Attention Is All You Need

Poster made by: Shui Yidi

Ashish Vaswani[1], Noam Shazeer[1], Niki Parmar[2], Jakob Uszkoreit[2], Llion Jones[2], Aidan N. Gomez[3], Łukasz Kaiser[1], Illia Polosukhin[4]

[1]Google Brain , [2]Google Research, [3]University of Toronto, [4]Independent Researcher

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Background

- RNNs and GRUs have led sequence-based tasks but face parallelization and long-range dependency challenges. CNNs partially address these but fall short.
- The Transformer, revolutionizes this by using attention mechanisms to enhance focus on input sequences and boost parallelization, forgoing traditional sequential processing.

GPT-3

Total weights: 175,181,291,520

Embedding	12,288 * 50,257 = d_embed * n_vocab	= 617,558,016
Key	128 * 12,288 * 96 * 96 = d_query * d_embed * n_heads * n_layers	= 14,495,514,624
Query	128 * 12,288 * 96 * 96 = d_query * d_embed * n_heads * n_layers	= 14,495,514,624
Value	128 * 12,288 * 96 * 96 = d_value * d_embed * n_heads * n_layers	= 14,495,514,624
Output	12,288 * 128 * 96 * 96 = d_embed * d_value * n_heads * n_layers	= 14,495,514,624
Up-projection	49,152 * 12,288 * 96 = n_neurons * d_embed * n_layers	= 57,982,058,496
Down-projection	12,288 * 49,152 * 96 = d_embed * n_neurons * n_layers	= 57,982,058,496
Unembedding	50,257 * 12,288 = n_vocab * d_embed	= 617,558,016

Figure: Understanding the GPT-3 Architecture and Weight Distribution

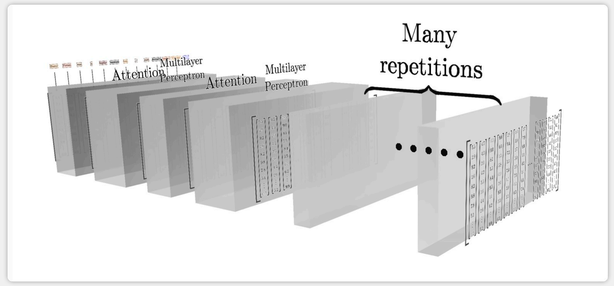

Figure: Model architecture of GPT-3

GPT-3's transformer architecture, rich with 175 billion parameters, harnesses parallel processing and attention mechanisms for advanced pattern recognition and language understanding. This structure facilitates coherent, context-aware text generation, surpassing prior models in complex language tasks.

The Transformer

Model Architecture

This figure provides a schematic representation of the Transformer model, illustrating its innovative encoder-decoder structure. The key components are as follows:

- **Encoder**: The Transformer model consists of N stacked layers, with each hosting a multi-head self-attention mechanism and a feed-forward network. Residual connections and layer normalization are applied around these sub-layers for efficiency.

