Deep Learning

- Transformers -

Eunbyung Park

Assistant Professor

School of Electronic and Electrical Engineering

Eunbyung Park (silverbottlep.github.io)

Transformers

- Revolutionize NLP
- Attention based model
- State-of-the-art methods in all NLP tasks and many vision tasks
- >20,000 citations in 4 years

Attention Is All You Need

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

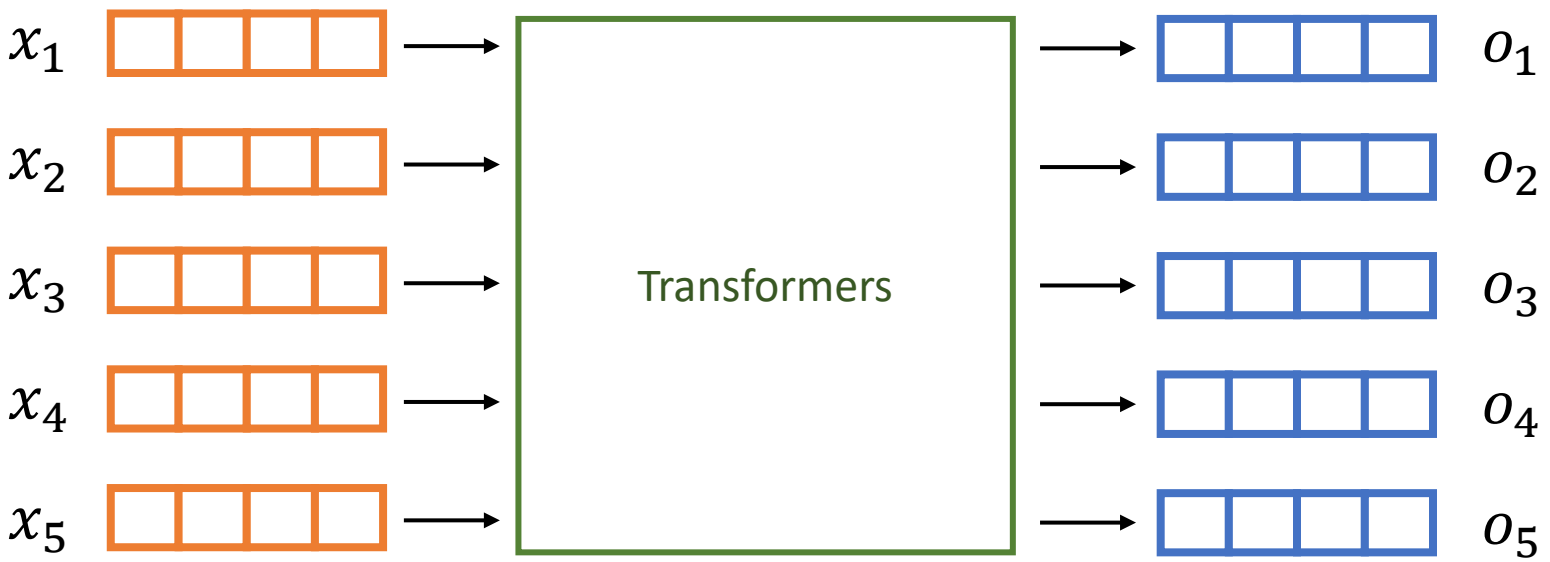
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

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 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

