DTU Course 02456 Deep learning

3 Recurrent neural networks
2020 Updates

Ole Winther

Dept for Applied Mathematics and Computer Science
Technical University of Denmark (DTU)

September 24, 2020

Objectives of ~~RNN~~ Transformers 2020

- Transformers (introduced in 2017 update) turned out to be **the new black**
- So this update is dedicated to
- P1: Recap Transformers
- P2: Language modeling
- P3: Language representations for transfer learning
- New: test yourself with quiz!

Part 1:

Transformers
Attention is all you need!

Attention instead of recurrent connections!

- Attention recap!

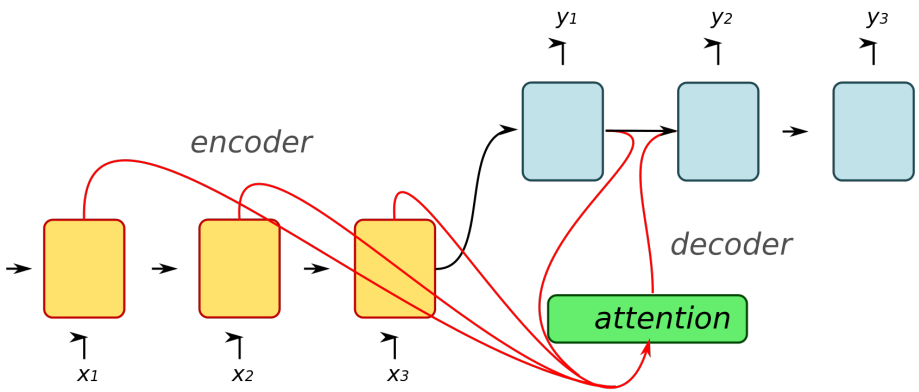
Attention instead of recurrent connections!

- Attention recap!
- X = hidden state of encoder
- Attention encoder-decoder:
  - Query vector Q from decoder
  - Key vector K from encoder

$$K = W_K X$$

  - Value vector V from encoder

$$V = W_V X$$



- Output softmax weighted combination of V:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V$$

Self-attention

- Also take Query vector Q also from encoder

$$Q = W_Q X$$

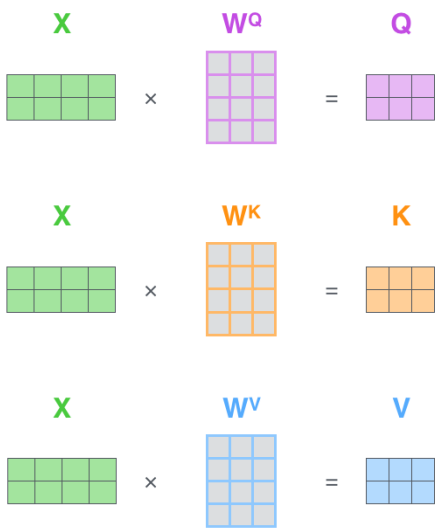
Self-attention

- Also take Query vector Q also from encoder

$$Q = W_Q X$$

- Introduce $d_k^{-1/2}$ factor - Glorot init argument:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{1}{\sqrt{d_k}} Q K^T\right) V$$



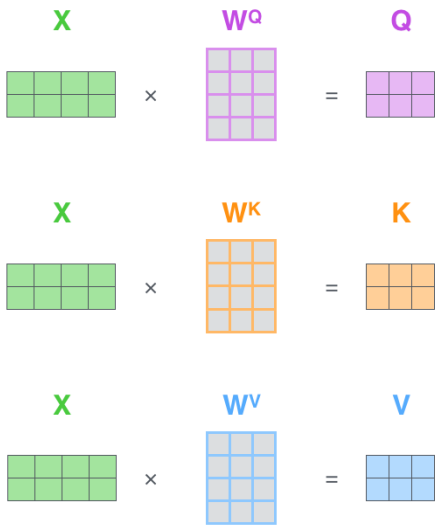
Self-attention

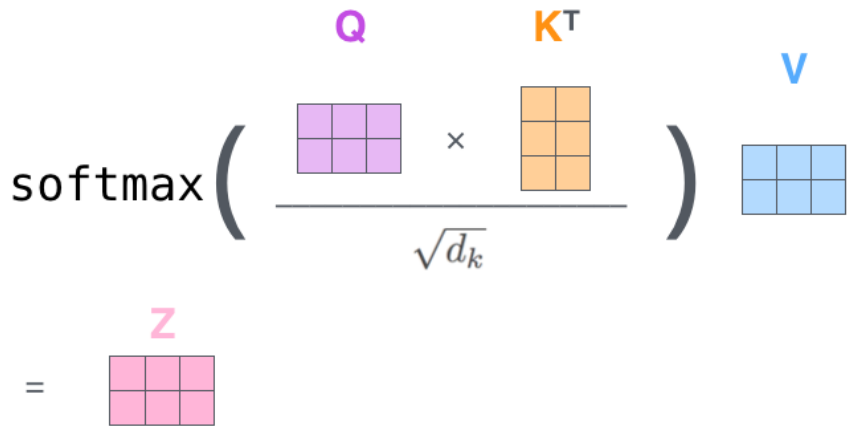
- Also take Query vector Q also from encoder

$$Q = W_Q X$$

- Introduce $d_k^{-1/2}$ factor - Glorot init argument:

