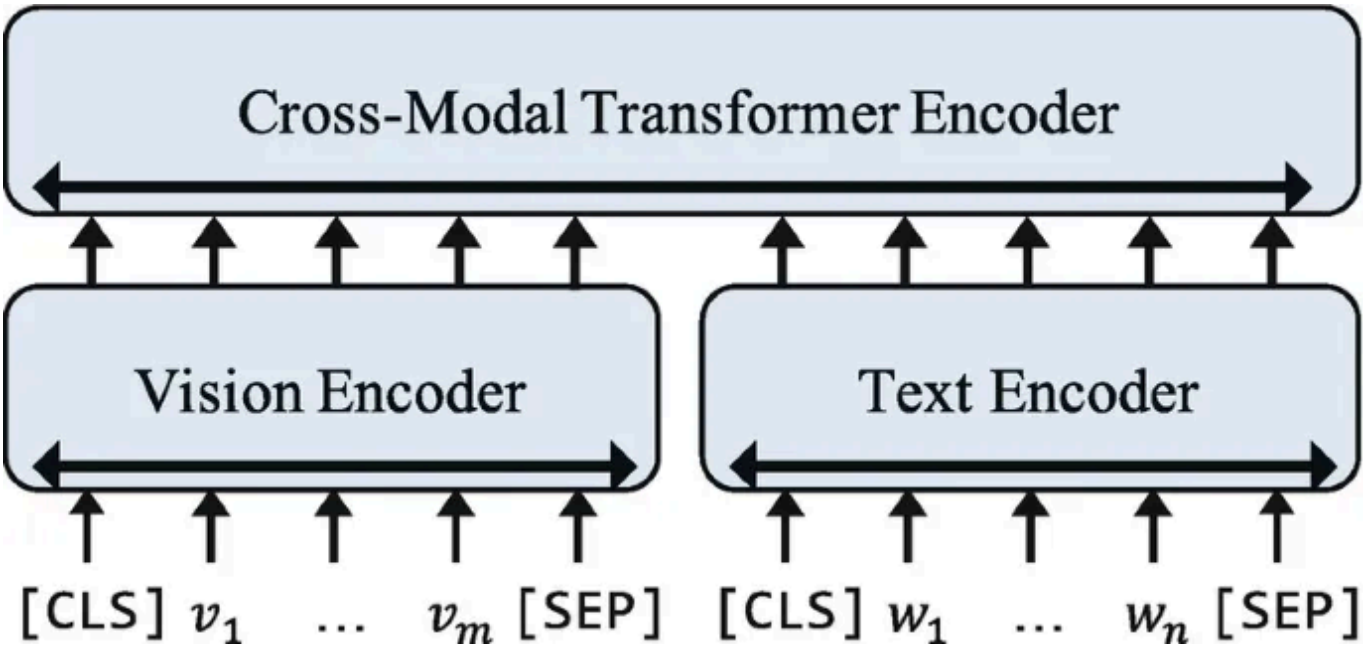
Cross-Modal Attention

Abdulkader Helwan · Follow
3 min read · Jan 21, 2025



Cross-Modal Attention

Cross-modal attention is a mechanism used in multimodal models to dynamically align and fuse information from different modalities (e.g., text, images, audio). It allows the model to focus on the most relevant parts of one modality when processing another, enabling effective interaction between

modalities. Cross-modal attention is a key component in many state-of-the-art multimodal models, such as **DALL-E**, **CLIP**, and **ViLT**.

What is Cross-Modal Attention?

While self-attention computes relationships within a single modality (e.g., between words in a sentence or patches in an image), cross-modal attention computes relationships **between different modalities** (e.g., between text and image features).

The idea is to Use one modality as a **query** to attend to the other modality's **key-value pairs**.

This allows the model to dynamically align and combine information from both modalities.

How Does Cross-Modal Attention Work?

Cross-modal attention operates similarly to self-attention but involves two modalities. Here's how it works:

- **Modality A (Query):** Represents one modality (e.g., text embeddings).
- **Modality B (Key-Value):** Represents the other modality (e.g., image embeddings).

Projections:

The embeddings from both modalities are projected into three spaces:

- **Query (Q):** Derived from Modality A.

- **Key (K):** Derived from Modality B.
- **Value (V):** Derived from Modality B.