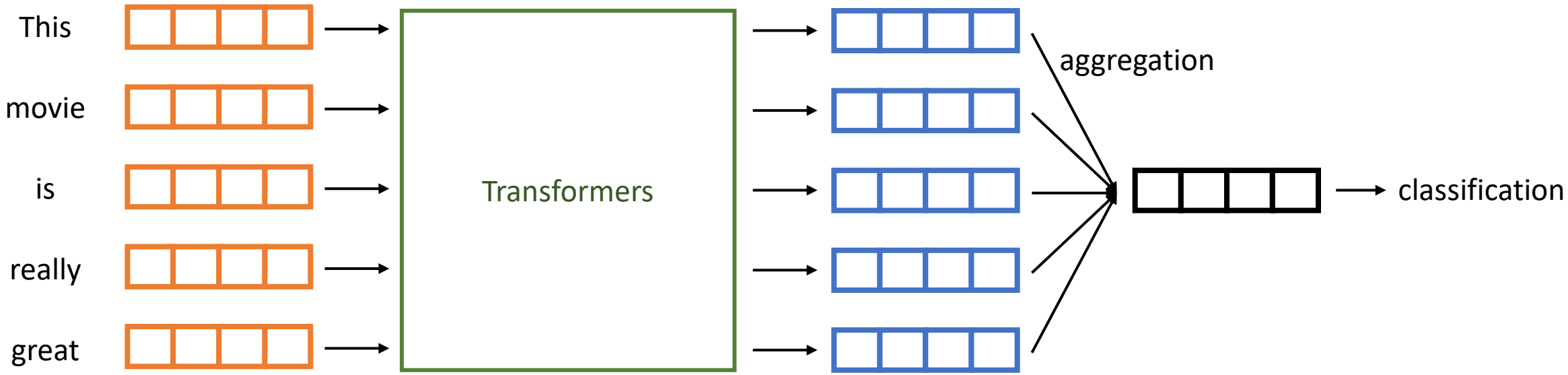
Transformers

- ***Transforms*** Input vectors to output vectors
- Attentional modeling



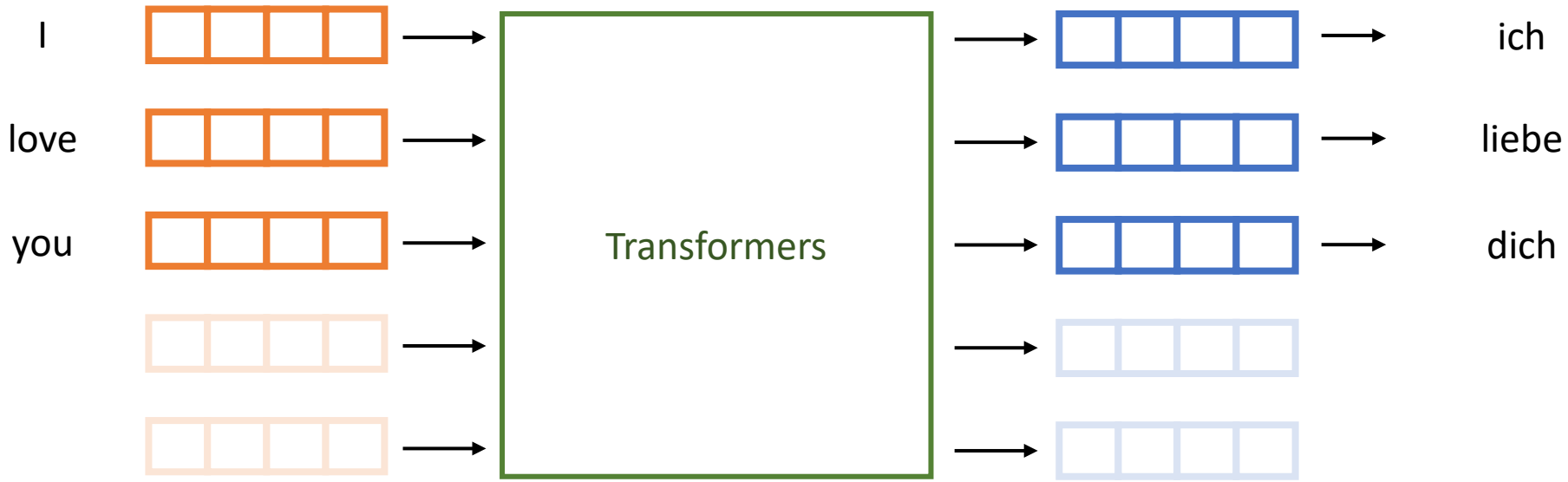
Transformers Applications

- Sentence Classification



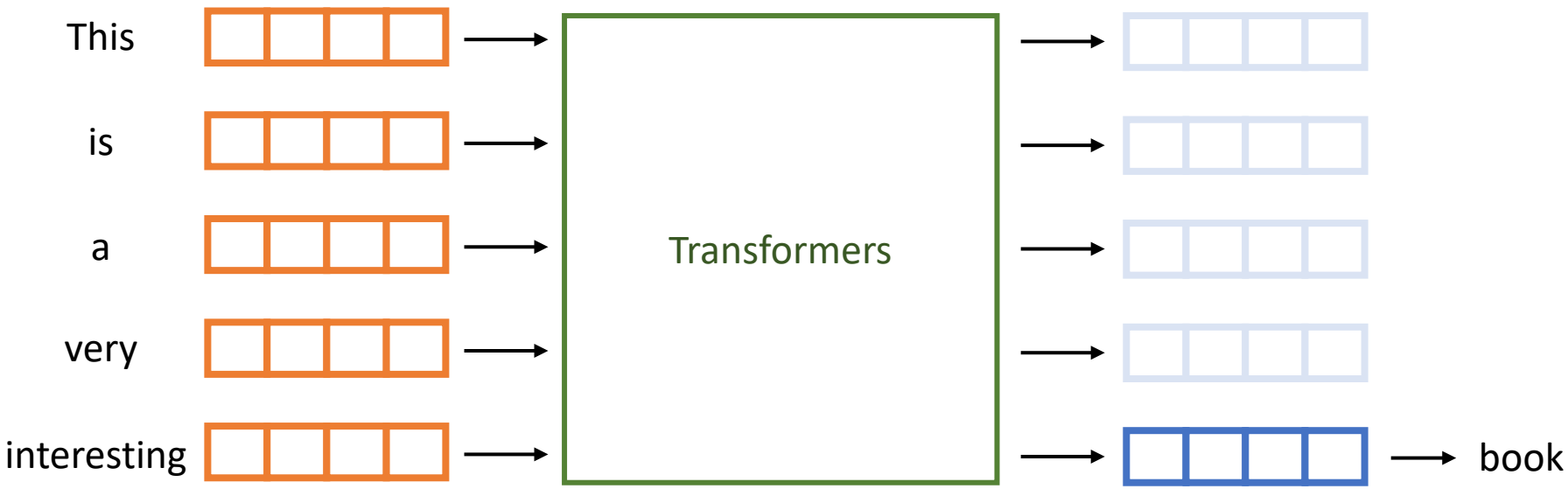
Transformers Applications

- Machine Translation



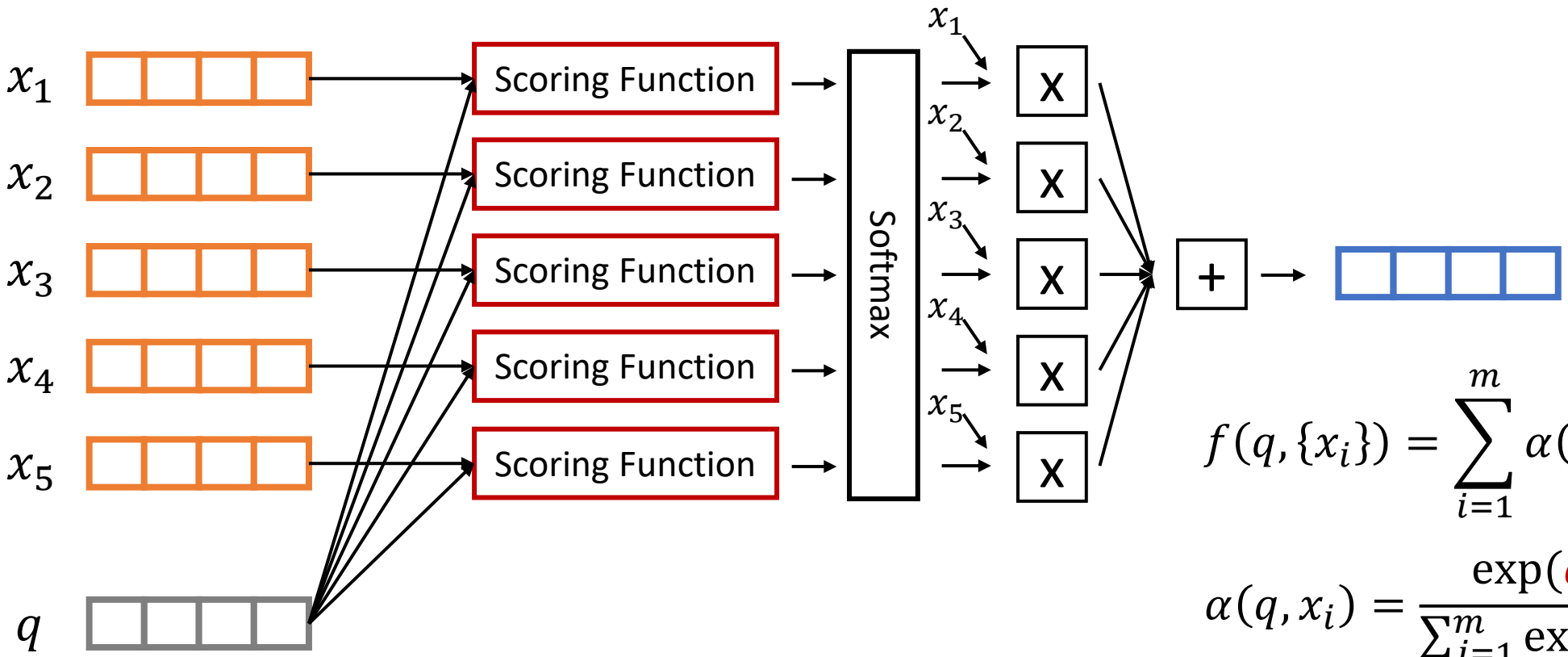
Transformers Applications

- Language generation (next word prediction)



Attention

- Given the query, extract and aggregate relevant information



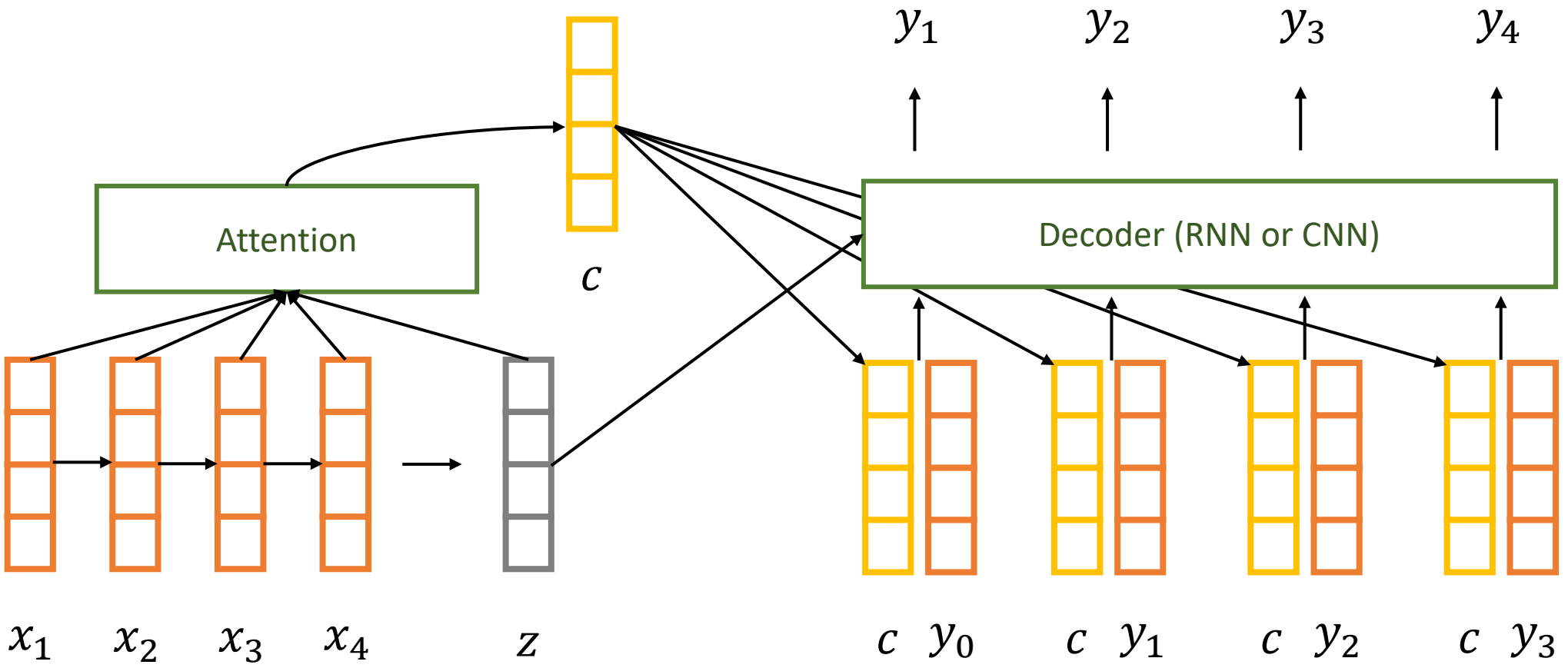
$$f(q, \{x_i\}) = \sum_{i=1}^{m} \alpha(q, x_i) x_i$$

$$\alpha(q, x_i) = \frac{\exp(a(q, x_i))}{\sum_{j=1}^{m} \exp(a(q, x_j))}$$

$$a(q, x) = q^\top x$$

Attention Model Example

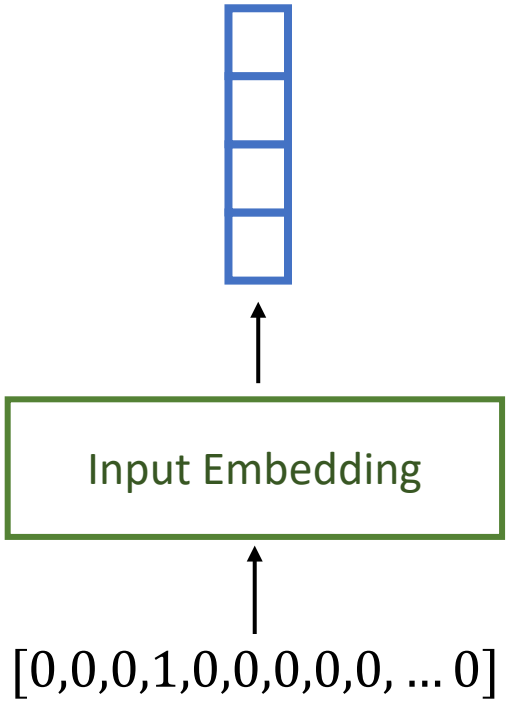
- Attention for sequence-to-sequence
  - E.g. machine translation

