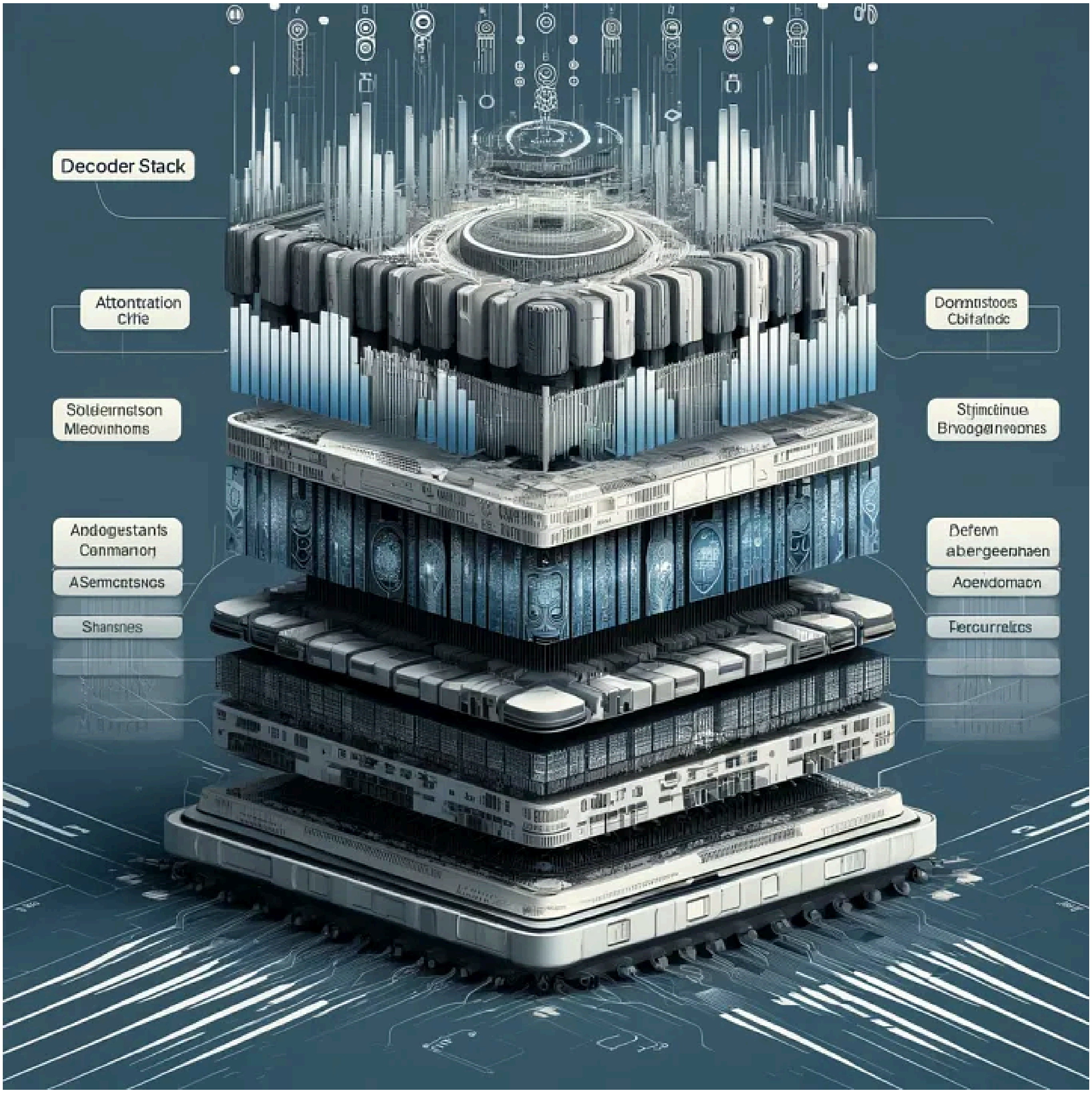
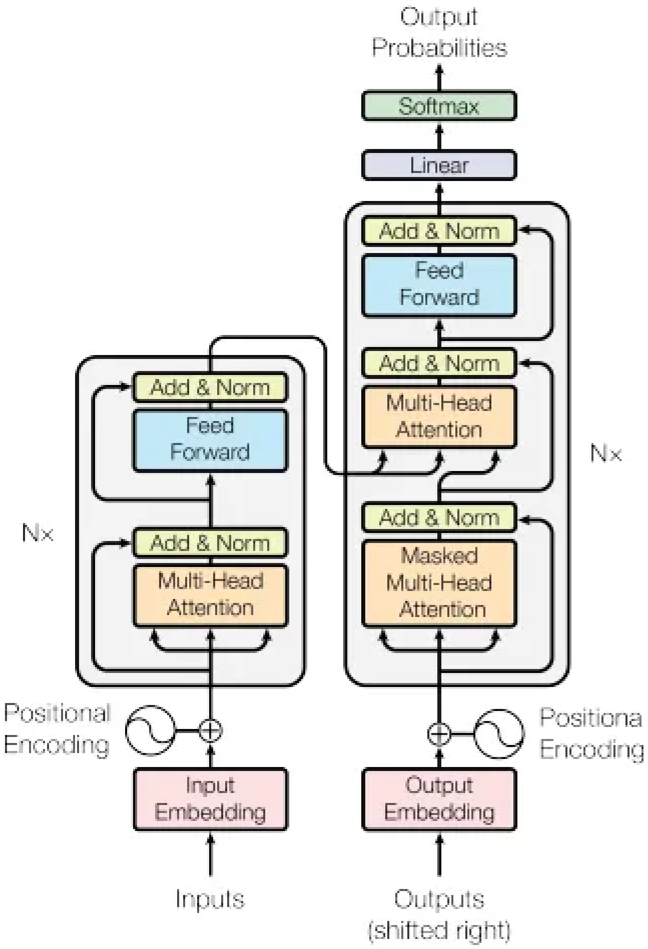
The decoder stack in the Transformer model