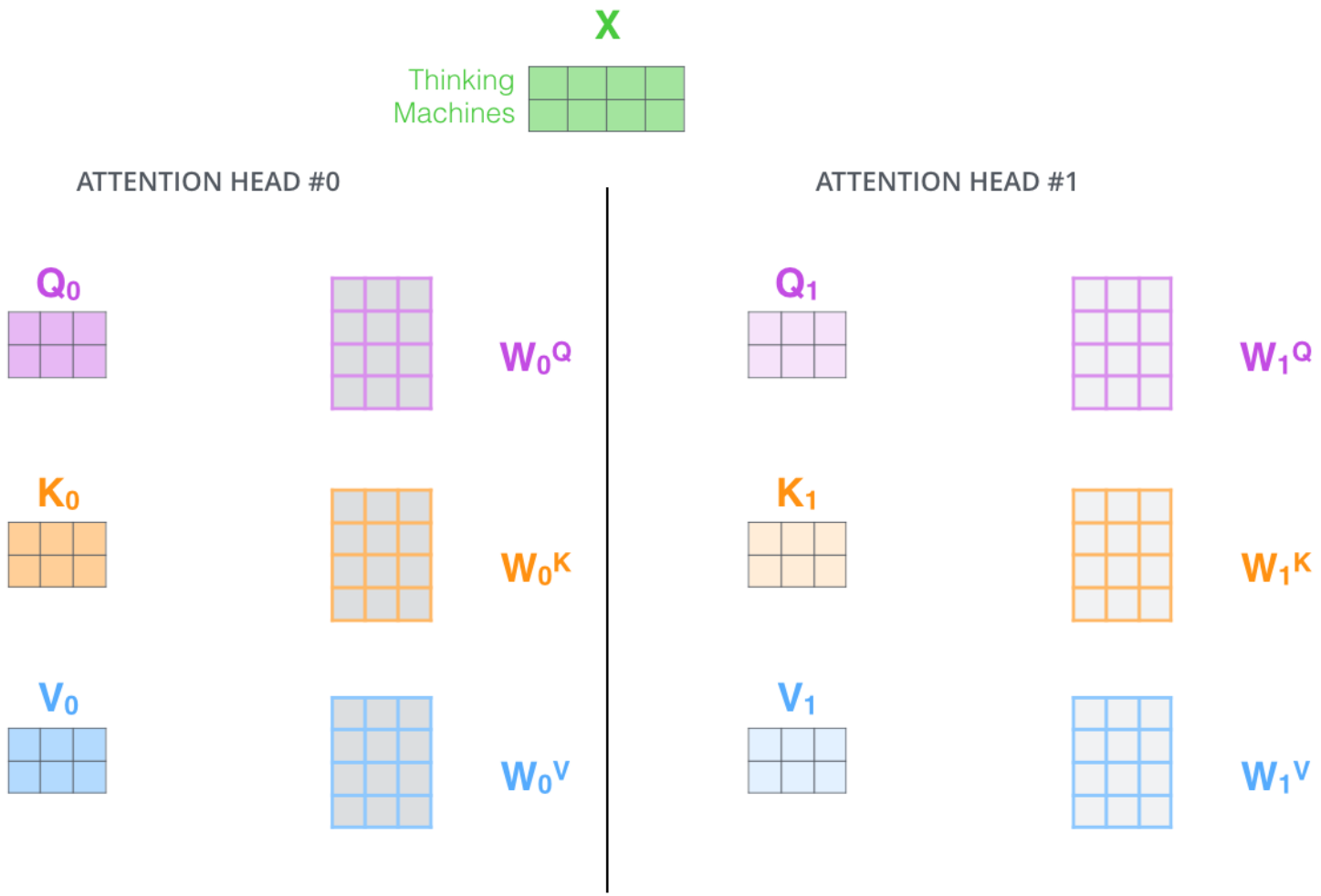
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{1}{\sqrt{d_k}} Q K^T\right) V$$