Attention Scores:

- Compute the dot product between the Query (Q) and Key (K) to measure the similarity between elements of Modality A and Modality B.
- This results in an **attention score matrix** that indicates how much each element of Modality A should attend to each element of Modality B.
- Apply a softmax function to the attention scores to normalize them into probabilities.

Weighted Sum:

- Use the attention probabilities to compute a weighted sum of the Values (V) from Modality B.
- This produces a **context vector** that represents the most relevant information from Modality B for each element of Modality A.

Output:

- The context vectors are combined with the original embeddings from Modality A to produce the final cross-modal representation.

Mathematical Formulation:

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

Where:

- Q: Query matrix (from Modality A).
- *K*: Key matrix (from Modality B).
- V: Value matrix (from Modality B).
- *dk*: Dimensionality of the key vectors (used for scaling).

Example: Cross-Modal Attention in DALL-E

In **DALL-E**, cross-modal attention is used to condition image generation on text:

Text Encoder:

- A Transformer encodes the input text into a dense embedding (Query).

Image Tokenizer:

- The input image is tokenized into discrete tokens (Key-Value).

Cross-Attention:

- The text embedding (Query) attends to the image tokens (Key-Value) to condition the image generation process.

Image Decoder:

- A VQ-VAE decoder generates the final image based on the conditioned tokens.

Attention Cross Attention Multimodal AI Transformers

Written by Abdulkader Helwan

827 Followers · 235 Following

Research Scientist, AI in Medicine, Technical Reviewer at ContentLab.io.
https://bio.site/AbdulkaderH

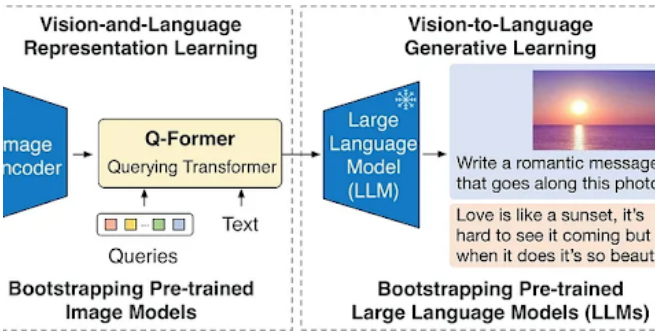
Follow

No responses yet

Write a response

What are your thoughts?

More from Abdulkader Helwan



Abdulkader Helwan

Q-Former

The ability to seamlessly integrate and process information from both visual and...

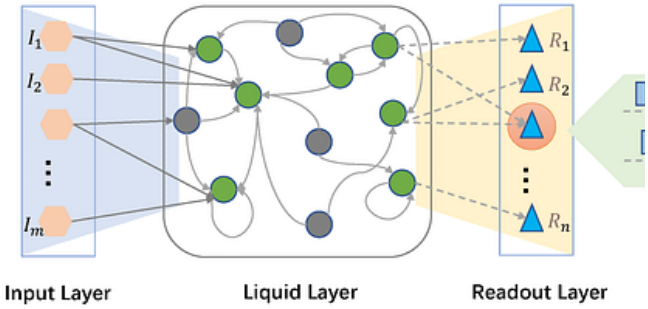
Dec 22, 2023 69

Abdulkader Helwan

Introduction to Image Embeddings

This blog post discusses image embeddings and its implementation in Python. I hope you...

Oct 1, 2023 217 2

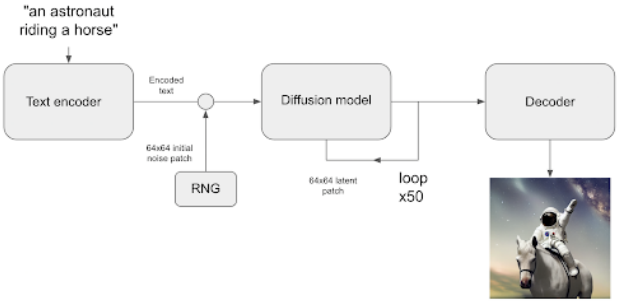


In Python in Plain English by Abdulkader Helwan

Liquid Neural Networks: Simple Implementation

Implementing Liquid Neural Network in TensorFlow

Feb 23, 2024 504 3



Abdulkader Helwan

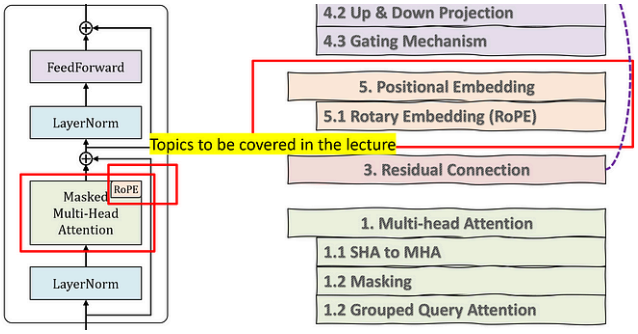
Text-to-Image Generation Model with CNN

Text-to-image models are a type of machine learning model that takes an input natural...

