AI | MACHINE LEARNING INTERNSHIP

Omar Saad

NEURAL NETWORKS

-Activation Functions. Gradient

-ANNs

-RNNs. Recurrency

-CNNs. Padding & Strides

TRANSFORMERS

SELF ATTENTION MECHANISM

Reads different parts of the sequence at the same time and focuses on important context instead of the traditional word by word reading.

ENCODING

Input Embeddings converts input sequence to vector representations.

Positional Encoding generates unique signal for each word to preserve it's order.

ATTENTION IS ALL YOU NEED

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

Where Q (query), K (key), and V (value)

TRANSFORMERS VS CNN VS RNN

Transformers have a longer sequence memory unlike RNNs, On the other hand RNNs require less computational power.

CNNs excel Transformer vision models in the image classification domain hence to their convolutional layers

BERT

Architecture

Based on Transformers architecture.

Where BERT Base has 12 Transformer encoder layers and 24 in BERT Large.

Transformers vs BERT

While Transformers focus on next word generation that depends on the previous tokens.

BERT focuses on understanding the sequence's context.

Bidirectionality

Bidirectional Encoder Representations from Transformers.

Reads text from both sides at the same time.

Contextual embedding & context sensitivity

Contextual Embedding are dynamic based on context compared to Transformer's static embedding.

Context Sensitivity found in BERT incorporates context from the very beginning of sequence rather than Transformers ignoring initial embedding.

RAG

Retrieval **A**ugmented **G**eneration

Advantages

Domain specific, Real-time updates and Secure.

Disadvantages

High Computational Cost, Latency Issues, Complexity in Fine-Tuning and Dependency on Clean Data.

Vector Data Bases

Conduct similarity search with stored vector data dimensions by calculating shortest context distance.

LangChain

Frame-Work for LLMs with the ability to chain multiple different models together to form a pipeline.

Provides multiple features like Agents, Vector DB, Indexes, Prompts, Memory, LLMs and lastly Chains.

EVALUATIONS

Overfitting and Underfitting:
•**Overfitting** model learns the training data too well, including noise and outliers, resulting in poor generalization to new data.
•**Underfitting** model is too simple to capture the underlying patterns in the data, leading to poor performance on both training and test sets.

Confusion Matrix:
• A confusion matrix is a table that summarizes the performance of a classification model. It shows the counts of True positives, false positives, true negatives, and false negatives, providing insight into where the model is making errors.

Confidence Score:
• A confidence score indicates the likelihood that a model's prediction is correct. It reflects the model's certainty in its prediction.
• Typically represented as a probability value between 0 and 1.

RAG Evaluation:
• RAG models are evaluated based on both retrieval accuracy (how well the relevant documents are retrieved), Generation quality (how coherently and accurately the response is generated).
• This dual evaluation makes the process more complex compared to traditional models.

THANK YOU

Special thanks to Alaa and Yara for their efforts