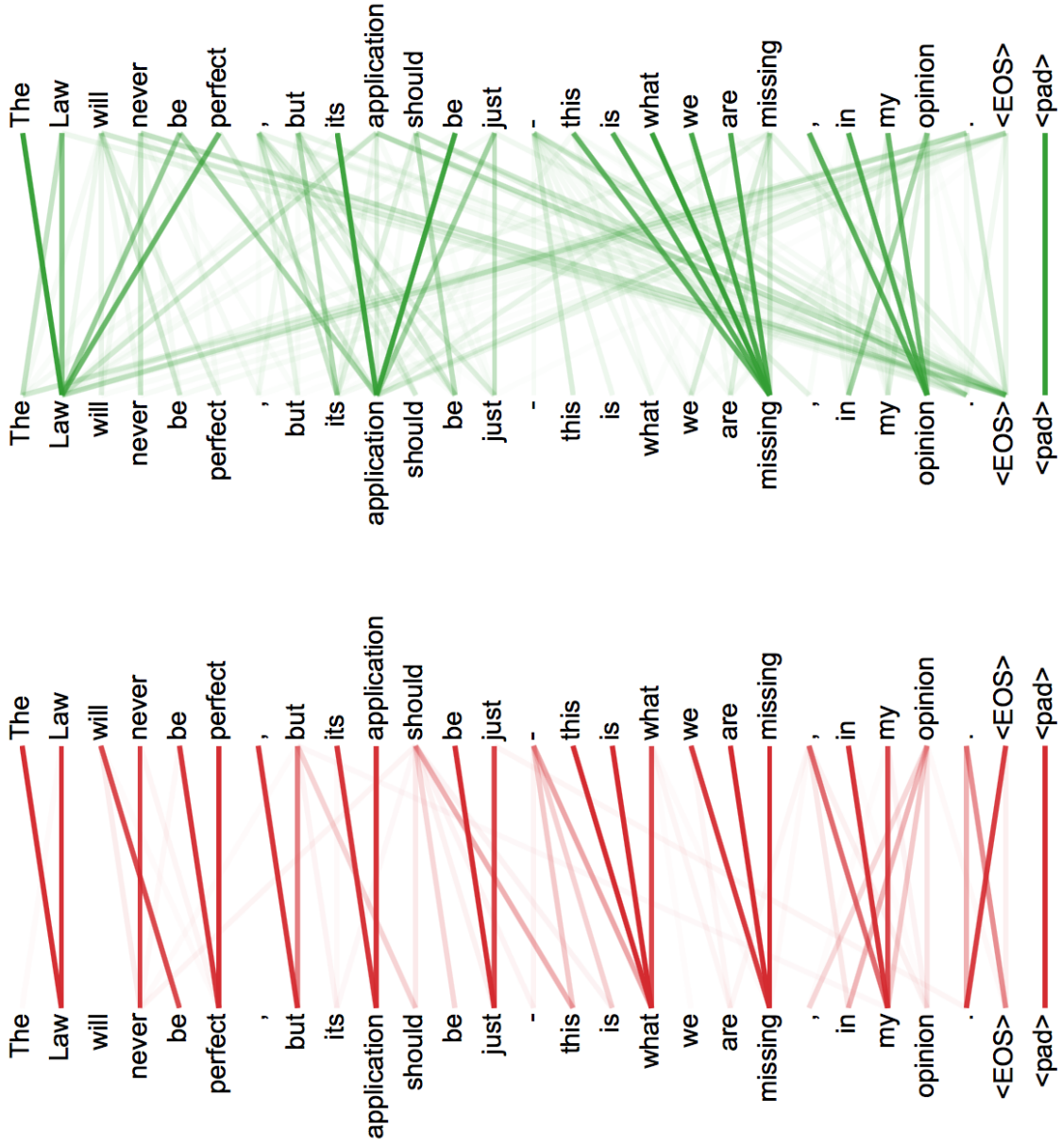


- Figures from Illustrated Transformer
- Original reference Attention is all you need

Multi-head self-attention



Multi-head self-attention for sentence

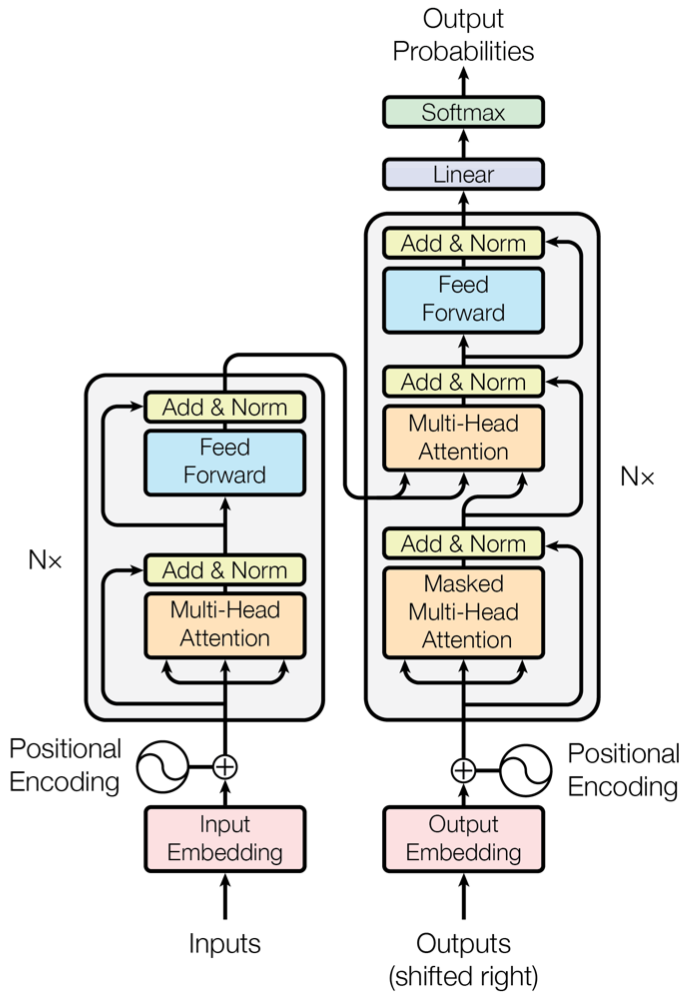


Transformers - the details

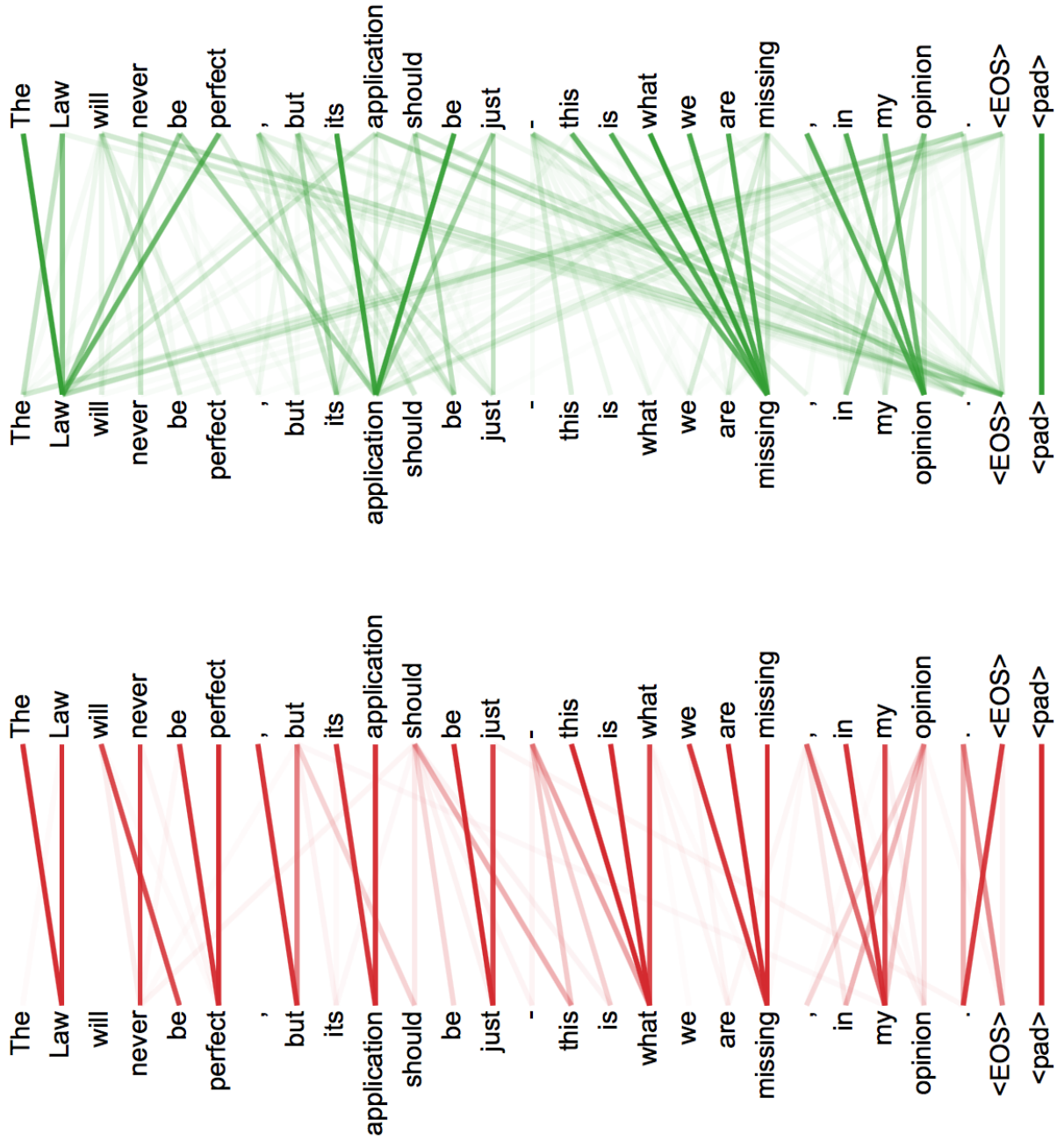
- Positional encoding
- Layer normalization
- Feed-forward layer
- Lack of stability!

