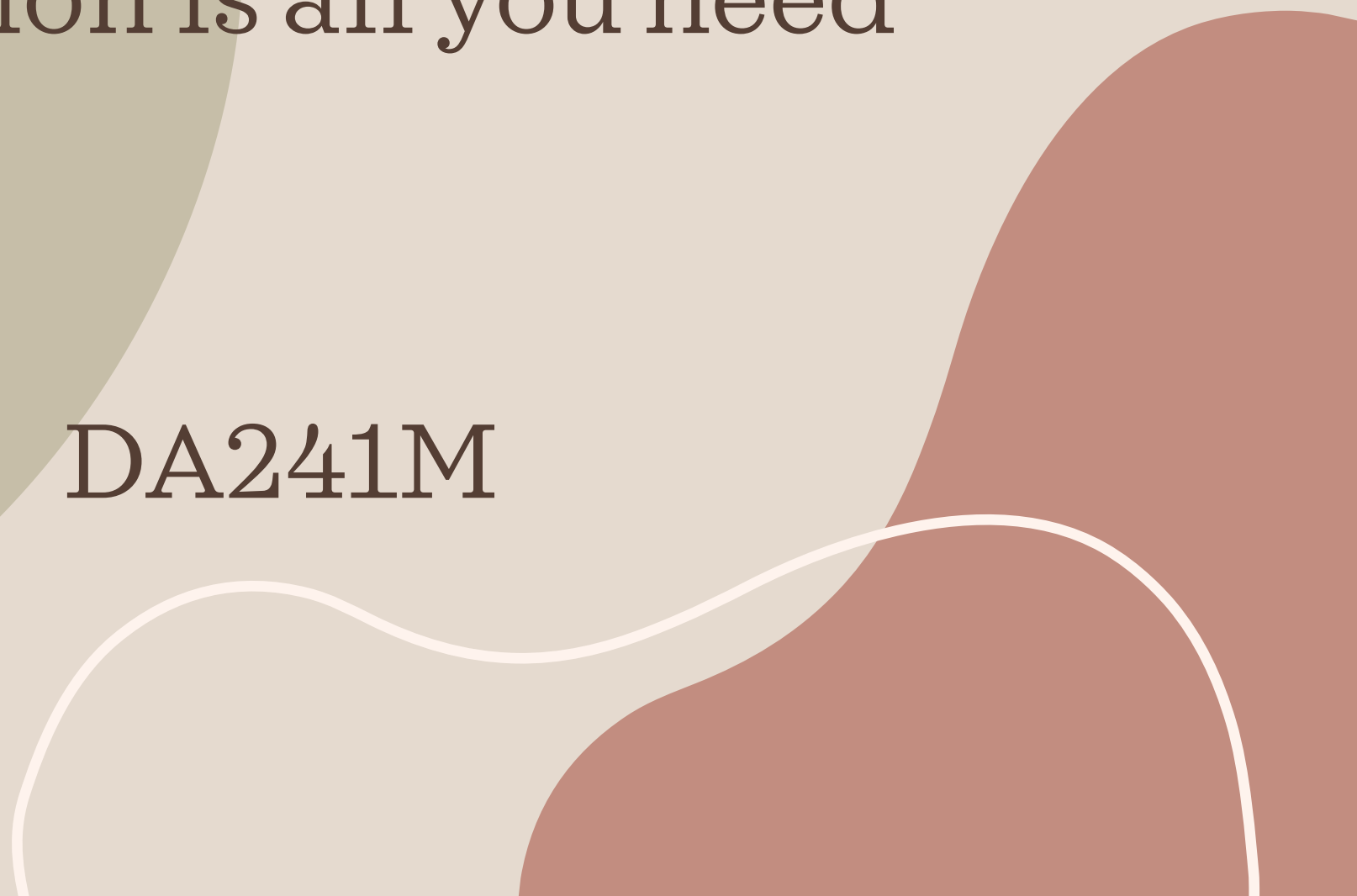




Attention is all you need

DA241M



A large, stylized leaf graphic in a muted blue-grey color, positioned in the upper left corner of the page. It has several elongated, pointed leaves branching out from a central stem.