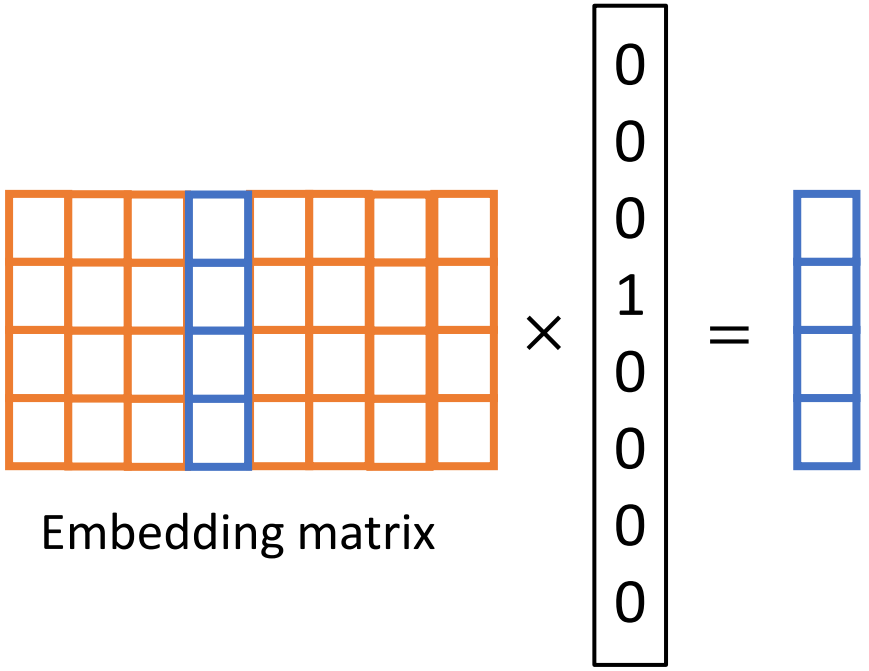


Bahdanau et al, Neural machine translation by jointly learning to align and translate, ICLR 2015

Input Embedding

- Mapping ‘one-hot’ to ‘vectors’
  - Input token -> one-hot encoding -> vector



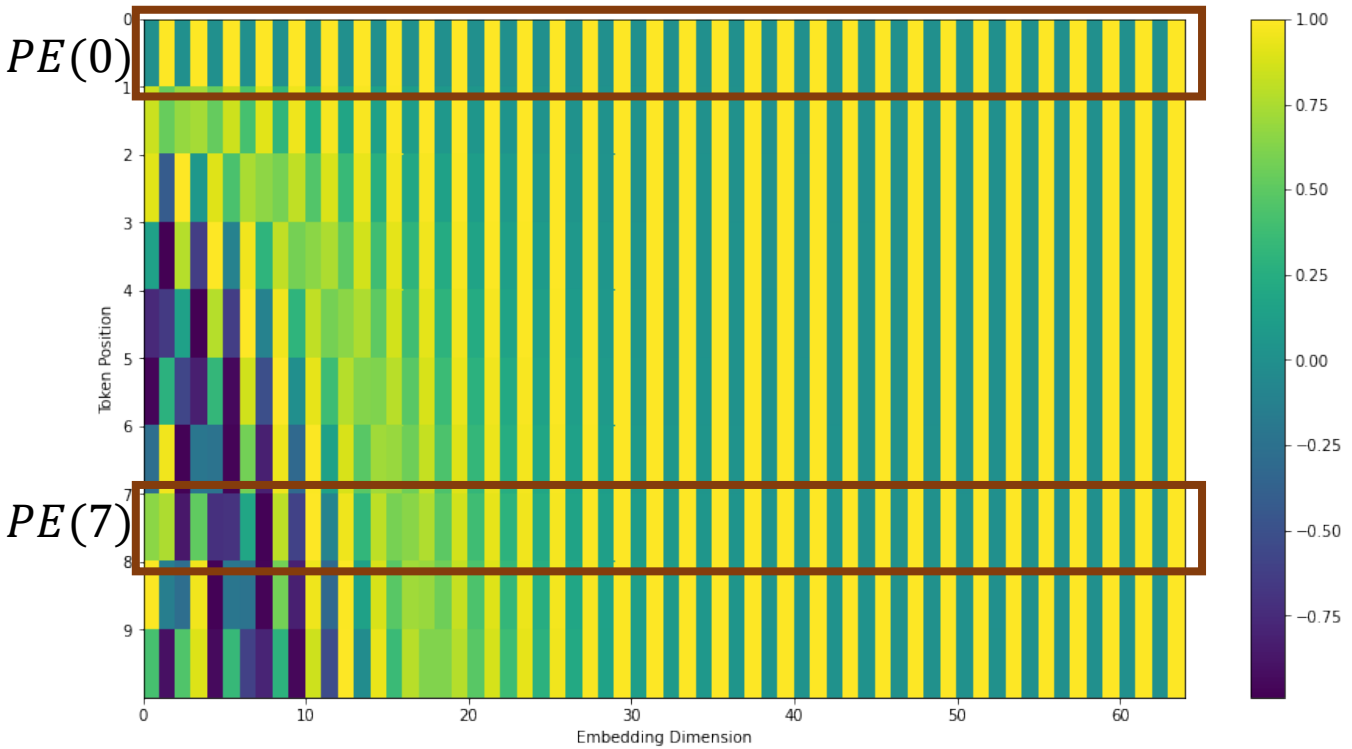


Positional Encoding

- Transformers are ‘orderless’ architecture
  - Additional time (order) information are needed

$$PE(t) = [\dots, \cos\left(2\pi\sigma^{j/m}\mathrm{t}\right), \sin\left(2\pi\sigma^{j/m}t\right), \dots]^{\mathrm{T}}$$

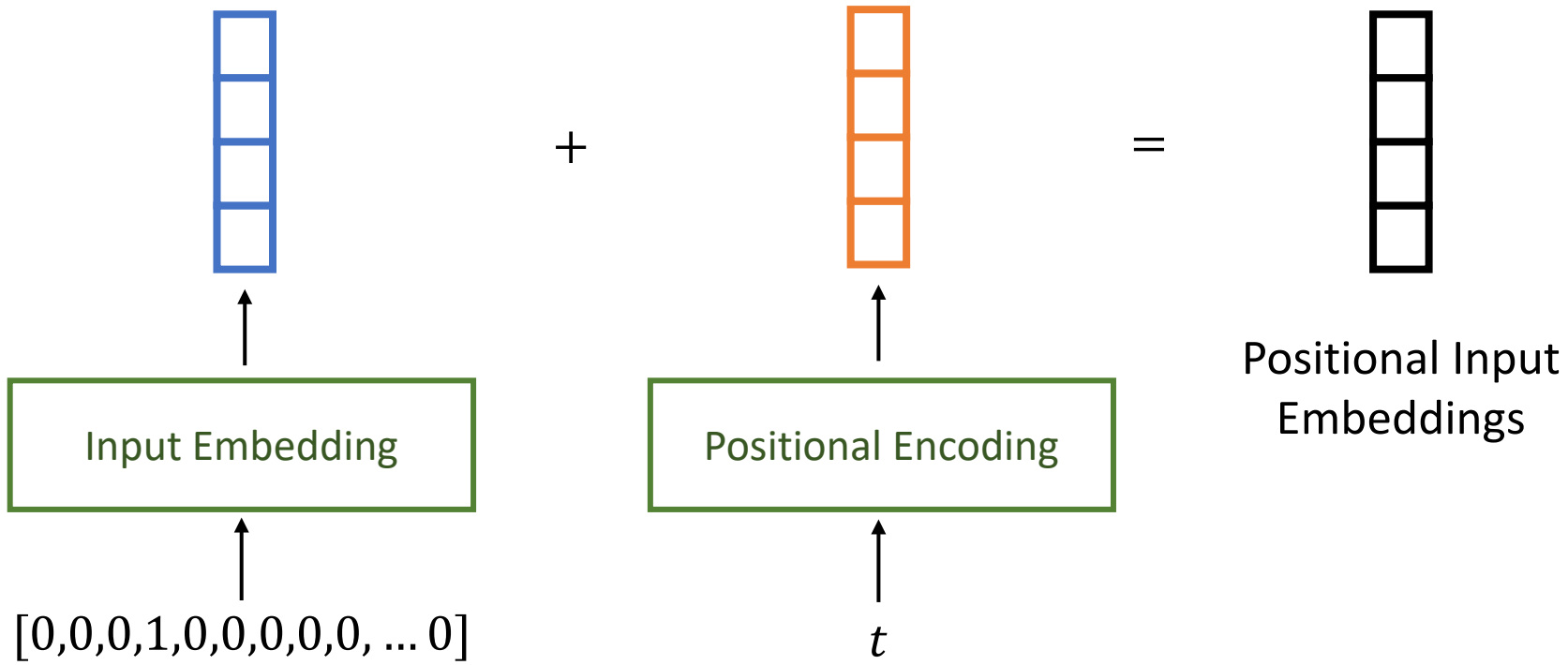
$$(\text{for } j = 0, \dots, m-1)$$



The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)