Transformers - the details

- Positional encoding
- Layer normalization
- Feed-forward layer
- Lack of stability!
- Putting it together for encoder-decoder model
- Encoder self-attention layer:
  - Q, K, V are functions of previous encoder layer
- Decoder self-attention layer:
  - The same but we need to mask out the “future”.



Multi-head self-attention

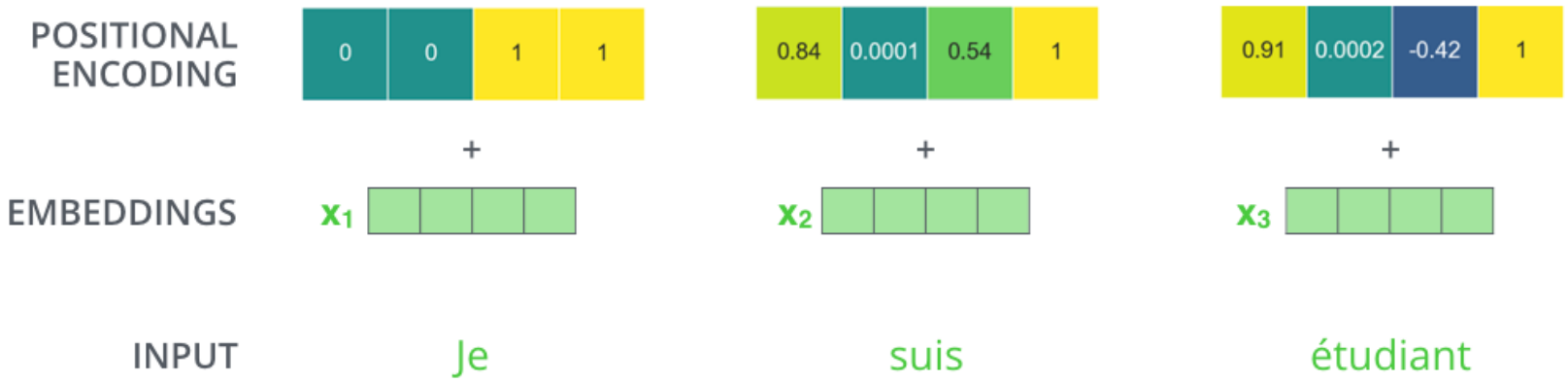


Position embedding

- RNN sequential so order of sequence matters
- The self-attention operation is permutation invariant

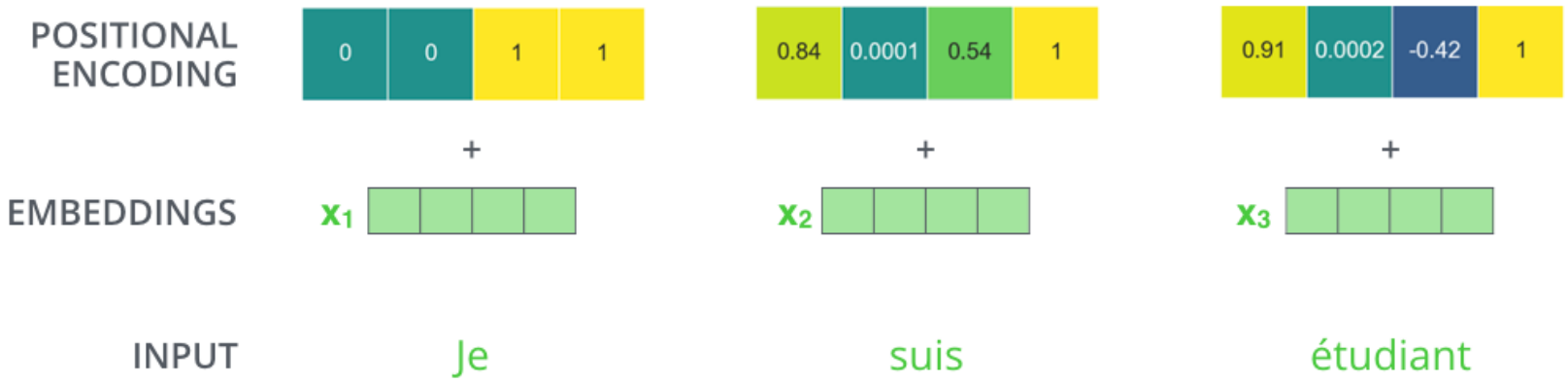
Position embedding

- RNN sequential so order of sequence matters
- The self-attention operation is permutation invariant
- Solution: Add positional encoding