Index

INTRODUCTION
3

METHODOLOGY
4

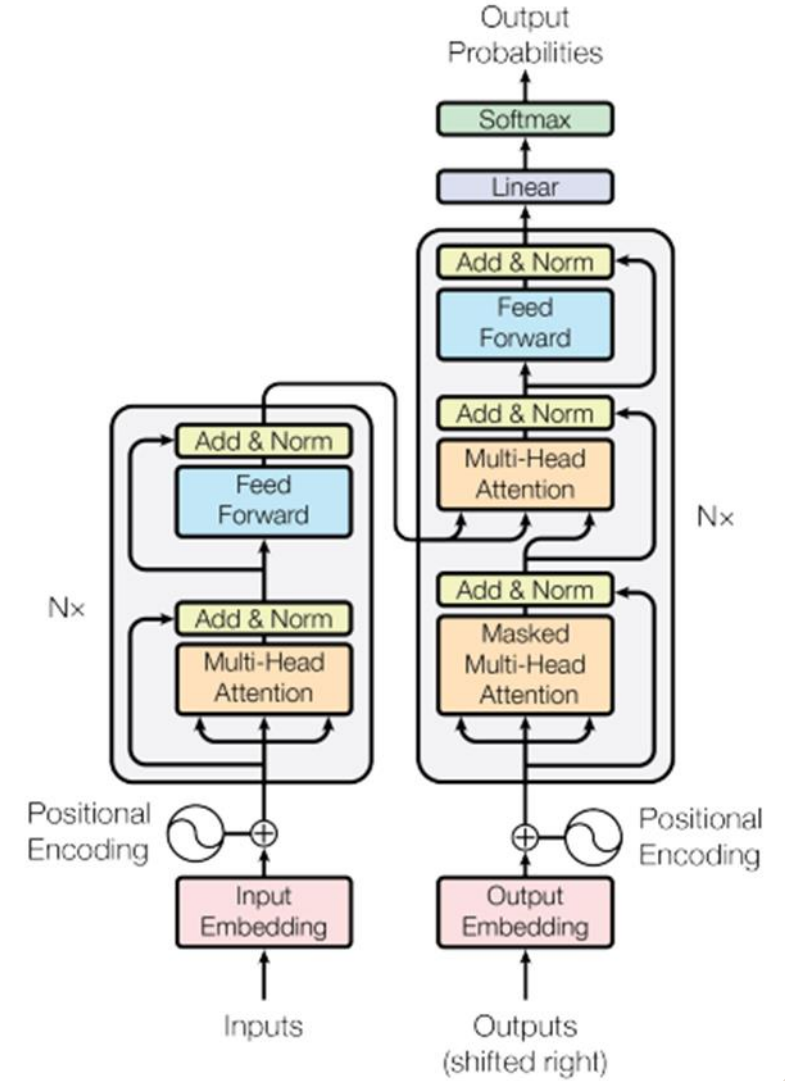
RESULTS AND DISCUSSION
6

FUTURE DISCUSSIONS
7

REFERENCES
8

INTRODUCTION

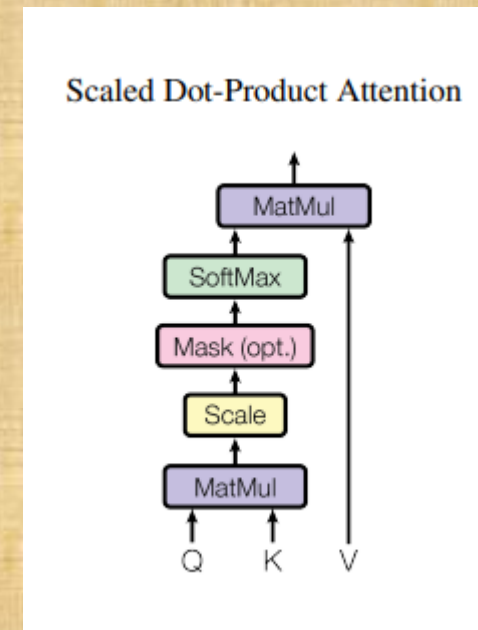
- The Transformer model, introduced in the landmark paper “Attention is All You Need” by Vaswani et al., fundamentally transformed the landscape of sequence transduction tasks, including machine translation and text summarization
- This innovation stems from replacing traditional recurrent architectures with an attention mechanism, addressing key limitations of RNNs, such as sequential dependency and slow training time.
- The self-attention mechanism allows for better parallelization, significantly improving both accuracy and computational efficiency



METHODOLOGY

- The Transformer architecture fundamentally relies on the self-attention mechanism, which allows each token in the input sequence to generate three distinct vectors: the query (Q), key (K), and value (V) vectors. This process is crucial for calculating the attention score, which determines how much focus a token should give to other tokens in the sequence
- The self-attention mechanism addresses significant limitations of traditional recurrent neural networks (RNNs), such as their sequential dependency and slow training times. By employing self-attention, the Transformer can capture long-range dependencies more effectively, which is essential for tasks like machine translation and text summarization