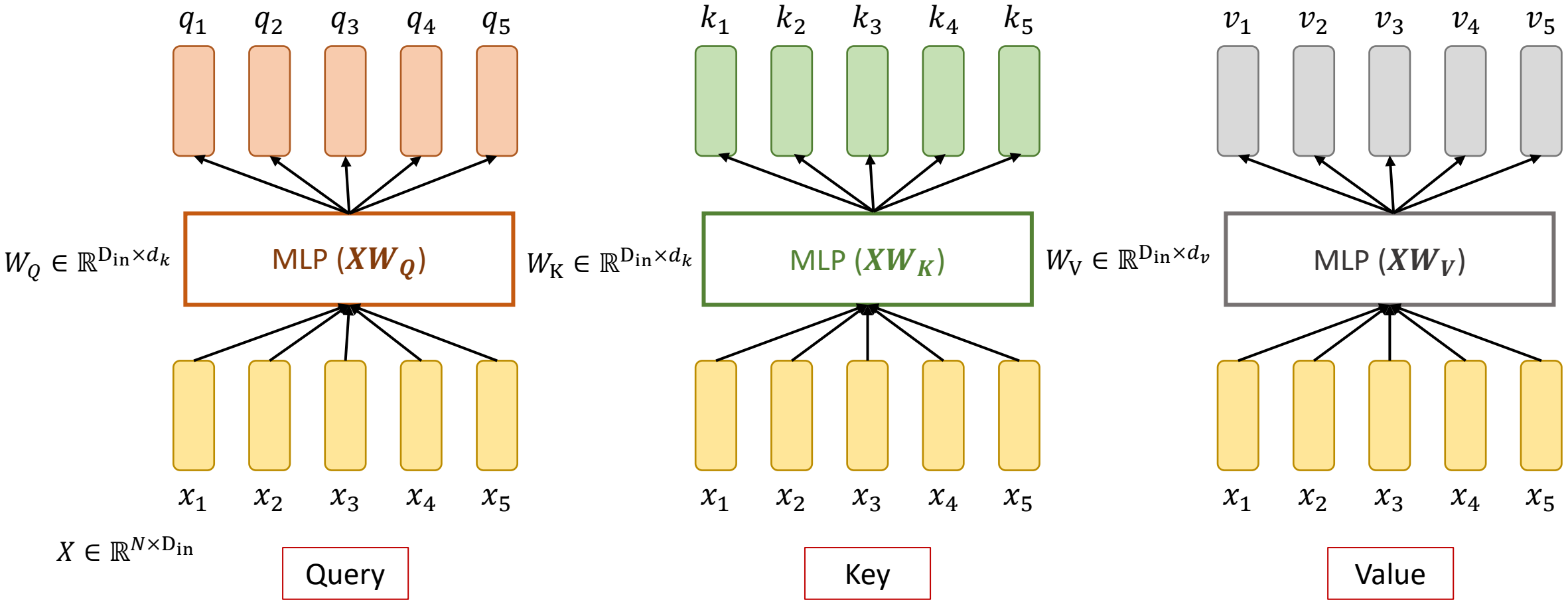
Positional Input Embeddings

- Function of input embeddings and positional encodings



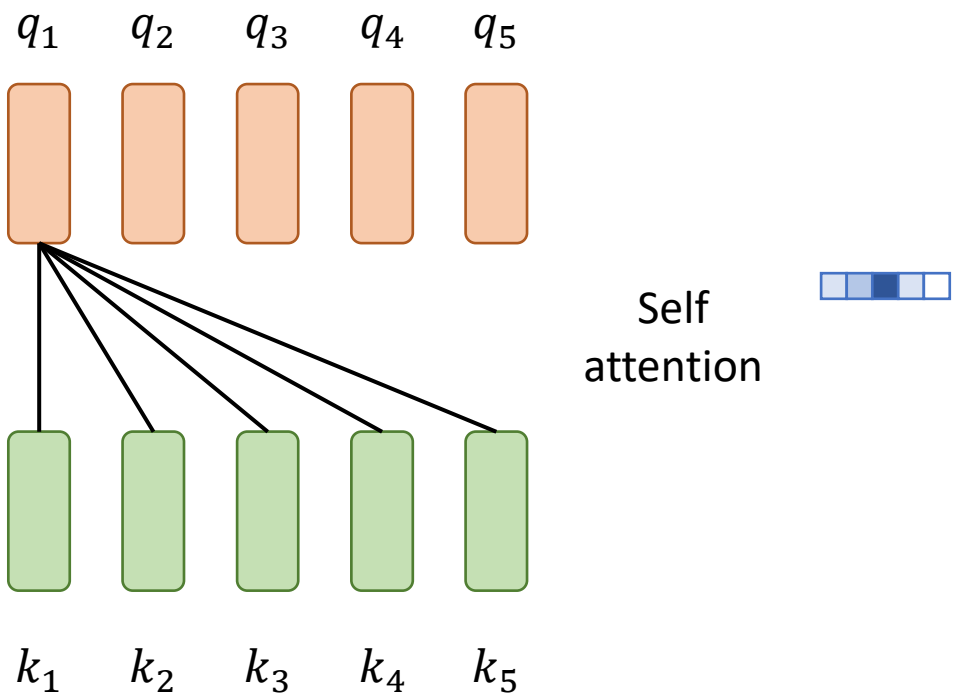
Self-Attention

- (Input) -> (query, key, value)