Dec 28, 2023 60

See all from Abdulkader Helwan

Recommended from Medium

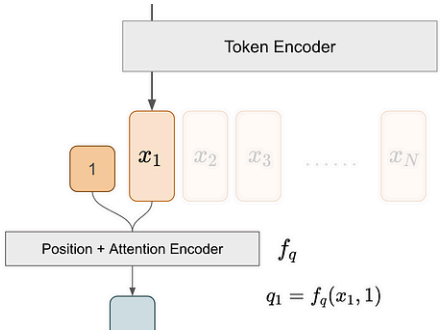


Hugman Sangkeun Jung

Mastering LLama —Understanding Rotary Positional Embedding...

Deep Dive into RoPE: Advanced Positional Encoding in Language Models

Nov 22, 2024 10

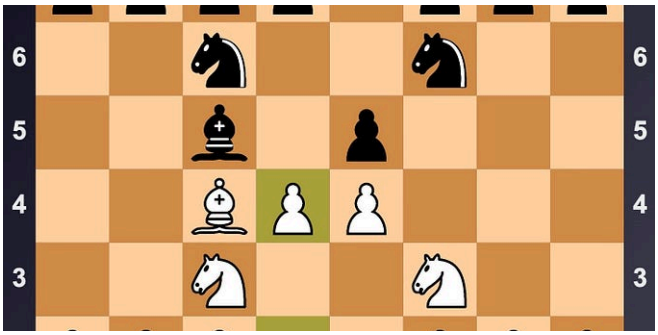


Allen Liang

RoPE: A Detailed Guide to Rotary Position Embedding in Modern...

Rotary Position Embedding (RoPE) has been widely applied in recent large language...

Dec 10, 2024 31 1

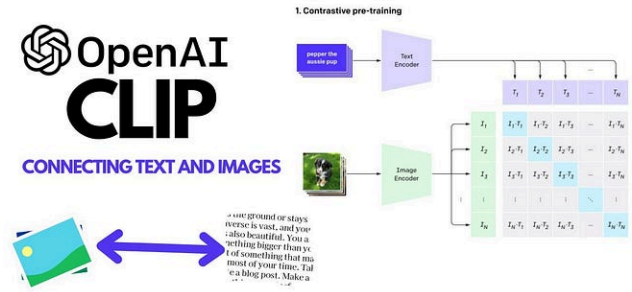


In AI Advances by Harys Dalvi

LLMs Do Not Predict the Next Word

RLHF forces us to view LLMs as agents in an environment, not just statistical models.

Apr 3 1.6K 31



Mohammed Lubbad

A Beginner's Guide to the CLIP Model

How It Brings Images and Text Together

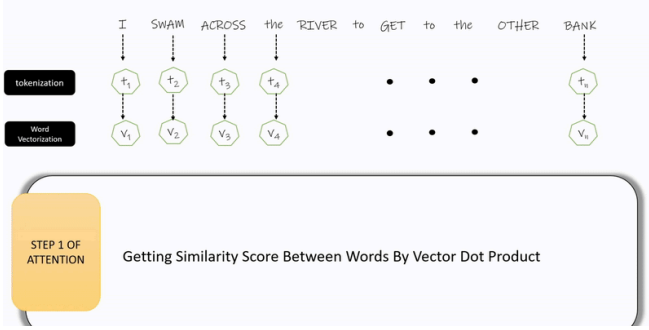
Nov 9, 2024 2

In Level Up Coding by Sahib Dhanjal

How To Train Your PyTorch Models With Less Memory

Strategies I regularly use to reduce GPU memory consumption by almost 20x

Feb 24 432 3



Maninder Singh

The Detailed Explanation of Self-Attention in Simple Words

The attention mechanism is one of the most groundbreaking concepts in deep learning. I...

Apr 1 132

See more recommendations