Summer study group - Transformer

Ziqi Kang

University of Helsinki

2024.07.31

The GREAT Transformer!

- Transformer is the **first sequence transduction model based entirely on attention**, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.
- The basis of Large Language Model!

The structure of Transformer

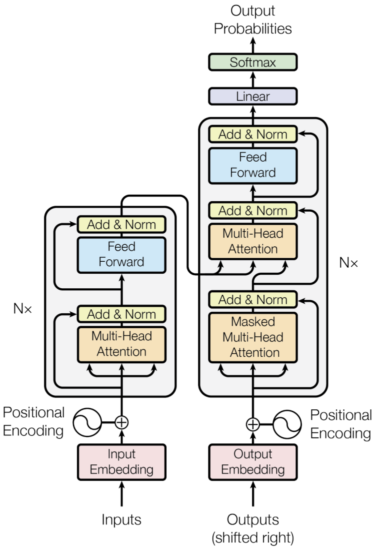
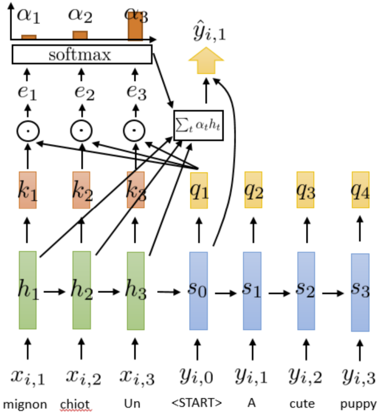


Figure 1: The Transformer - model architecture.

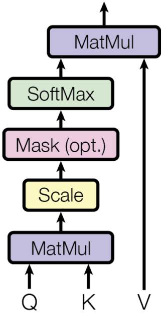
- Embedding.
  - Context embedding (word2vec, token)
  - Positional embedding
- Left: Encoder (N=6).
  - Multi-head self-attention
  - Feed forward
  - LayerNorm
  - Residual connection
- Right: Decoder (N=6).
  - Masked Multi-head attention
  - Feed forward
  - LayerNorm
  - Residual connection
- Output.
  - Linear + Softmax

What is attention?



- Get **Value** via **Query** based on **Key**.
  - $V = f(Q, K)$
  - $Attention = f(Q, K, V)$, Attention is a weight matrix.

What is **self**-attention?



- Calculate all tokens at the same time.
- Q, K, V all based on input matrix X with certain linear transformation.
- Use scaled dot product to calculate the **similarity** of *Query* and *Key*.

$$S(K, Q) = \frac{Q \cdot K^T}{\sqrt{d_k}}$$

, where d_k is the dimension of Key (Q, K, V all have the same dimension in self-attention).

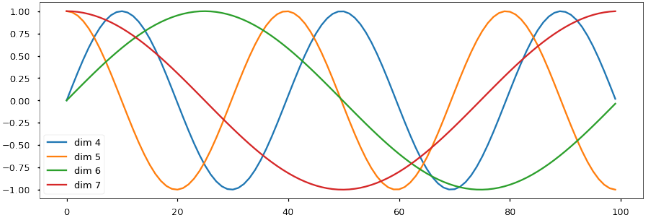
- Attention is based on similarity:

$$Attention(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}})V$$

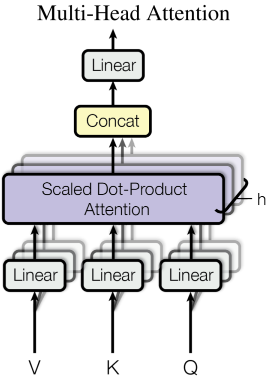
From self-attention to Transformer

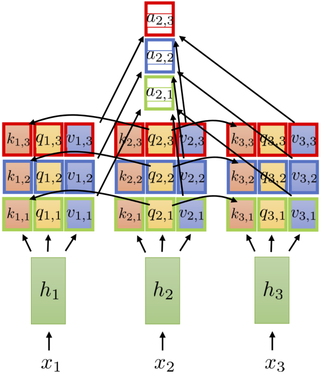
- Positional encoding
  - Lack of sequence information.
- Multi-headed attention
  - To get more information
- Adding nonlinearities
  - Self-attention is linear
- Masked decoding
  - When decoding, the output at step 1 can see the input in future steps