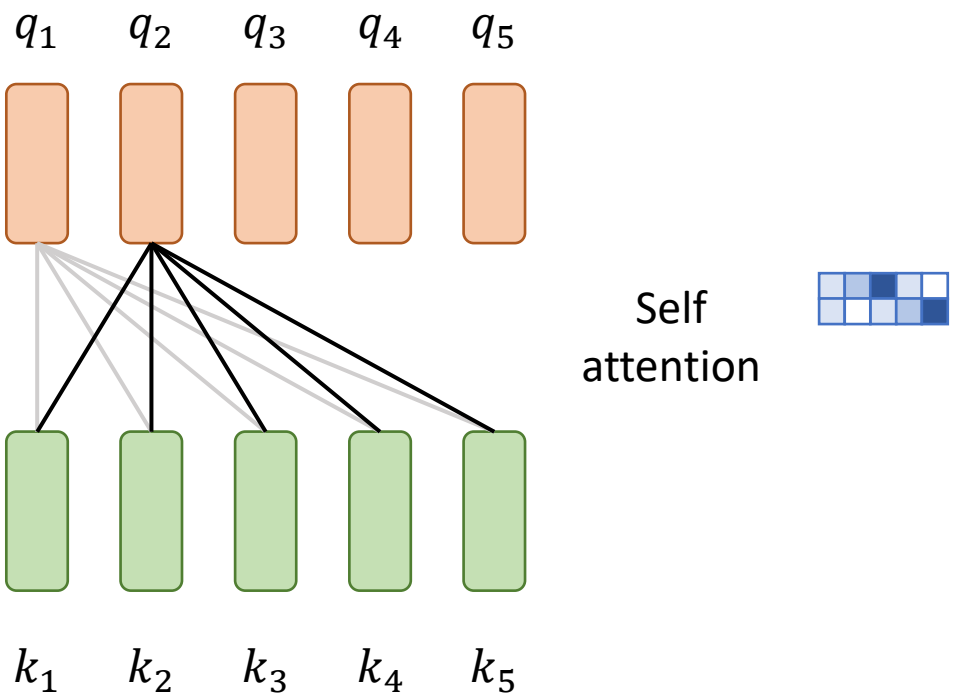
Self-Attention

- Attention over each other



Self-Attention

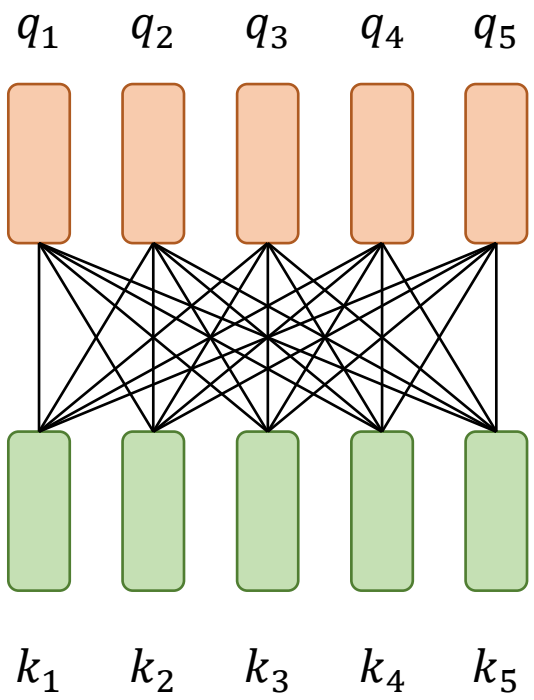
- Attention over each other



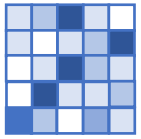
Self attention

Self-Attention

- Attention over each other



Self attention



$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)$$

$$Q \in \mathbb{R}^{N \times d_k}$$

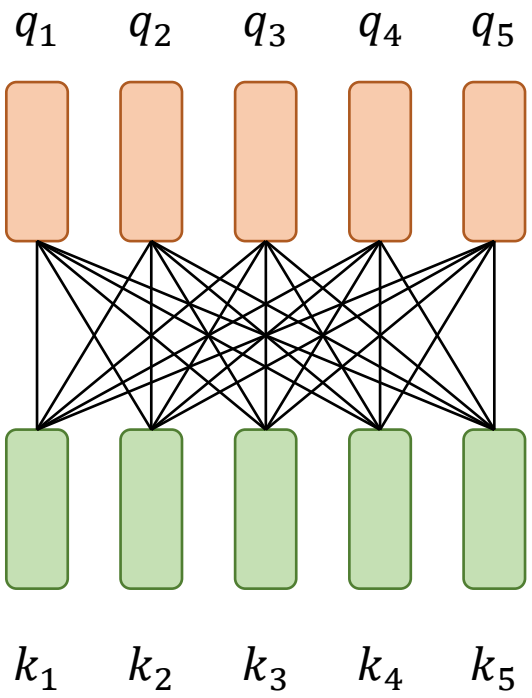
$$\mathrm{K} \in \mathbb{R}^{N \times d_k}$$

