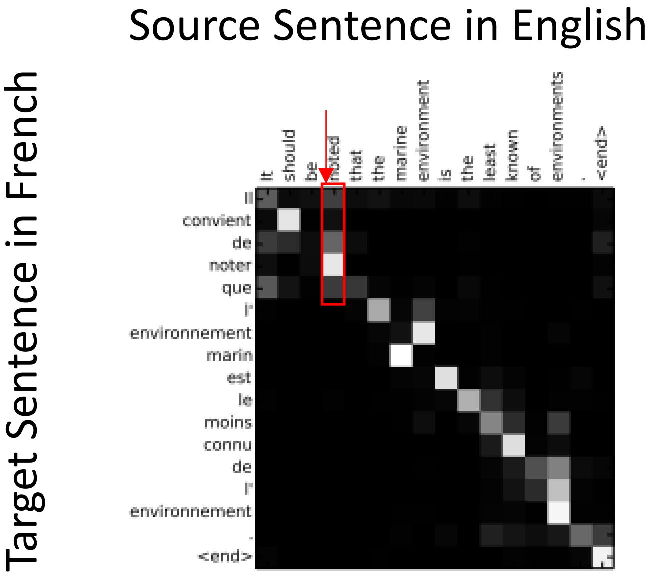
Positional Encoding

- Natural Language Processing word order important.
- RNN:
  - word order understanding based on sequential pass of words
- Transformers:
  - Use positional encoding
  - Simplified example: “I like to code” → [(“I”, 1), (“like”, 2), (“to”, 3), (“code”, 4)]
  - In original paper sine and cosine used for encoding

Transformers

Attention

- Attention concept introduced in 2015
- Attention focused on each word in source sentence to come up with translation
- Interdependency between words learned during training
- Helps to learn word order, plurality or grammar



Bahdanau, Cho, Bengio: "Neural Machine Translation by Jointly Learning to Align and Translate"

Transformers

Self-Attention

- Helps to improve internal representation
- Example of word with different meanings:
    - Orange is created by mixing yellow and red.
    - An orange is a fruit of various citrus species.
- Self-attention helps to get better context understanding.

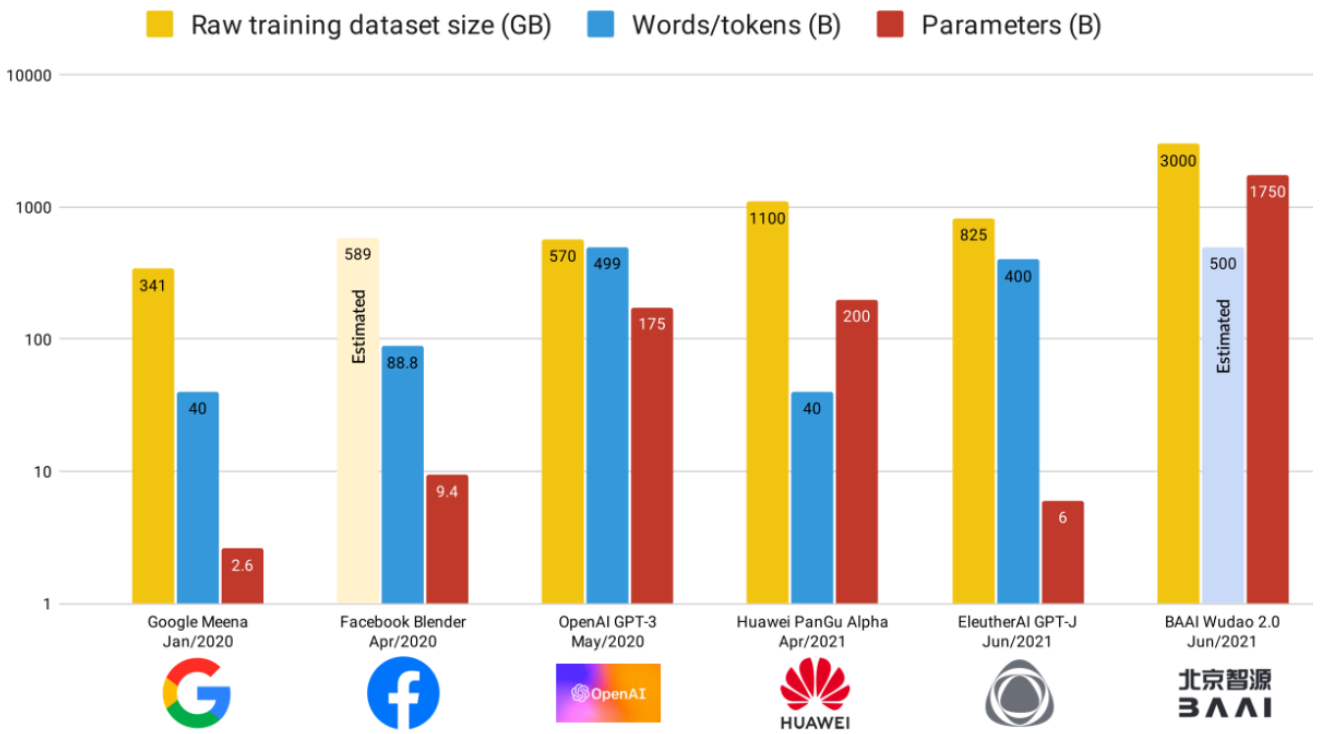
Transformers

Examples

- Google “BERT”
  - Bidirectional Encoder Representations from Transformers
  - 110 Million parameters
- OpenAI “GPT-3”
  - Generative Pre-trained transformers 3
  - 175 Billion parameters
- Google “LaMDA”
  - Language Model for Dialogue Applications
  - 137 Billion parameters