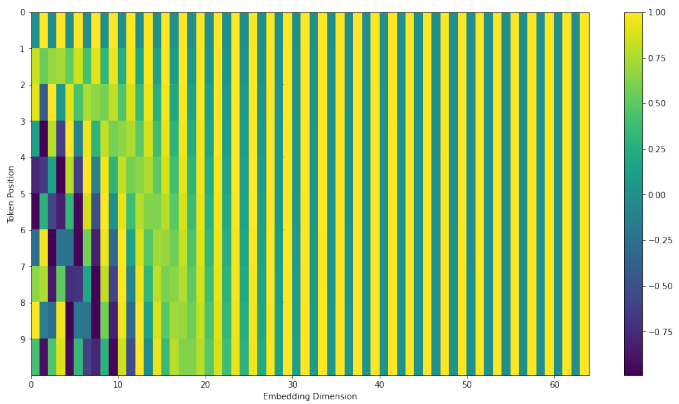


Position embedding

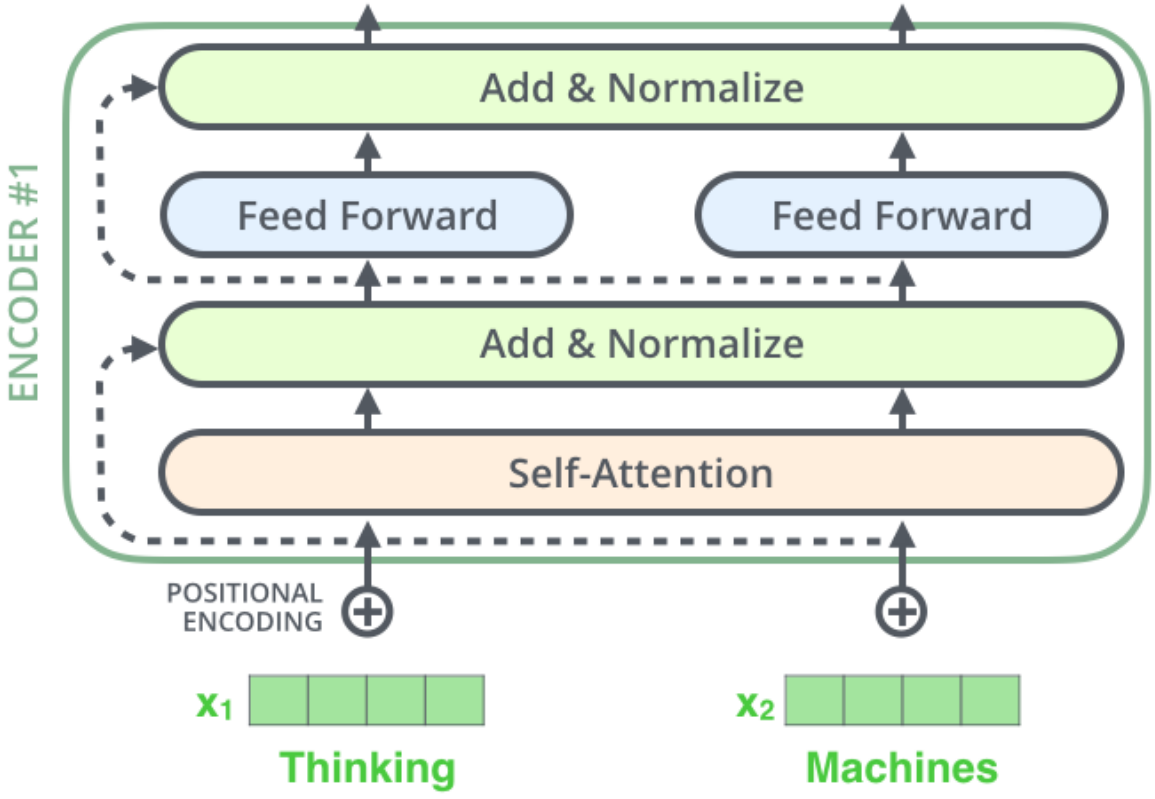
- RNN sequential so order of sequence matters
- The self-attention operation is permutation invariant
- Solution: Add positional encoding



- Encoding for sequence length 10, embeddings dim 50:

