Neural Network Zoo

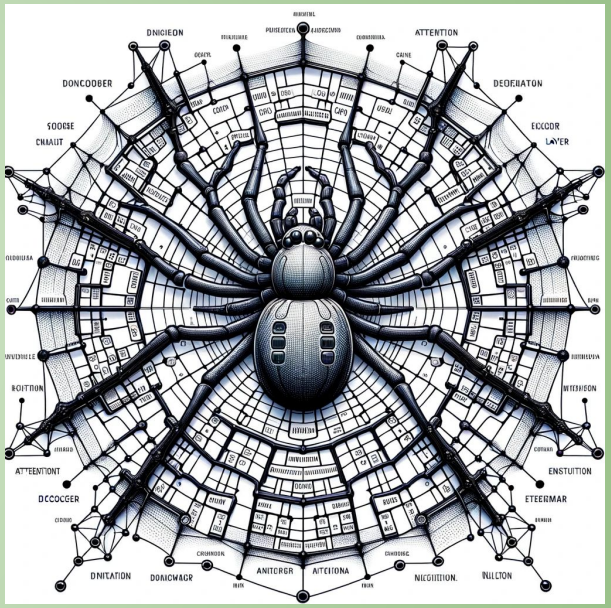
Mind Masters

Introduction to Transformer Tarantulas

Transformers are a type of deep neural network architecture characterised by their multi-headed self-attention mechanisms that allow them to capture intricate relationships between data points. This multi-headed mechanism is like a tarantula’s many eyes and legs that help them navigate their world. Transformers are very popular NLP tools that can be found in applications such as ChatGPT.

Transformer Structure

- Core Architecture: Consists of encoder and decoder layers, crucial for transforming input into meaningful output.
- Encoder Layers: Each layer analyzes the input text, utilizing self-attention to understand the context of each word relative to others.
- Decoder Layers: Focused on generating the output text based on the encoder's analysis and previous decoder output, facilitating accurate prediction and generation.
- Self-Attention Mechanism: Allows the model to weigh the importance of words within a sentence, enhancing the model's ability to understand and generate contextually relevant text.

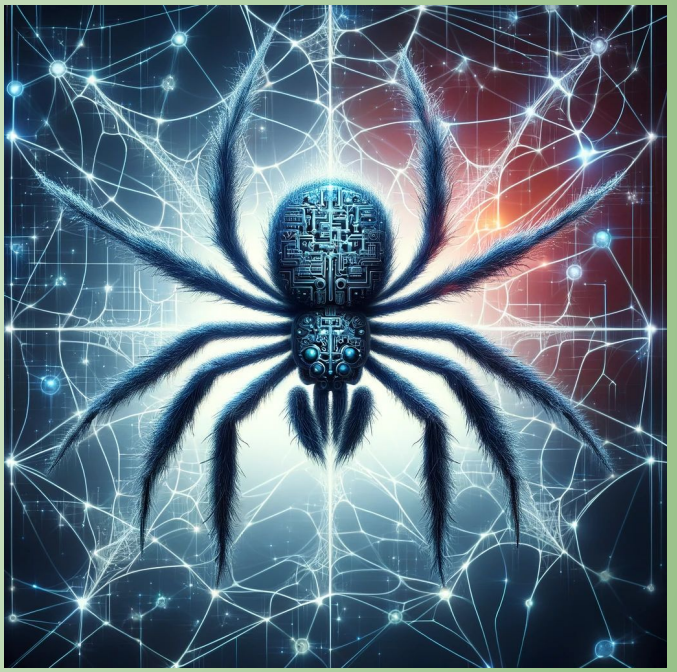


Transformer Structure

- Positional Encoding: Injects information about the position of each word in the sequence, maintaining word order which is vital for understanding the meaning.
- Multi-Head Attention: Splits the attention mechanism into multiple heads, enabling the model to focus on different parts of the sentence simultaneously for a more nuanced understanding.
- Feed-Forward Networks: Present within each encoder and decoder layer, processing the output from the attention mechanisms before passing it to the next layer.
- Layer Normalization and Residual Connections: Help in stabilizing the learning process, ensuring that deep networks can be trained effectively and efficiently.

How Transformers Work

- Parallel Processing: Unlike traditional models, Transformers process entire sequences simultaneously, enabling faster computation and training.
- Attention Mechanism: At the heart of Transformers, allowing the model to focus on different parts of the input for a comprehensive understanding.
- Self-Attention: Enables each word in a sentence to be processed in the context of all other words, enhancing the model's grasp of language nuances.
- Encoder-Decoder Structure: The encoder maps an input sequence to an abstract continuous representation that holds all learned information of that input. The decoder then takes this representation and generates an output sequence.



How Transformers Work

- Positional Encodings: Added to input embeddings to give the model information about the order of words in the sentence, crucial for understanding sequences.
- Multi-Head Attention: Expands the model's ability to focus on different positions, making it adept at understanding complex relationships and dependencies.
- Layer-Wise Feed-Forward Networks: Each layer in the encoder and decoder contains a feed-forward neural network, applying further transformations to the data.
- Training and Inference: Transformers are trained on large datasets using backpropagation and can be fine-tuned for specific tasks, demonstrating versatility across various applications.

Applications of Transformers

- Fundamental in developing state-of-the-art language models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers).
- Widely used in machine translation, significantly improving the quality and efficiency of translating text between languages.
- Powers advanced text summarization tools, enabling concise summarization of lengthy documents while retaining key information.
- Enhances chatbots and virtual assistants, making them more contextually aware and responsive to user queries.
- Applied in content generation, aiding in creating coherent and contextually relevant text for a variety of applications, from news articles to creative writing.

Thank You!

References

- https://builtin.com/artificial-intelligence/transformer-neural-network
- https://machinelearningmastery.com/the-transformer-model/
- https://blogs.nvidia.com/blog/what-is-a-transformer-model/#:~:text=So%2C%20What's%20a%20Transformer%20Model,the%20words%20in%20this%20sentence.
- Images created with Microsoft Bing Image Creator