Transformers

Examples



Source: https://i.redd.it/lq69ol56kk971.png

Transformers

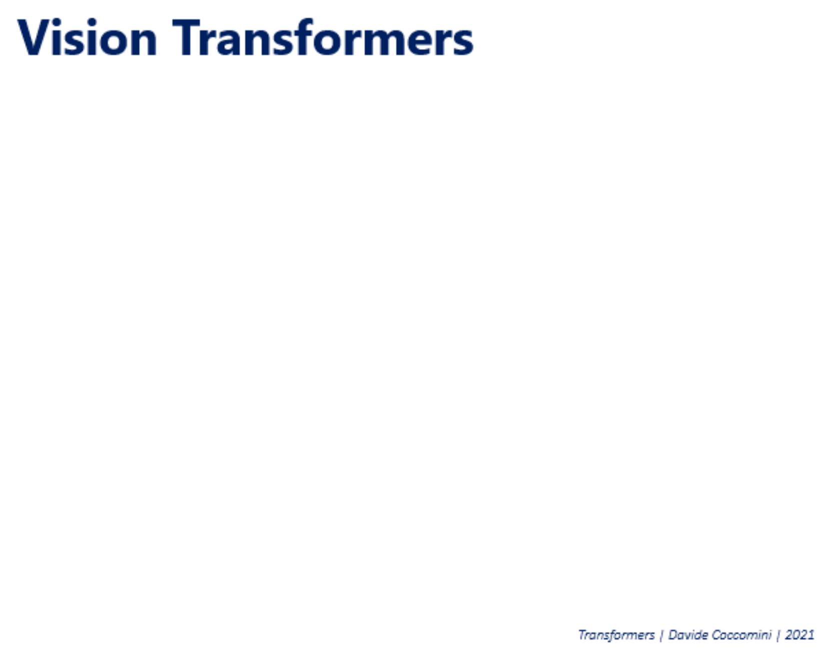
Domains of Expertise

- NLP
  - Text summarization
  - Classification
  - Sentiment analysis
- Computer Vision
- Time-Series Prediction

Transformers

Vision Transformers (ViT)

- Pixel - basic unit
- Relationships between pixels unfeasible
- Image sections analyzed (positional encoding)
- Relationship for sections calculated
- Patch size, e.g. 16x16
- Stride, e.g. 16x16
- Overlaps of images allowed



Source: https://en.wikipedia.org/wiki/Vision_transformer#/media/File:Vision_Transformer.gif

Transformers

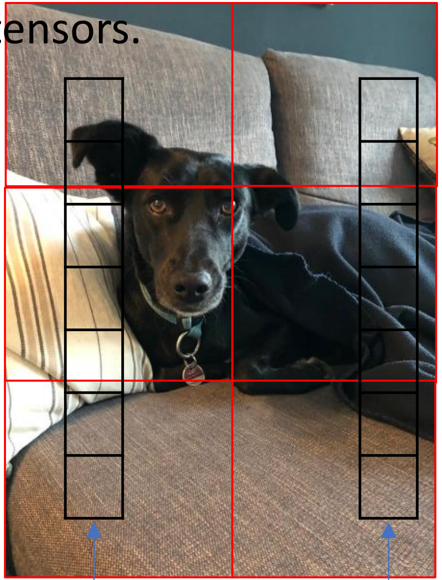
Vision Transformers (ViT)

Segments are flattened to 1D tensors.

Positional Encoding (1)

(2)

(3)