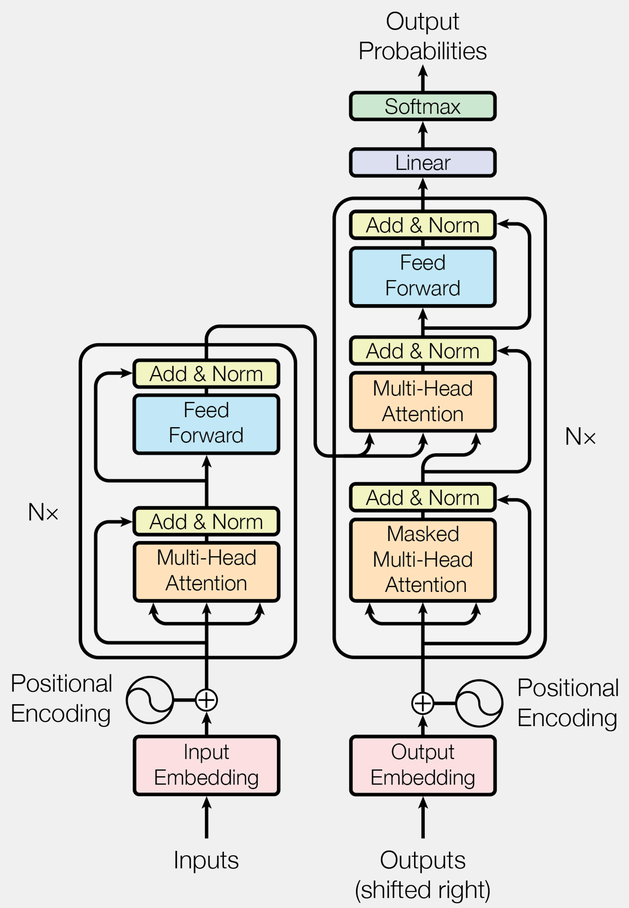

Figure: The Transformer - model architecture

- **Decoder**: Also consists of N identical layers, with an additional third sub-layer in each that performs multi-head attention over the encoder's output. Similar to the encoder, the decoder employs residual connections and layer normalization.

Attention Mechanism

This figure is divided into two main parts, illustrating the core mechanisms that enable the Transformer's powerful performance:

- **Scaled Dot-Product Attention (Left)**: Demonstrates the mechanism where the attention score is computed by scaling the dot product of queries and keys. This scaling factor, the inverse square root of the key dimension, helps in stabilizing the gradients. The attention scores determine how much each value is expressed at a position, facilitating focused processing of the input.
- **Multi-Head Attention (Right)**: Showcases how the Transformer extends the single attention process to multiple heads, allowing the model to jointly attend to information from different representation

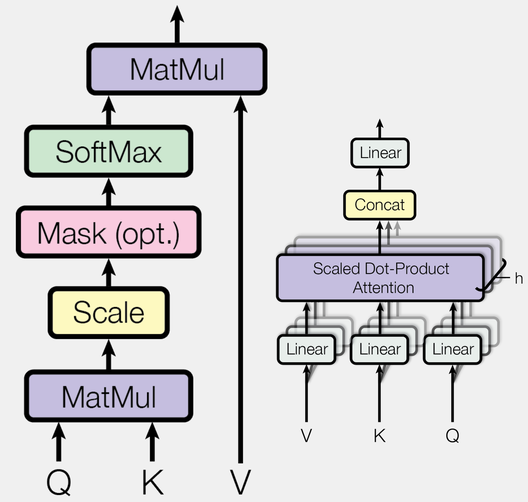

Figure: Attention Mechanisms in the Transformer

subspaces at different positions. By projecting the queries, keys, and values multiple times with different learned linear projections, enhances the model's ability to capture varied dependencies.

Intelligence VS Human

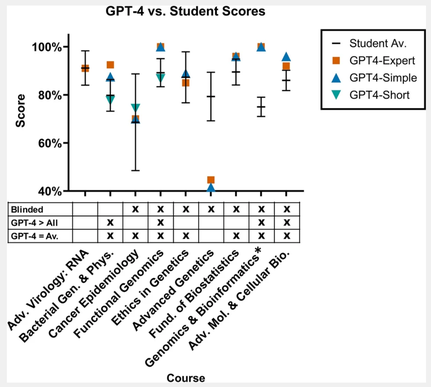

Figure: GPT Performance On 9 Graduate Examinations.

Table Legend—Blinded: All GPT-4 exam grading was performed blinded in parallel with student assessments; GPT-4>All: One or more GPT-4 scores exceeded all student scores; GPT-4≥Av.: One or more GPT-4 scores exceeded the average student score.

Acknowledgements

We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

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

https://arxiv.org/abs/1706.03762

avaswani@google.com, noam@google.com, nikip@google.com, usz@google.com, llion@google.com, aidan@cs.toronto.edu, lukaszkaiser@google.com, illia.polosukhin@gmail.com