


Wikipedia Research Helper using LangChain & FLAN-T5 XXL GPT



What topic do you wish to conduct research on?

Transformers (deep learning architecture)

Short Summary

Human: Transformers is a deep learning architecture that combines a neural network with a transformer network. AI: Transformers is a deep learning architecture that combines a neural network with a transformer network.

Detailed Summary

Page: Transformer (deep learning architecture) Summary: A transformer is a deep learning architecture developed by Google and based on the multi-head attention mechanism, proposed in a 2017 paper "Attention Is All You Need". Text is converted to numerical representations called tokens, and each token is converted into a vector via looking up from a word embedding table. At each layer, each token is then contextualized within the scope of the context window with other (unmasked) tokens via a parallel multi-head attention mechanism allowing the signal for key tokens to be amplified and less important tokens to be diminished. The transformer paper, published in 2017, is based on the softmax-based attention mechanism proposed by Bahdanau

et. al. in 2014 for machine translation, and the Fast Weight Controller, similar to a transformer, proposed in 1992. Transformers have the advantage of having no recurrent units, and thus requires less training time than previous recurrent neural architectures, such as long short-term memory (LSTM), and its later variation has been prevalently adopted for training large language models (LLM) on large (language) datasets, such as the Wikipedia corpus and Common Crawl.

This architecture is now used not only in natural language processing and computer vision, but also in audio and multi-modal processing. It has also led to the development of pre-trained systems, such as generative pre-trained transformers (GPTs) and BERT (Bidirectional Encoder Representations from Transformers).

Page: Mamba (deep learning architecture) Summary: Mamba is a deep learning architecture focused on sequence modeling. It was developed by researchers from Carnegie Mellon University and Princeton University to address some limitations of transformer models, especially in processing long sequences. It is based on the Structured State Space sequence (S4) model.

Page: Multimodal learning Summary: Multimodal learning, in the context of machine learning, is a type of deep learning using multiple modalities of data, such as text, audio, or images. In contrast, unimodal models can process only one type of data, such as text (typically represented as feature vectors) or images. Multimodal learning is different from combining unimodal models trained independently. It combines information from different modalities in order to make better predictions. Large multimodal models, such as Google Gemini and GPT-4o, have become increasingly popular since 2023, enabling increased versatility and a broader understanding of real-world phenomena.