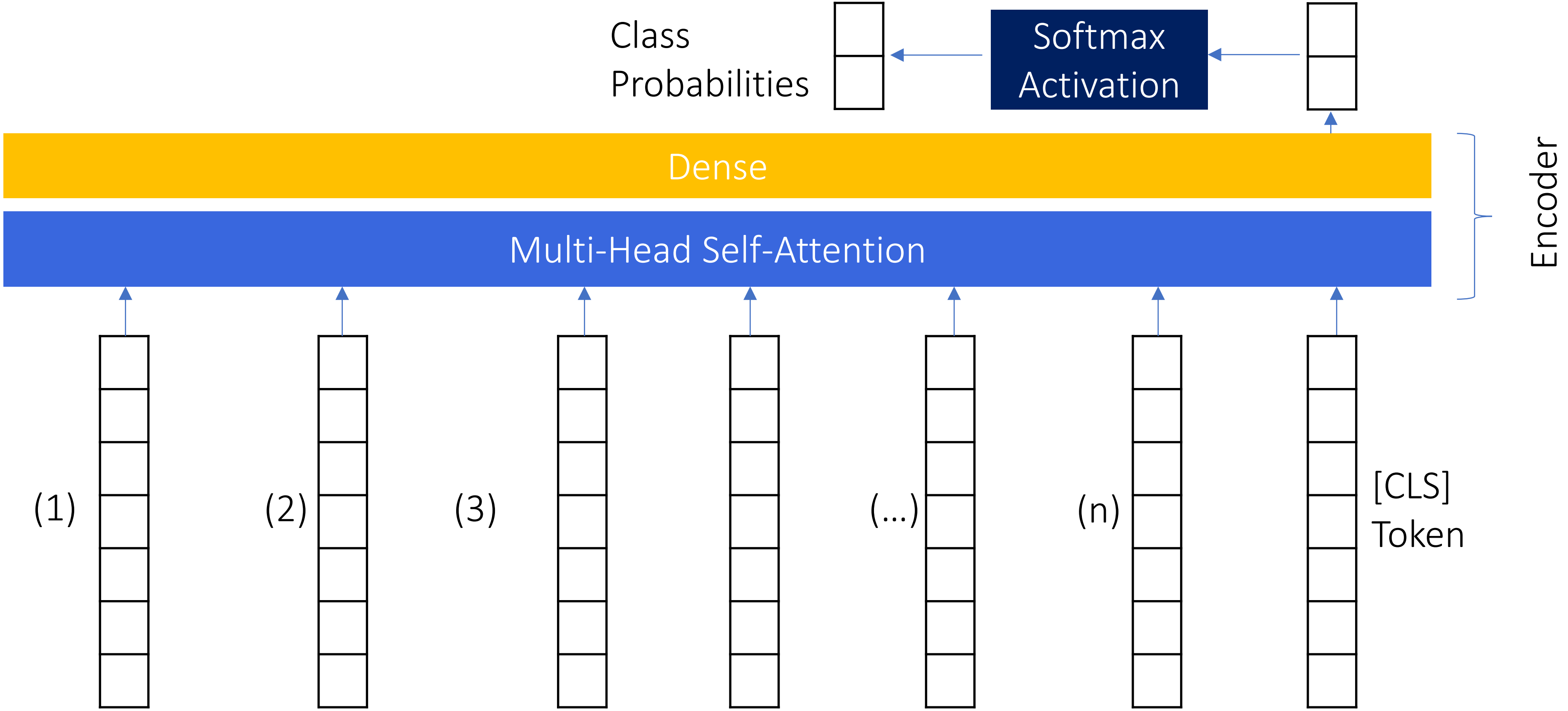
(...)

(n)

[CLS] Token

Transformers

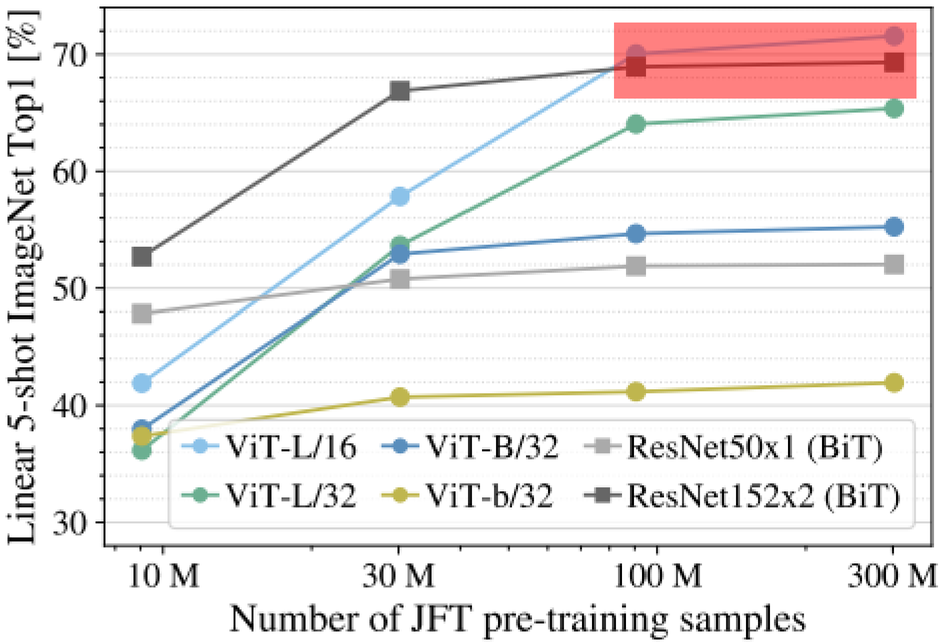
Vision Transformers (ViT)



Transformers

Vision Transformers (ViT)

- ViT outperform CNNs, but only with >100M images
- JFT (Imagenet dataset with up to 300 Million images and 18.000 classes, not public)



Source: https://arxiv.org/pdf/2010.11929.pdf,
Dosovitskiy et. al. „An image is worth 16x16 words: Transformers for Image Recognition at scale.“