$$\sum_j A_{ij} = 1$$

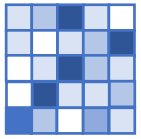
Row-wise softmax

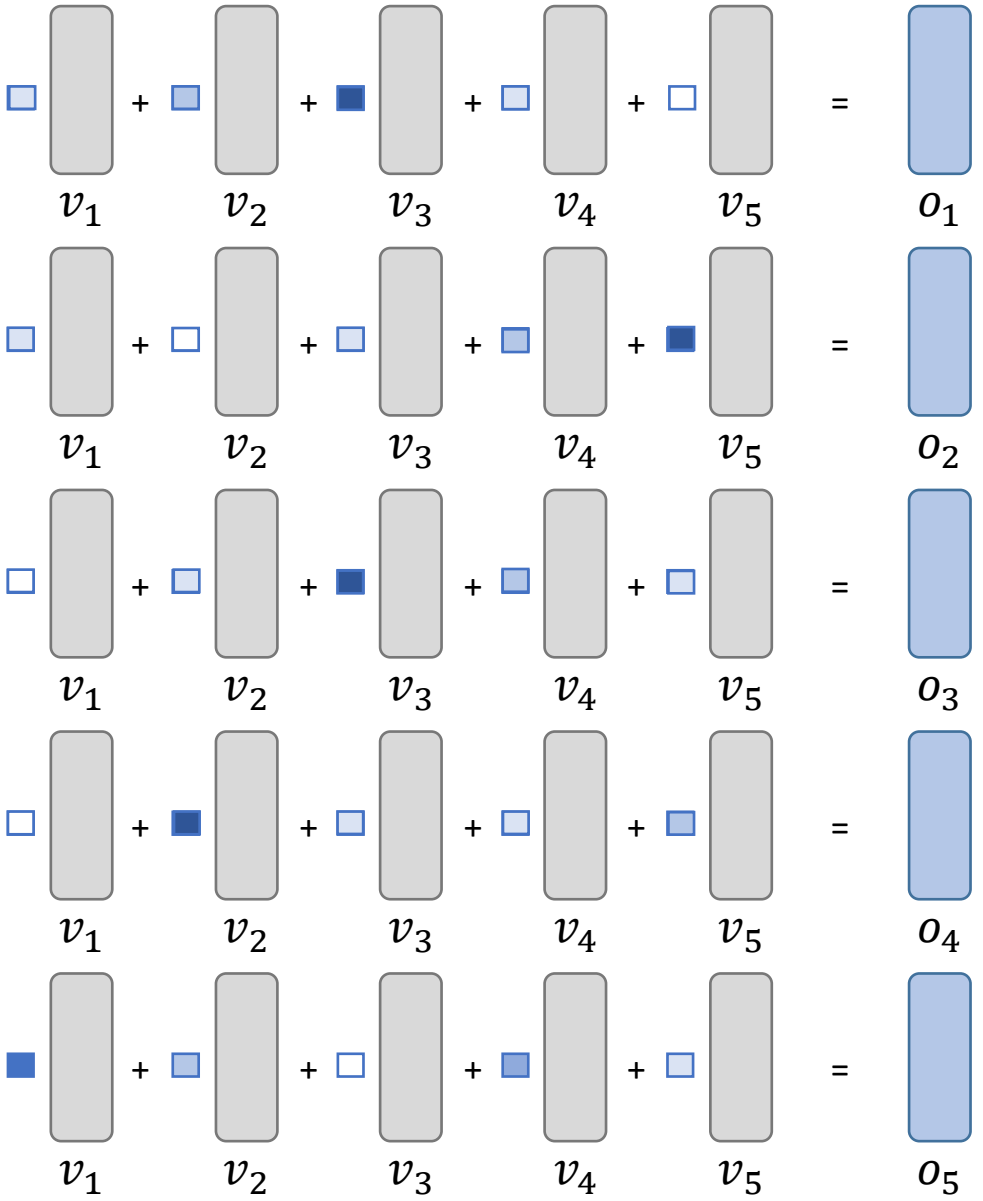
Self-Attention

- Attention over each other



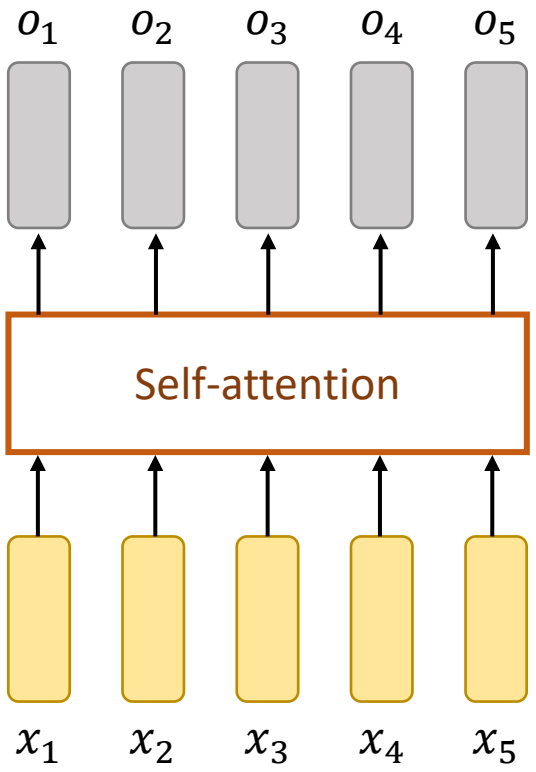
Self attention





Self-Attention

- Attention over each other



$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)$$

$$\text{SA}(X; W_Q, W_K, W_V) = AV$$

$$Q = XW_Q$$

$$\text{K} = XW_K$$

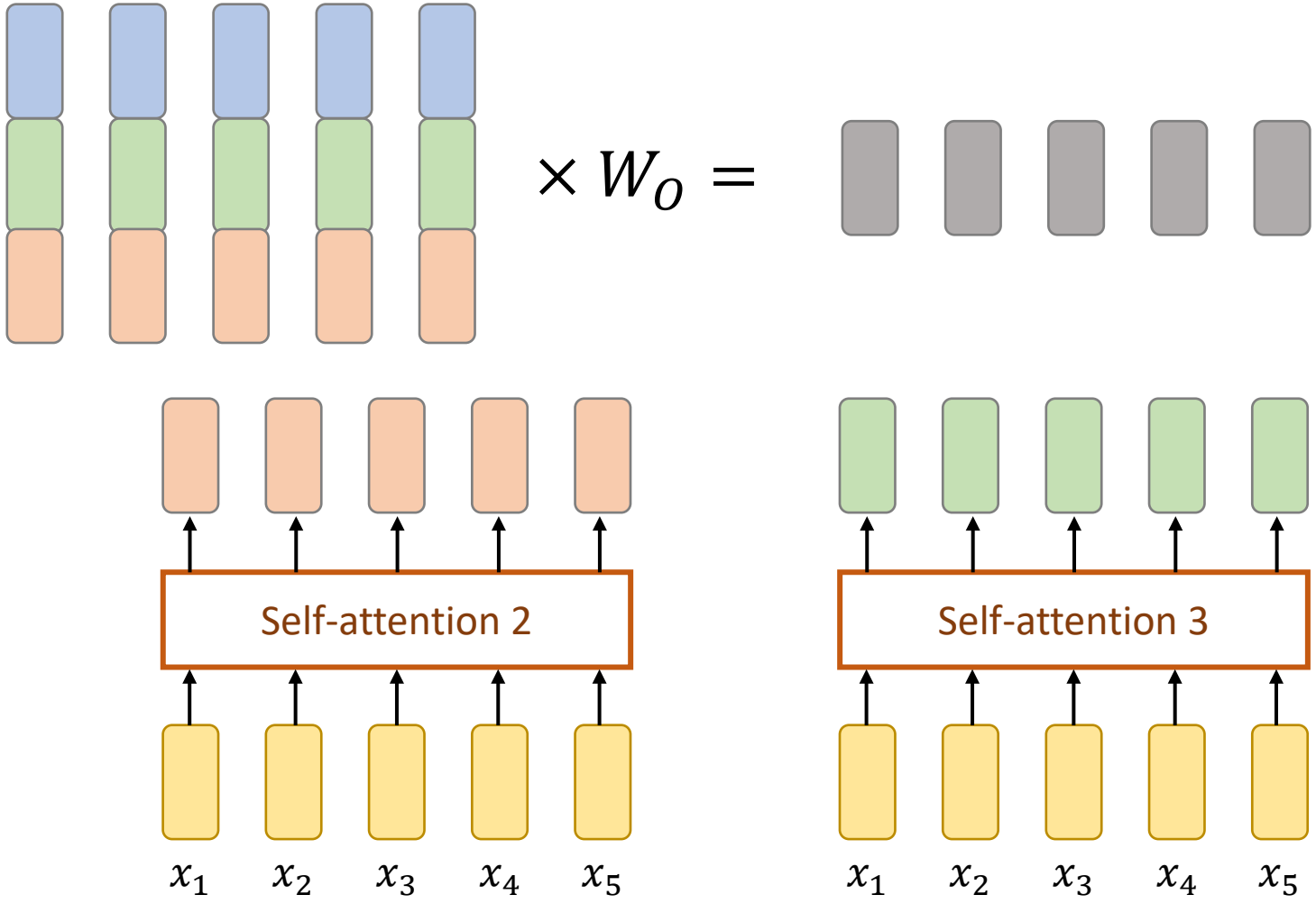
$$\text{V} = XW_V$$

d_k: dimension of query and keys

Multi-Head Attention

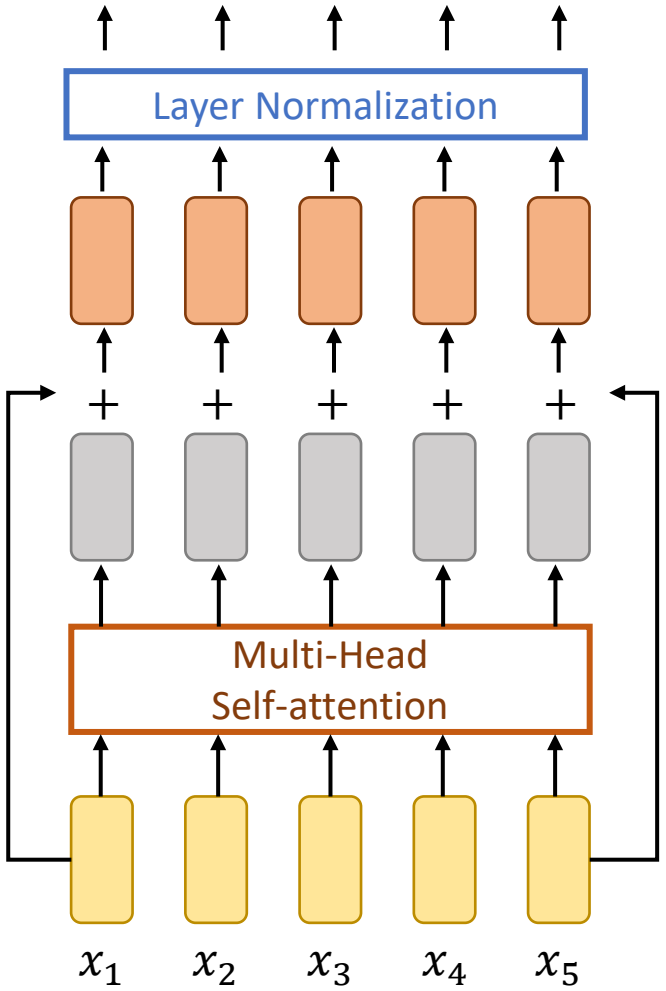
- Multiple self-attention modules

$W_O \in \mathbb{R}^{d_v N_h \times d_o}$



Multi-Head Attention

- Multi-head + residual connection + layer normalization



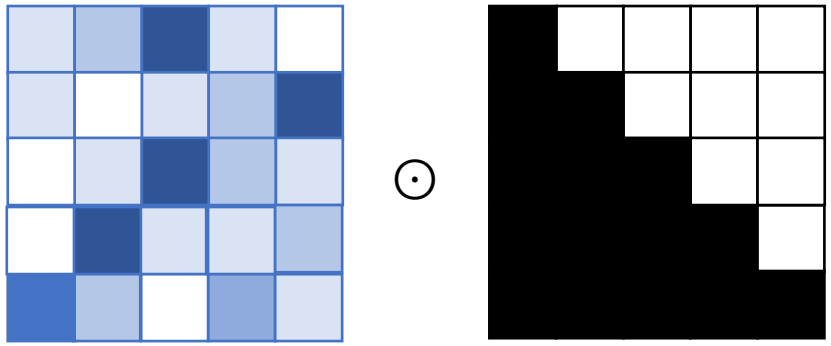
$$LN_{\gamma,\beta}(x_i) = \gamma\hat{x} + \beta$$

$$\hat{x}_{i,k} = \frac{x_{i,k} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

$$\mu_i = \frac{1}{K}\sum_{k=1}^{K} x_{i,k}\,, \qquad \sigma_i^2 = \frac{1}{K}\sum_{k=1}^{K}\left(x_{i,k} - \mu_i\right)^2$$

Transformers for Language Modeling

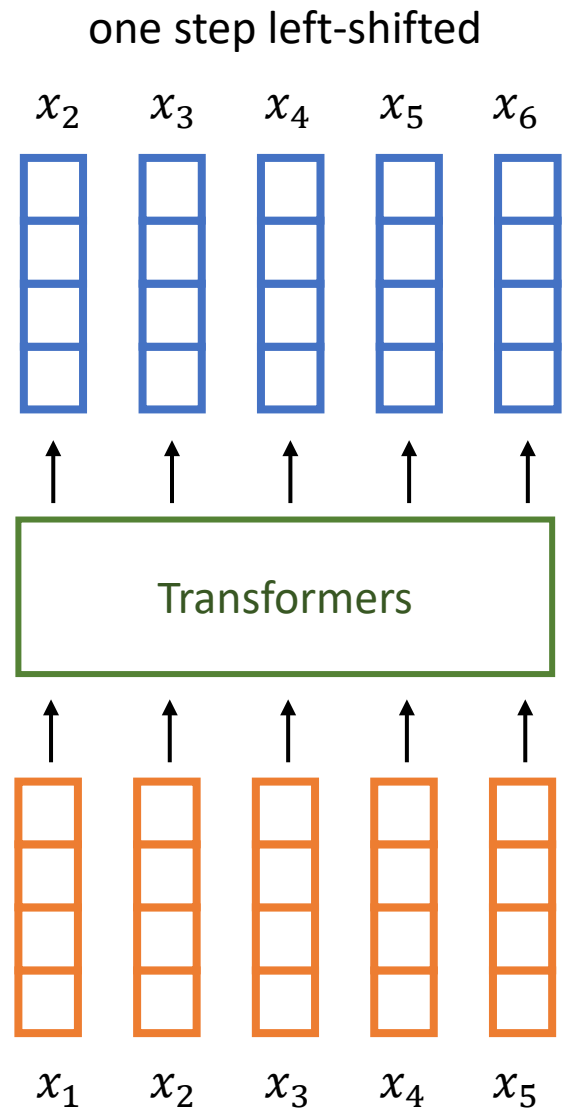
- Language modeling as next word prediction
  - Autoregressive generation
  - Training is parallelizable w/ attention mask



Attention Score ⊙ Attention Mask

$p(x_2|x_1)$
$p(x_3|x_2, x_1)$
$p(x_4|x_3, x_2, x_1)$
$p(x_5|x_4, x_3, x_2, x_1)$
$p(x_6|x_5, x_4, x_3, x_2, x_1)$