Positional Encoding

- Add to word embedding.
- The positional encoding shouldn't be too large to infect the word embedding itself.
- Ensure the position encoding has a certain value range, also keep the information of relative positions (in various length of texts, the words should have the same position encoding value difference if the distance is the same). – Use sine and cosine functions!
- 

Multi-headed attention




$a_{2,3}$
$a_{2,2}$
$a_{2,1}$
$k_{1,3}$ $q_{1,3}$ $v_{1,3}$
$k_{2,3}$ $q_{2,3}$ $v_{2,3}$
$k_{3,3}$ $q_{3,3}$ $v_{3,3}$
$k_{1,2}$ $q_{1,2}$ $v_{1,2}$
$k_{2,2}$ $q_{2,2}$ $v_{2,2}$
$k_{3,2}$ $q_{3,2}$ $v_{3,2}$
$k_{1,1}$ $q_{1,1}$ $v_{1,1}$
$k_{2,1}$ $q_{2,1}$ $v_{2,1}$
$k_{3,1}$ $q_{3,1}$ $v_{3,1}$
h_1
h_2
h_3
x_1
x_2
x_3

- Repeat the operation (WQ, WK, WV) several times for better optimization.
- Usually 8-head is quite enough.

Adding nonlinearities

- Because real world is non-linear!
- Feed-forward layer.
- The activation function is non-linear (e.g. ReLu).

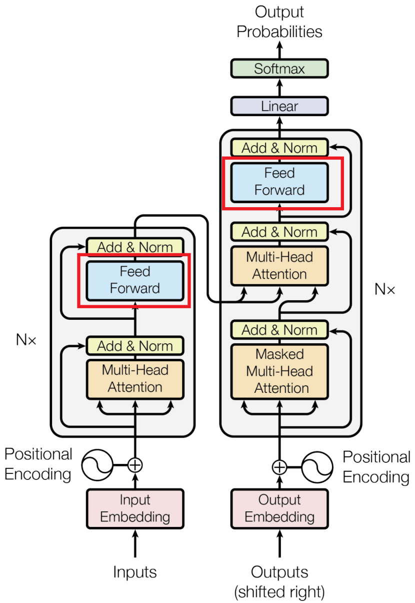
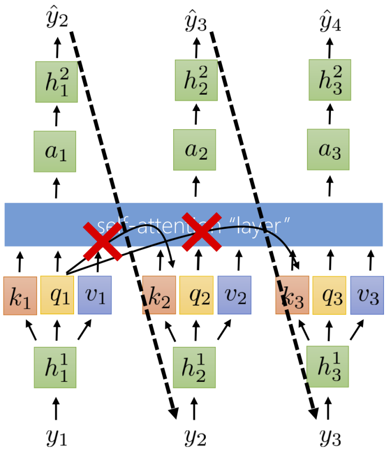


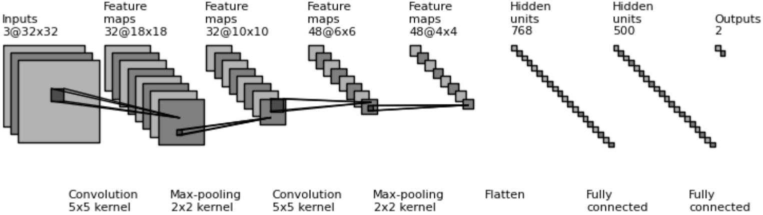
Figure 1: The Transformer - model architecture.

Masked decoding



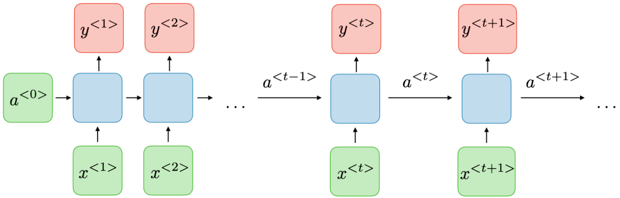
- When decoding, make sure that the model can only see the first n-1 words in the output sequence when the nth word is generated.
- Masked through Attention matrix. (Set value as -inf)

Compare with convolutional neural networks (CNNs) and recurrent neural networks (RNNs)