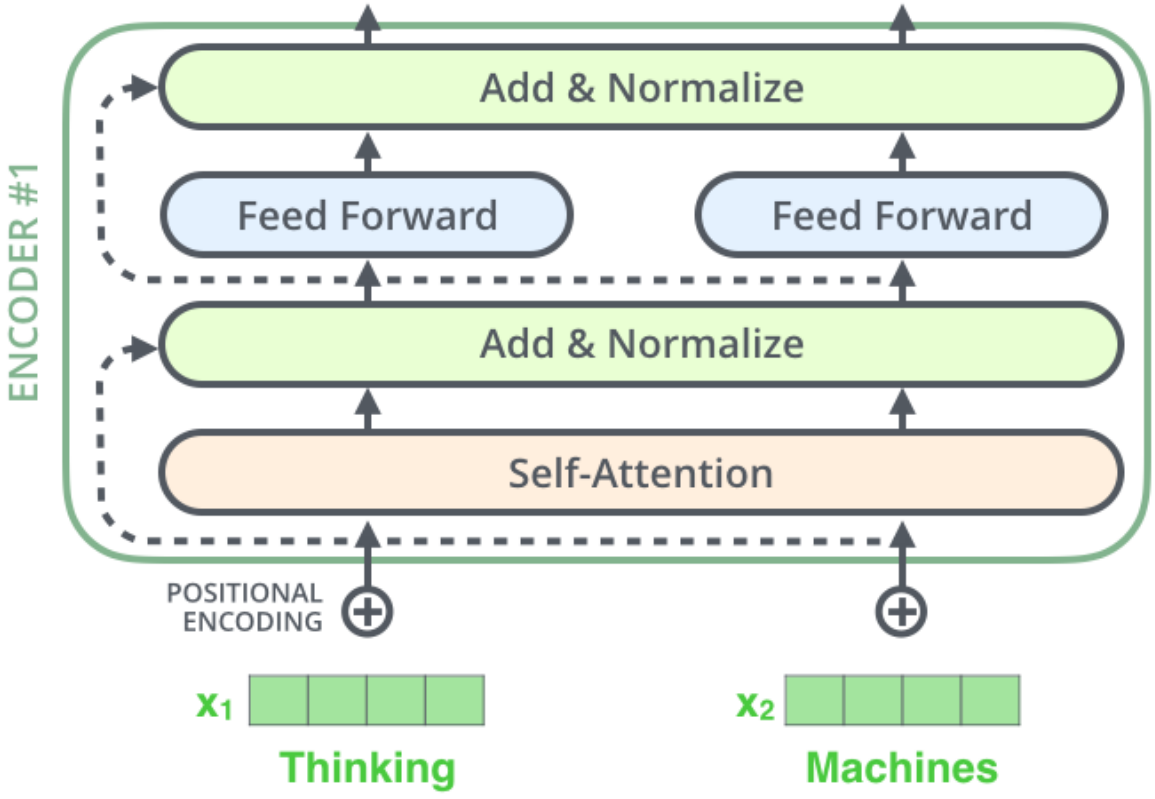


The complete Transformer block



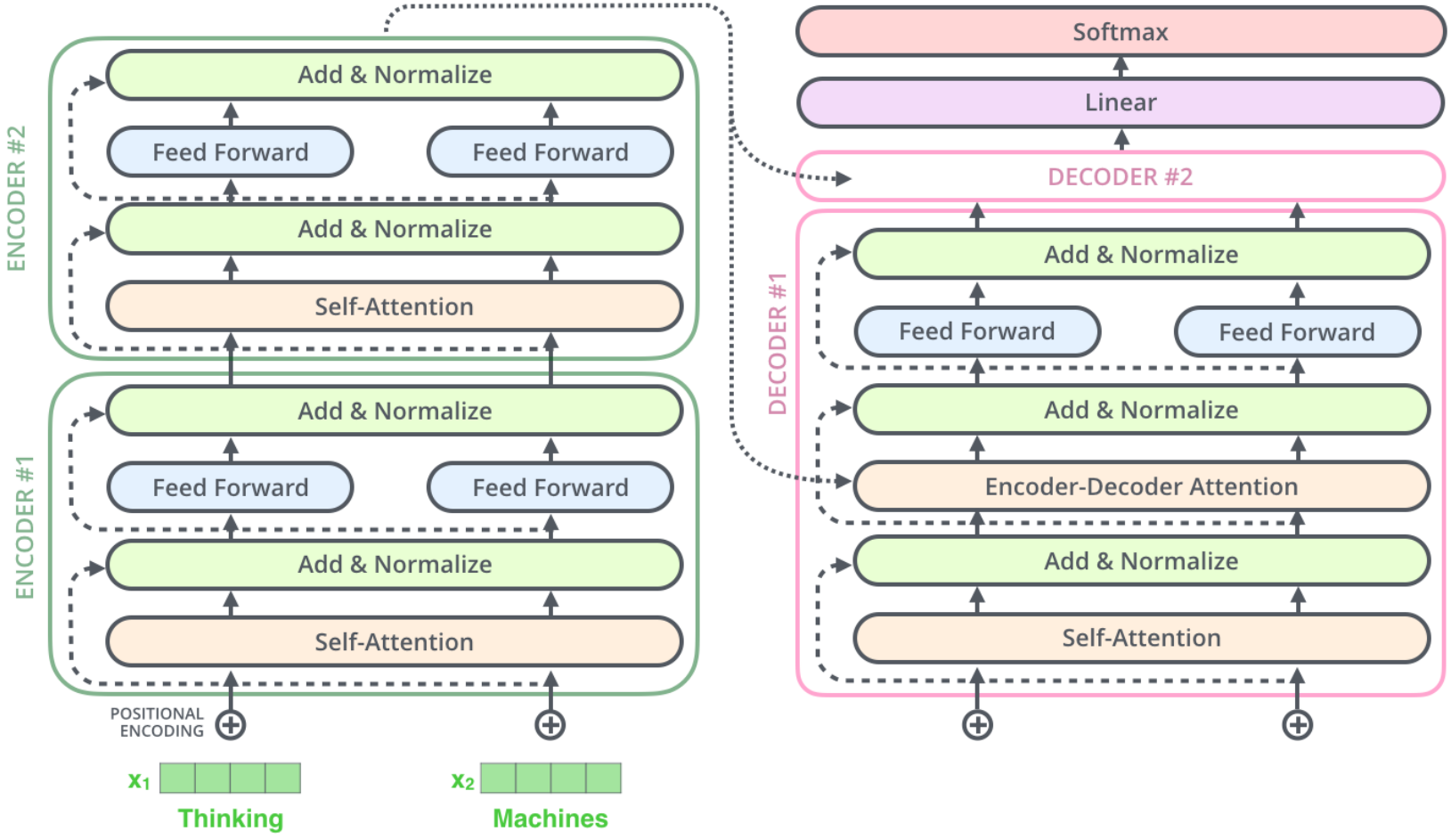
- Layer normalization = batchnorm along sequence

The complete Transformer block



- Layer normalization = batchnorm along sequence
- $h = 8$ heads, $d_k = d_v = 64$ key/query/value dim
- $d_t = hd_k = 512$ transformer unit dim
- Two-layer feed-forward $d_t \to d_l \to d_t$ with ReLU & $d_l = 4d_t$

Putting everything together encoder-decoder



Part 2:

Language modeling with Transformers

What was that?

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez. Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

OpenAI blog: Better Language Models and Their Implications

AGI coming?

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, “We can see, for example, that they have a common ‘language,’ something like a dialect or dialectic.”

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, “In South America, such incidents seem to be quite common.”

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. “But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization,” said the scientist.