1 batch -> 5 training examples



GPT3

- An autoregressive transformer language model w/ 175 billion parameters

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

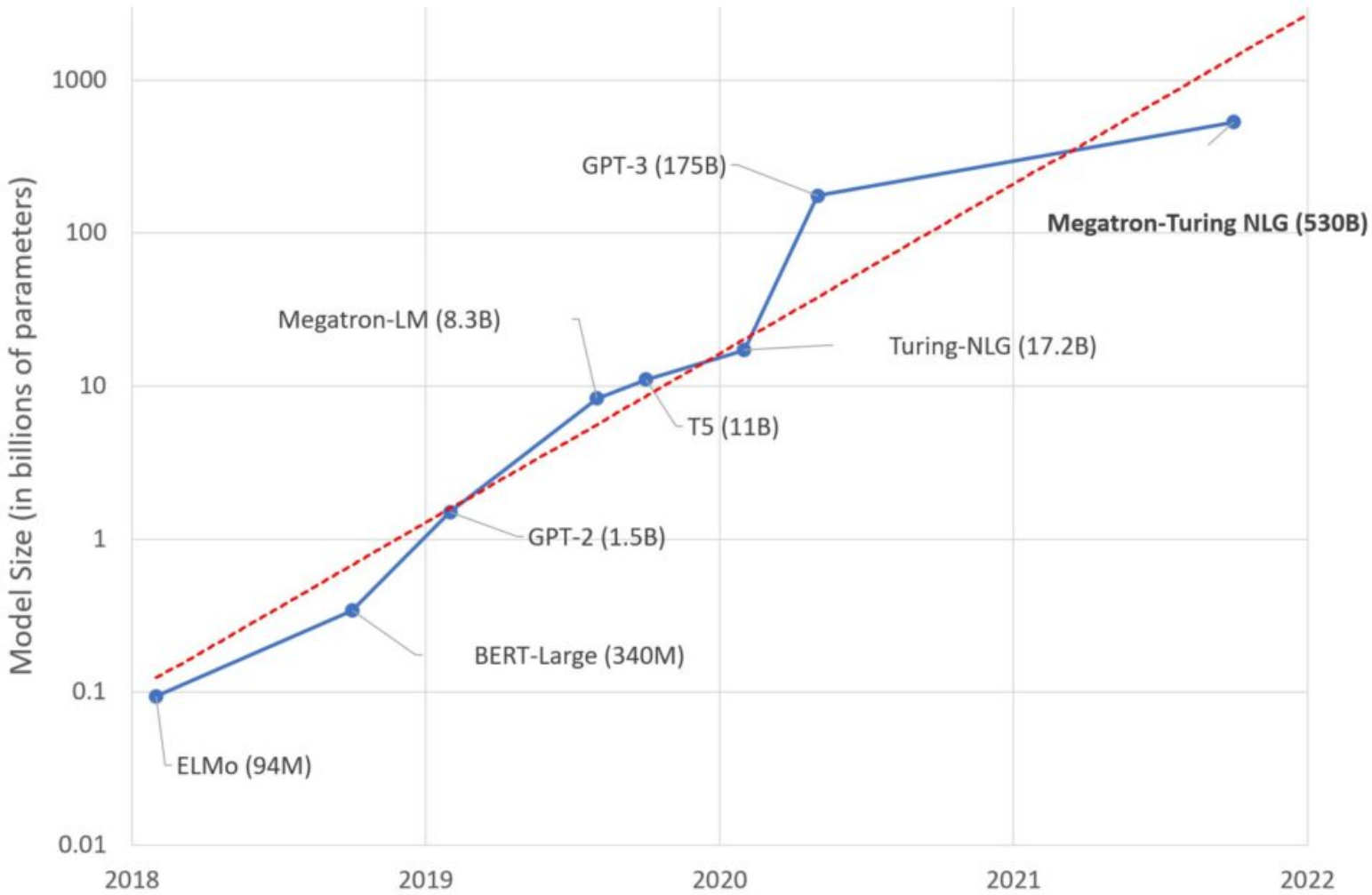
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

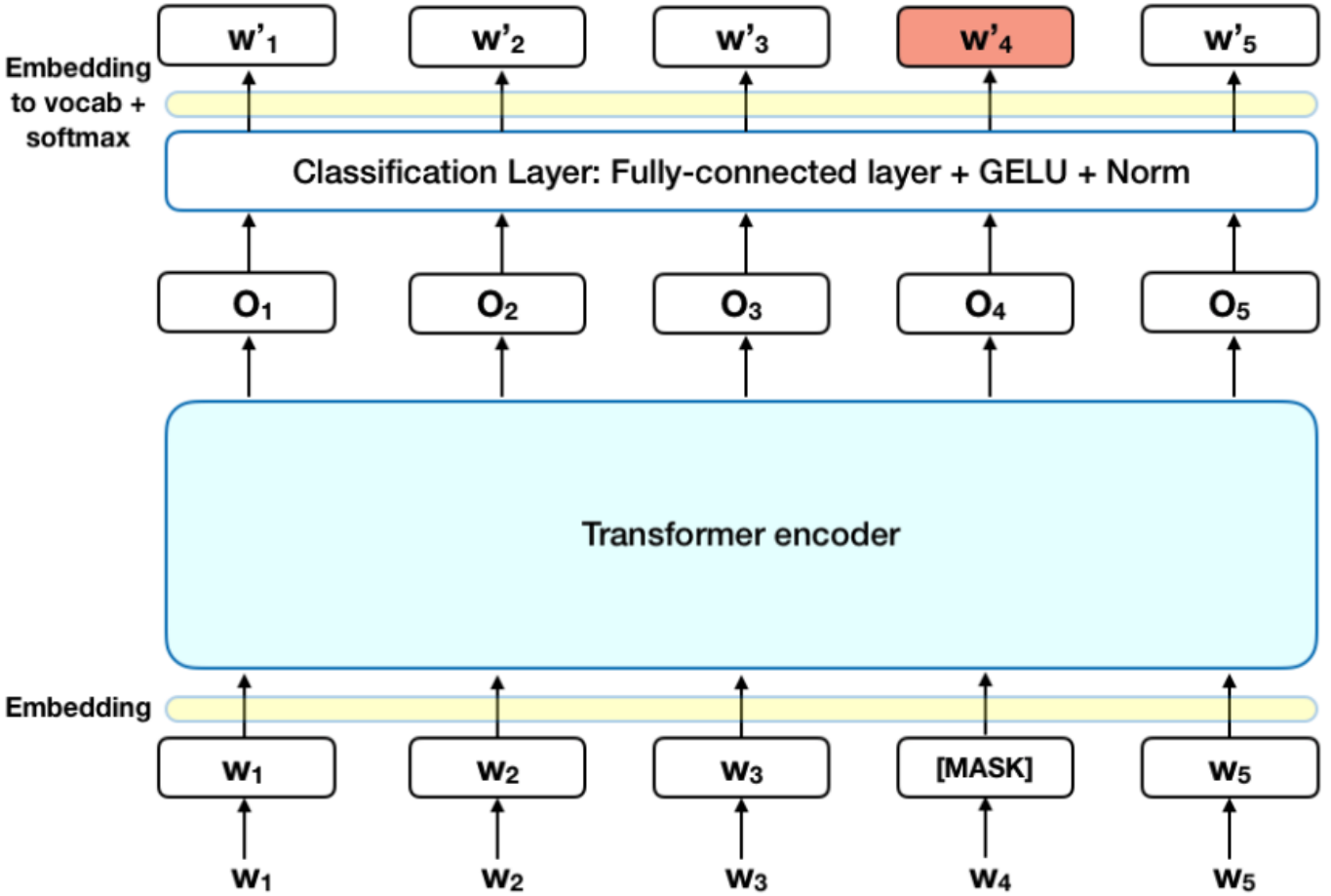
Scaling Up Language Models



Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model - Microsoft Research

BERT (Pre-Training Bidirectional Transformers)