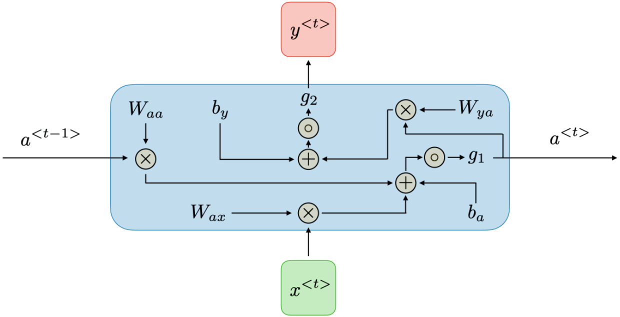


- Process with image data
- Can't process sequence data (the previous step infect the next step)

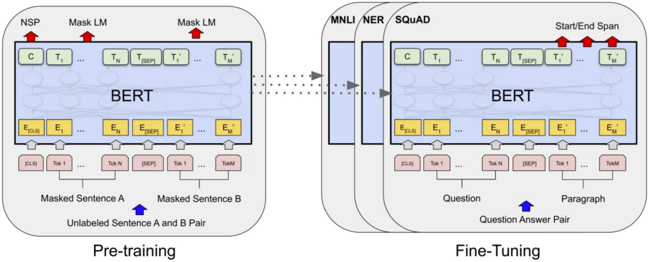
Compare with convolutional neural networks (CNNs) and recurrent neural networks (RNNs)



One RNN block:

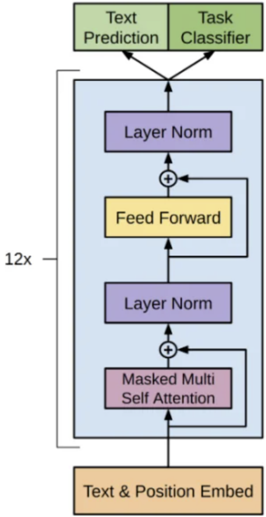


Variants of Transformer - BERT



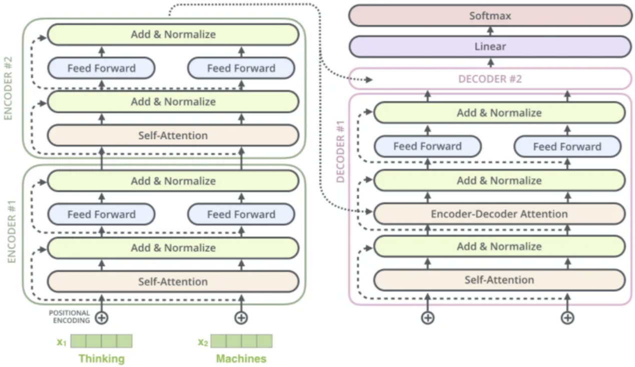
- Encoder-only architecture
- Adapts the final hidden states of the pre-trained model for downstream tasks with additional output layers as needed
- Requires task-specific head layers and fine-tuning for each downstream task

Variants of Transformer - GPT



- Decoder-only architecture.
- Adds a linear layer on top of the transformer and fine-tunes on the downstream task using the same causal language modeling objective.
- Is generative in nature and can be prompted to perform tasks with minimal changes to its structure.

Variants of Transformer - T5



- Text-to-Text Transfer Transformer, Treats every task as a "text-to-text" problem.
- Encoder-decoder architecture.
- Implements a variant of the transformer that uses relative position biases instead of positional embedding.

The potential harms of Language Models

- Language models may say false things, a problem called hallucination. (No way to enforce that the text that is generated is correct or true
- Language models can generate toxic language.
- Language models also present privacy issues since they can leak information about their training data.
- The issue of copyright.

https://web.stanford.edu/j̃urafsky/slp3/10.pdf

Thanks for listening!

Reference

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 30.

Mittal, A. (2023) NLP rise with Transformer models: A comprehensive analysis of T5, Bert, and GPT, Unite.AI. Available at: https://www.unite.ai/nlp-rise-with-transformer-models-a-comprehensive-analysis-of-t5-bert-and-gpt/ (Accessed: 29 July 2024).

Introduction to Machine Learning (CS 189/289A), University of California, Berkeley

Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright © 2023. All rights reserved. Draft of February 3, 2024.