Sandaruwan Herath · Follow

Published in Data Science and Machine Learning · 3 min read · May 1, 2024



The decoder stack in the Transformer model, much like its encoder counterpart, consists of several layers, each featuring three main components. These are a multi-headed masked attention mechanism, a multi-headed attention mechanism, and a fully connected feedforward network. This structure is consistently repeated across all layers of the decoder.



Encoder-Decoder model[1]

The Decoder Architecture

The decoder in the Transformer model mirrors the encoder with layers that are stacked upon each other, but with additional components tailored for generating predictions:

1. Multi-Headed Masked Attention Mechanism: The first sublayer in each decoder layer is crucial for ensuring the predictions for each token are made based only on the previous tokens. This is achieved through masking future tokens in the sequence during training, which simulates the real-world scenario of predicting the next word in a sentence without knowing it.
2. Multi-Headed Attention Mechanism: Following the masked attention, this layer helps the decoder focus on relevant parts of the input sentence, irrespective of their position. It does this by attending to the outputs of

the encoder, facilitating the incorporation of pertinent context into the prediction process.

3. Position-wise Feedforward Networks: Each layer also includes a feedforward neural network, which applies the same weights to each position separately, processing the data identically across the sequence. This uniform treatment helps maintain consistency in data handling.

How It Works: Translating a Sentence

To illustrate, consider translating the sentence “The quick brown fox jumps over the lazy dog” into French. Here’s how the decoder contributes to this process:

- Starting the Sequence: The decoder generates the translation with an initial token.
- Word by Word Translation: For each subsequent word in the output:

> The masked multi-head attention layer ensures that the prediction for the current word only considers previously translated words.
>
> The encoder-decoder attention layer pulls relevant context from the entire input sequence, helping to focus on the appropriate input words.
>
> The feedforward network processes this information to contribute to the prediction of the next word.

- Ending the Sequence: This continues until the model predicts an `<end>` token, marking the end of the translation.

Practical Example of Output Generation

For the sentence above, the Transformer might generate a sequence in French:

- Input: `<start>` + Context from the encoder
- Prediction sequence: “Le”, “renard”, “brun”, “rapide”, “saute”, “par-dessus”, “le”, “chien”, “paresseux”, `<end>`

Training and Performance

Training the Transformer involves large datasets and computational power. Models are often trained on millions of sentence pairs, leveraging powerful GPUs. The performance is evaluated using the BLEU score, where higher scores indicate better translations. The Transformer’s ability to handle parallel data processing significantly reduces training time without sacrificing accuracy.

Conclusion

The decoder of the Transformer model plays a vital role in the model’s ability to perform complex translation tasks. By leveraging masked and encoder-decoder attention mechanisms, along with feedforward networks, it effectively generates coherent and contextually appropriate translations. The Transformer architecture not only marks a significant advancement in machine translation but also sets a foundation for future innovations in NLP.

Further Exploration

As NLP continues to evolve, the Transformer model remains at the forefront, inspiring new architectures and models. Its versatile framework is adaptable for a range of tasks beyond translation, including summarization, text generation, and more. The ongoing research and development in this area promise even more exciting advancements in the future.

References

[1]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

[2]. Rothman, D. (2024). *Transformers for Natural Language Processing and Computer Vision*. Packt Publishing.

Deep Learning

Transformer Model