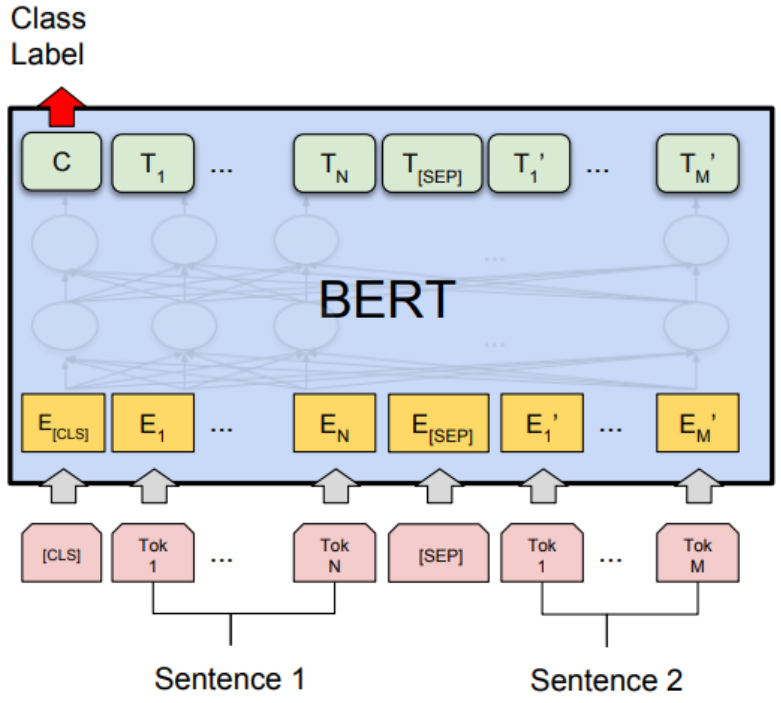
- Masked-LM



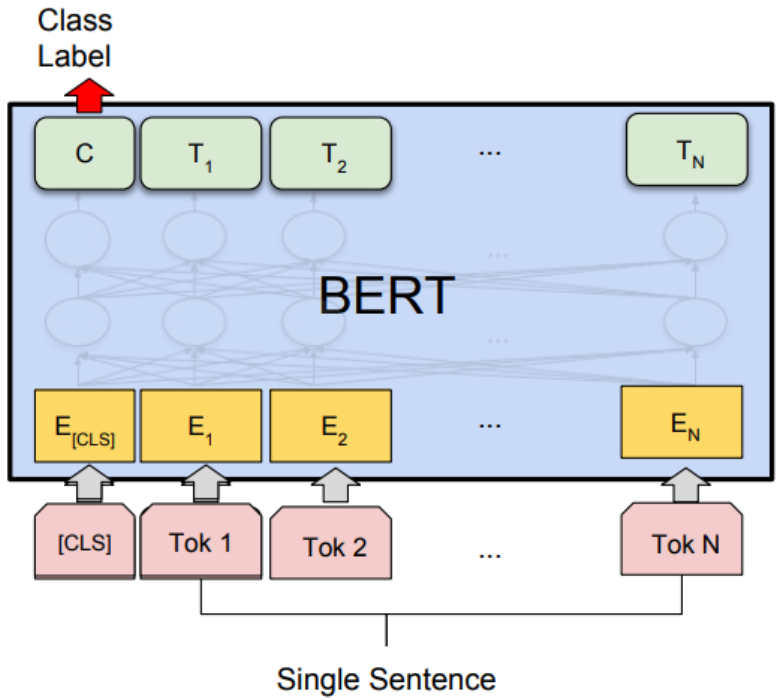
BERT Explained: State of the art language model for NLP | by Rani Horev | Towards Data Science

BERT (Pre-Training Bidirectional Transformers)

- Masked-LM pre-training, then fine-tuning



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

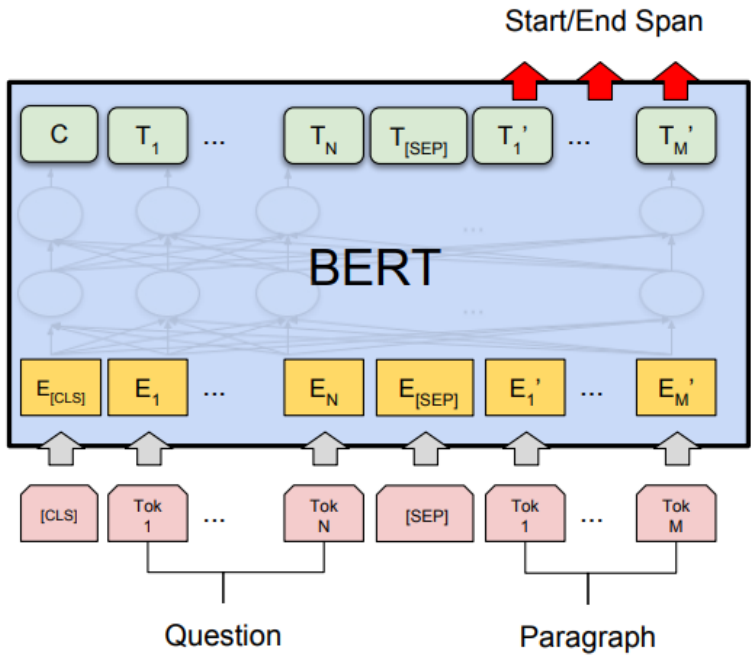


(b) Single Sentence Classification Tasks: SST-2, CoLA

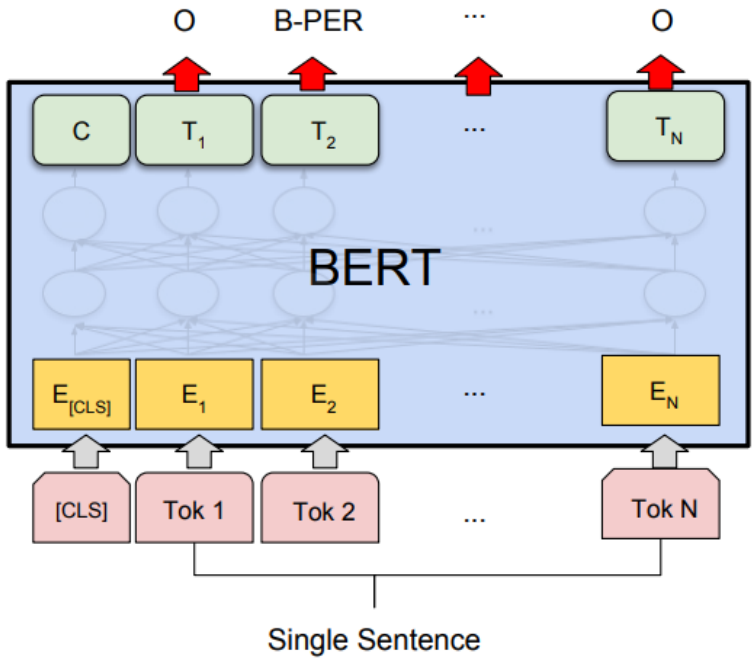
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al, NAACL 2019

BERT (Pre-Training Bidirectional Transformers)

- Masked-LM pre-training, then fine-tuning



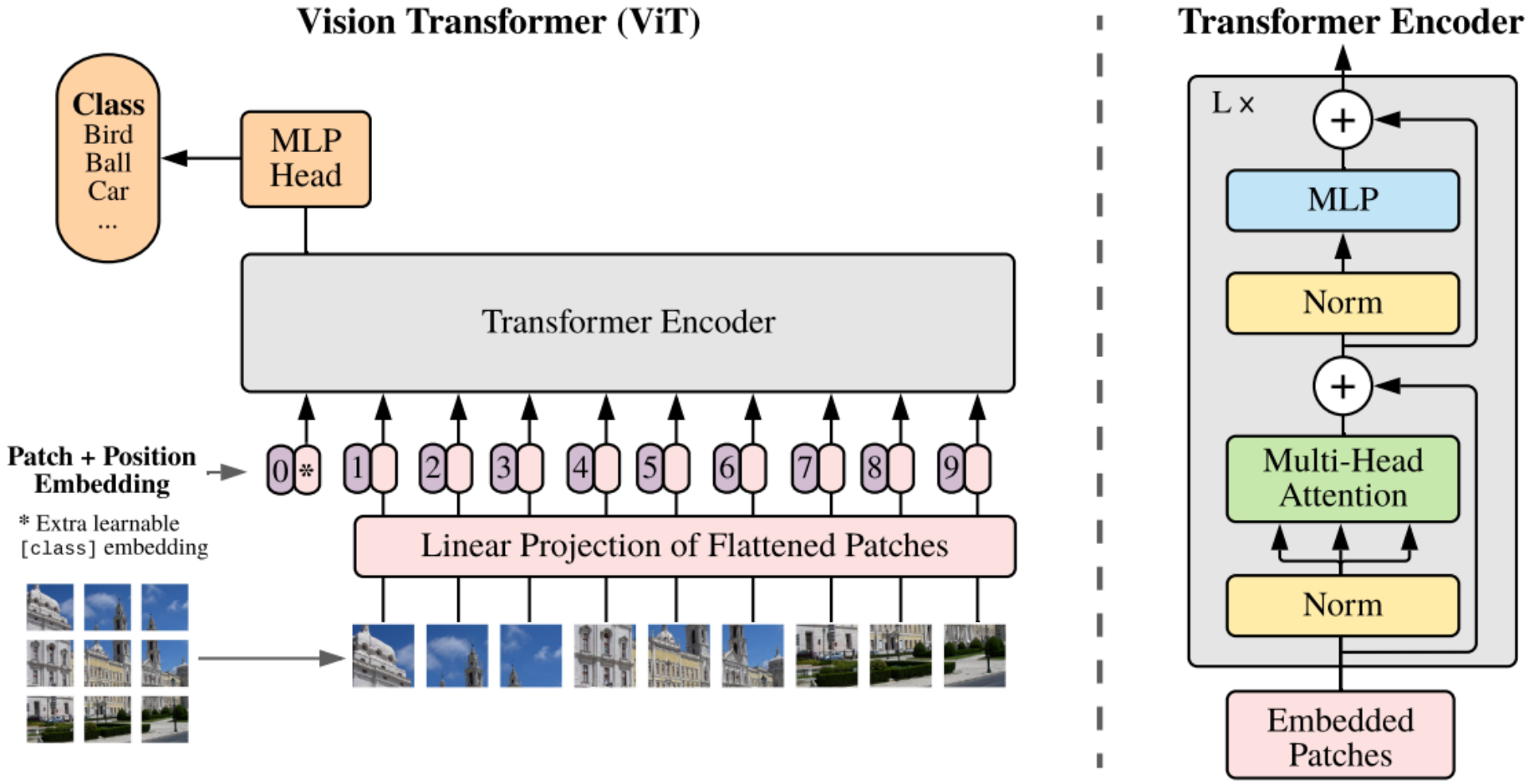
(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al, NAACL 2019

Vision Transformers (ViT)



An Image is worth 16x16 words: Transformers for image recognition at scale, Dosovitskiy et al, ICLR 2021