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

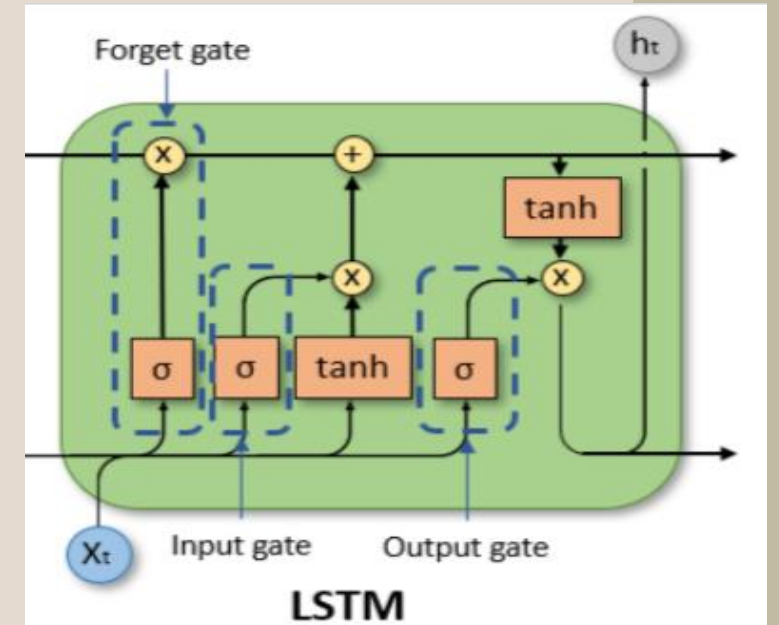


METHODOLOGY

- In addition to self-attention, the Transformer utilizes multi-head attention, which allows the model to apply multiple attention mechanisms in parallel. This is achieved by learning separate sets of Q, K, and V matrices for each attention head, enabling the model to focus on different parts of the input sequence simultaneously. The outputs from all heads are then concatenated and linearly transformed to produce the final output, enhancing the model's ability to capture diverse contextual information
- Furthermore, the Transformer incorporates positional encodings to account for the order of tokens in the sequence, as it lacks an inherent understanding of sequence order. The positional encoding is calculated using sine and cosine functions, allowing the model to maintain the sequential relationships between tokens. This combination of self-attention, multi-head attention, and positional encoding forms the backbone of the Transformer's architecture, setting it apart from previous models

RESULTS AND DISCUSSIONS

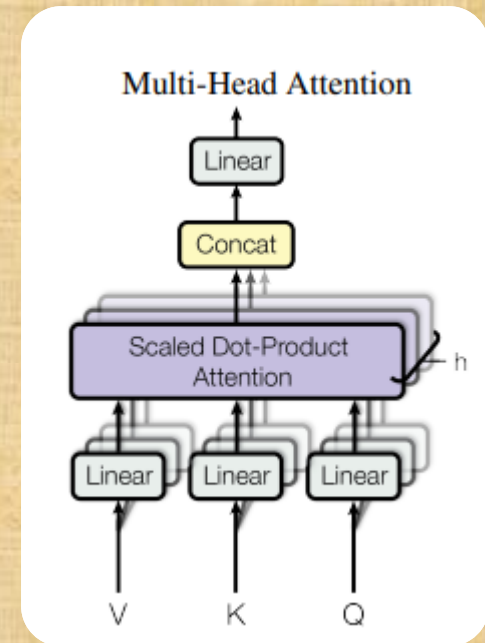
- The Transformer model was evaluated on multiple translation tasks, achieving state-of-the-art results on the WMT 2014 English-to-German and English-to-French datasets, with a BLEU score of 28.4 for English-to-German.
- This performance surpasses traditional RNN-based models like LSTMs and GRUs, demonstrating the model's efficiency in training and its ability to handle long-range dependencies
- The results indicate that the Transformer's architecture sets a new standard for sequence-to-sequence learning, making it suitable for various applications beyond translation



CONCLUSION AND FUTURE DISCUSSIONS



- The introduction of the Transformer model marked a turning point in NLP and deep learning research, paving the way for models like BERT and GPT
- Future research could focus on optimizing the Transformer for complex tasks involving hierarchical and multi-modal data
- Additionally, improving the efficiency of self-attention mechanisms remains an active area of research, suggesting that the architecture will continue to impact many fields of AI



REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). "Attention is all you need." Advances in Neural Information Processing Systems, 30, 5998–6008
2. Bahdanau, D., Cho, K., & Bengio, Y. (2015). "Neural machine translation by jointly learning to align and translate." in StatPearls, StatPearls Publishing, 2023
3. <https://alok-shankar.medium.com/understanding-googles-attention-is-all-you-need-paper-and-its-groundbreaking-impact-c5237043540a>
4. <https://medium.com/@alejandro.itoaramendia/attention-is-all-you-need-a-complete-guide-to-transformers-8670a3f09d02>



THANK YOU

SAKSHI GUPTA

230103088

RITHVIK PONNAPALLI

230123052