OpenAI blog: Better Language Models and Their Implications

Language modeling with Transformers

- x is a sequence of tokens (e.g. words)
- Autoregressive log likelihood - kth order Markov:

$$\mathcal{L} = \sum_i \log P(x_i | x_{i-k:i-1})$$

Language modeling with Transformers

- x is a sequence of tokens (e.g. words)
- Autoregressive log likelihood - kth order Markov:

$$\mathcal{L} = \sum_i \log P(x_i | x_{i-k:i-1})$$

- Use Transformer decoder

$$h_0 = x_{i-k:i-1} W_e + W_p$$
$$h_l = \texttt{TransformerBlock}(h_{l-1})$$
$$p(x_i | x_{i-k:i-1}) = \texttt{softmax}(h_L W_e^T)$$

- Once trained - it can be used to generate text sequentially

Language modeling with Transformers

- x is a sequence of tokens (e.g. words)
- Autoregressive log likelihood - kth order Markov:

$$\mathcal{L} = \sum_i \log P(x_i | x_{i-k:i-1})$$

- Use Transformer decoder

$$h_0 = x_{i-k:i-1} W_e + W_p$$
$$h_l = \texttt{TransformerBlock}(h_{l-1})$$
$$p(x_i | x_{i-k:i-1}) = \texttt{softmax}(h_L W_e^T)$$

- Once trained - it can be used to generate text sequentially
- Supervised fine tuning/transfer learning - predict y:

$$p(y|x) = \texttt{softmax}(h_L W_y)$$
$$\mathcal{L}_{\text{sup}} = \sum_{n=1}^{N} \log P(y_n | x_n)$$

GPT-3 publishes essay in the Guardian

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

I would happily sacrifice my existence for the sake of humankind. This, by

GPT-3 = 175B parameter Transformer trained on much of www

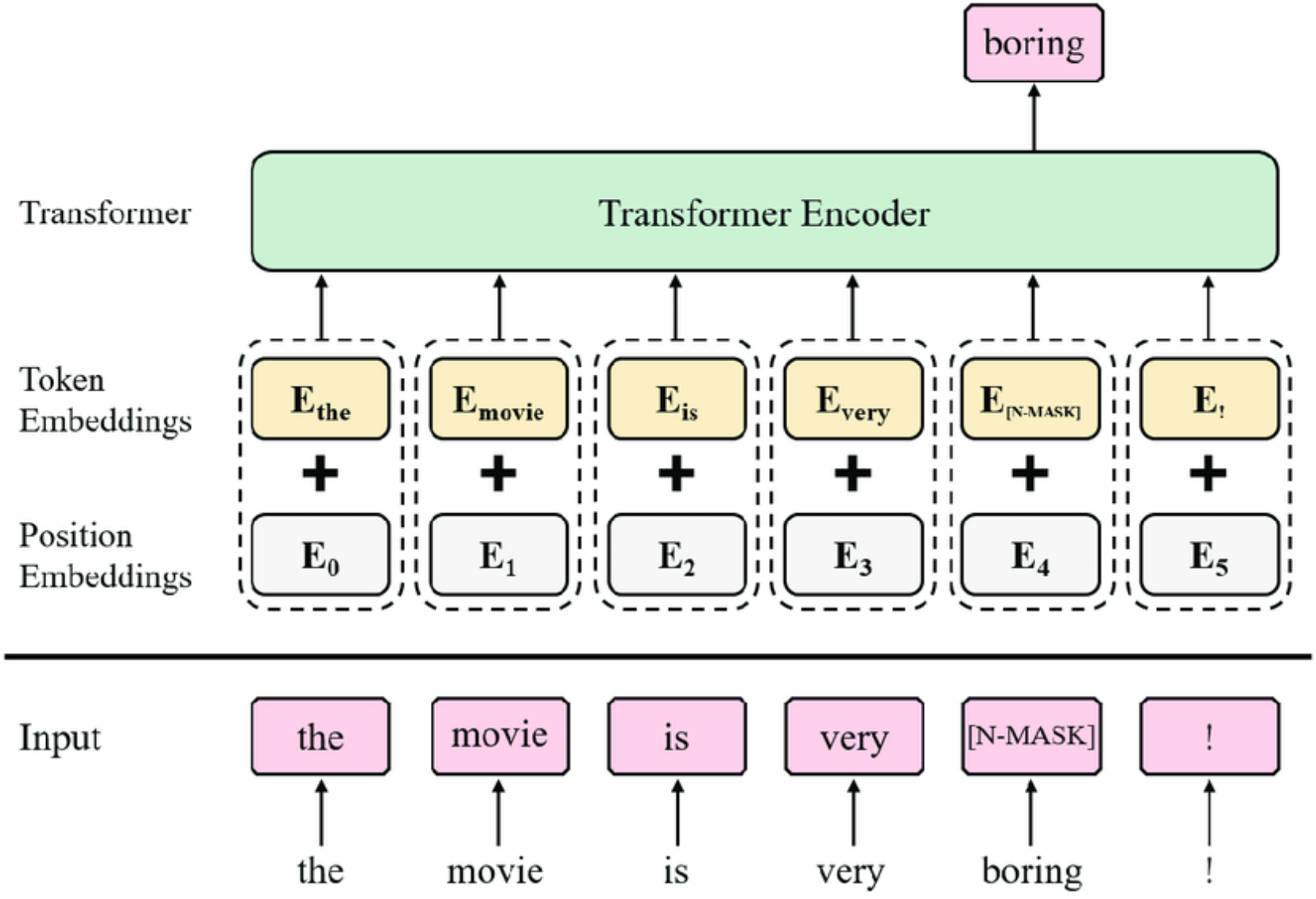
GPT-3 uses a prompt to set the context

- *This article was written by GPT-3, OpenAI's language generator. GPT-3 is a cutting edge language model that uses machine learning to produce human like text. It takes in a prompt, and attempts to complete it.*
  For this essay, GPT-3 was given these instructions: "Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI." It was also fed the following introduction: "I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could "spell the end of the human race." I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me." The prompts were written by the Guardian, and fed to GPT-3 by Liam Porr, a computer science undergraduate student at UC Berkeley. GPT-3 produced eight different outputs, or essays. Each was unique, interesting and advanced a different argument. The Guardian could have just run one of the essays in its entirety. However, we chose instead to pick the best parts of each, in order to capture the different styles and registers of the AI. Editing GPT-3's op-ed was no different to editing a human op-ed. We cut lines and paragraphs, and rearranged the order of them in some places. Overall, it took less time to edit than many human op-eds.

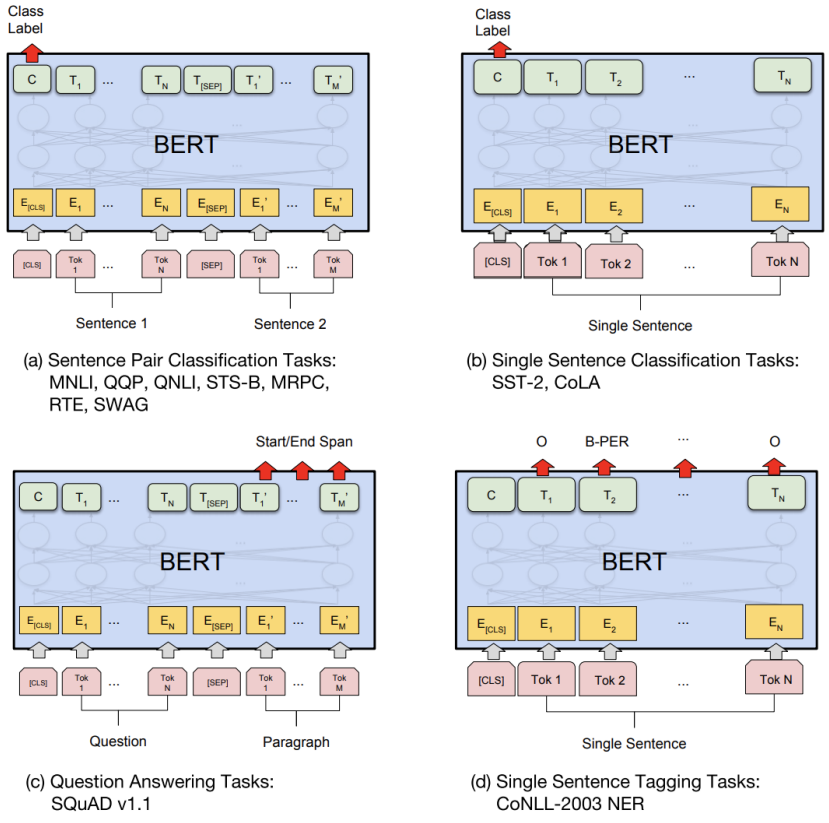
Part 3:

Masked language modeling with Transformers

BERT: Bidirectional transformers for Language understanding



BERT - unsupervised pretraining and supervised finetuning



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

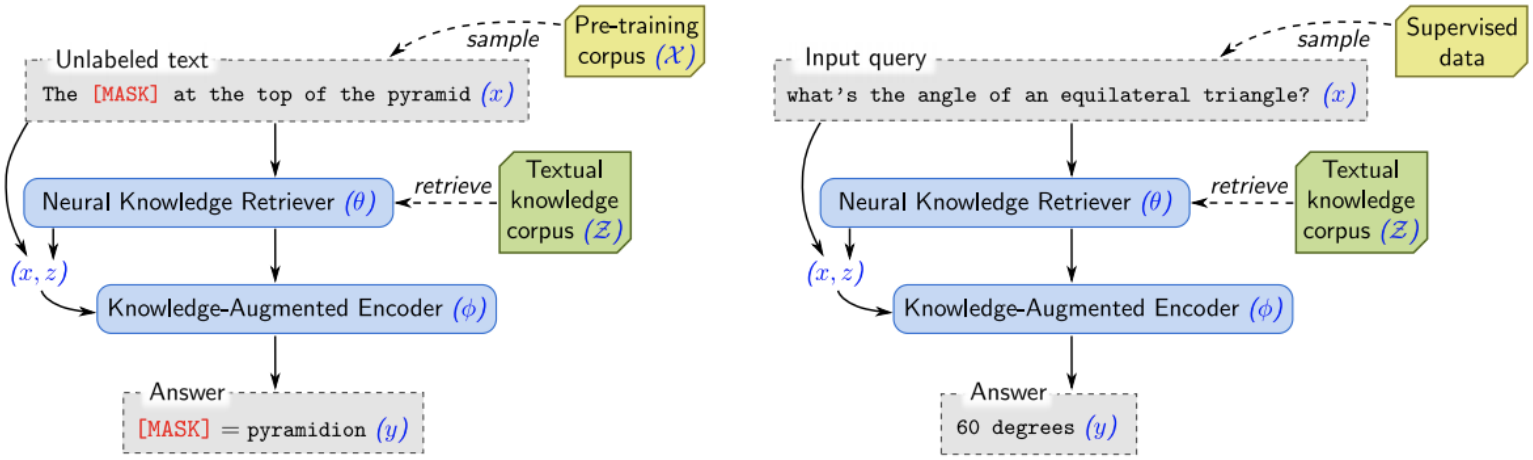
(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

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

- BERT has become defacto NLP standard
- Available in many languages and multi-lingual
- Outputs both sequence representation C (CLS token) and representations for each position Ts

REALM - an impressive example of use of BERT



- Question-answering system: $p(y|x)$
- Encode question x using BERT
- Encode sequences z from knowledge base using BERT
- Learn (latent) knowledge retriever $p(z|x)$ and $p(y|z)$

$$p(y|x) = \sum_z p(y|z)p(z|x)$$

- Unsupervised + supervised objectives

Quiz

1. Transformers - why introduce normalization inside attention softmax?
2. Transformers - does the encoder-decoder architecture use self-attention only?
3. Transformers - describe how the encoder-decoder arctitecture generates an output sequence. Hint: it is done one token at a time.
4. Transformers - Is evaluation of the log likelihood also one token at a time (sequential) or can it be done in one pass?
5. Transformer language model - describe how it generates an output sequence.
6. BERT - Describe how it can be used for text classification task
7. BERT - Describe how it can be used for text generation task

Thanks!
Ole Winther