ZTM RAG: Generative AI Fundamentos

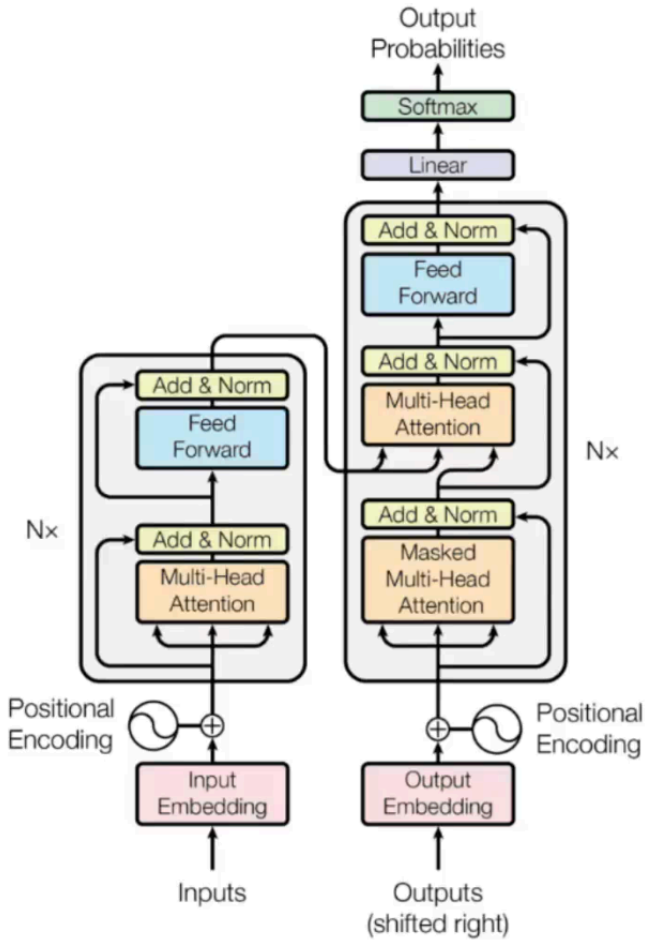
quinta-feira, 26 de junho de 2025 13:22

Fine Tuning

Fine-tuning (ou ajuste fino) é o processo de pegar um modelo pré-treinado (como GPT, BERT, T5, etc.) e treiná-lo mais um pouco em uma tarefa específica, com seus próprios dados.

Arquitetura Transformers

É uma arquitetura de rede neural. A ideia principal é usar mecanismos de atenção para entender relações entre palavras, sem depender de RNNs ou CNNs.
O transformers inclui um encoder e um decoder, transformando input sequences em representações detalhas e outputs.



A esquerda é o encoder e a direita é o decoder.
Obs:

- Em alguns casos, o modelo pula os processos "Multi Head Attention" e "Feed Forward" para melhorar o desempenho e preservar dados chaves sem processamento desnecessário.
- Nx = A sequencia é repetida para criar representações profundas dos dados

Componentes Principais / Camadas

Input / Output embedding

Adiciona informação de posição aos embeddings, ajudando o modelo a entender a ordem das palavras em uma sequencia.

Multi-Head Attention
É o core component. Permite o modelo focar em diferentes partes do input simultaneamente usando multiple "heads" que captura varios aspectos dos dados
Exemplo:
Input = "gato sentou no tapete"
head 1 = gato + tapete
head 2 = sentou + tapete
head 3 = gato + sentou

Add & Norm
Balanceia tudo. Adiciona atalhos de conexão e normaliza os outputs, garantindo estabilidade no aprendizado.

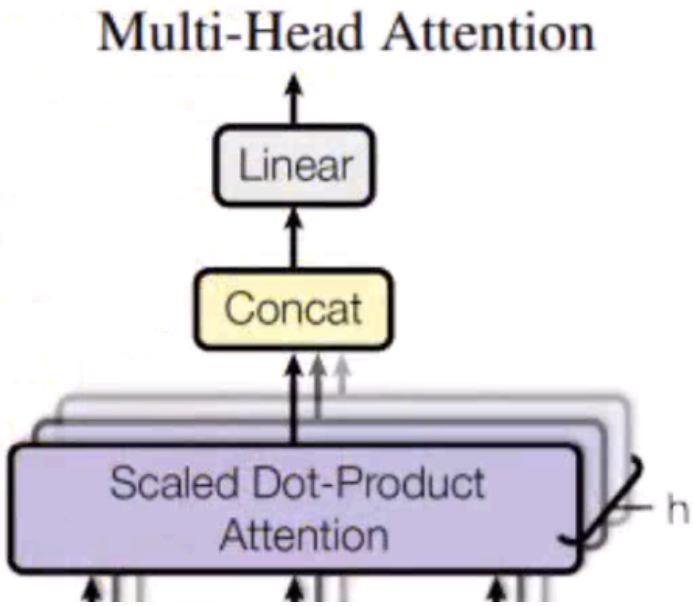
Feed Forward
Pega a informação e faz uma série de transformações para torna-los mais precisos e uteis. Os dados passam por uma refinaçao linear e depois não linear. Ajuda a entender parametros mais complexos.

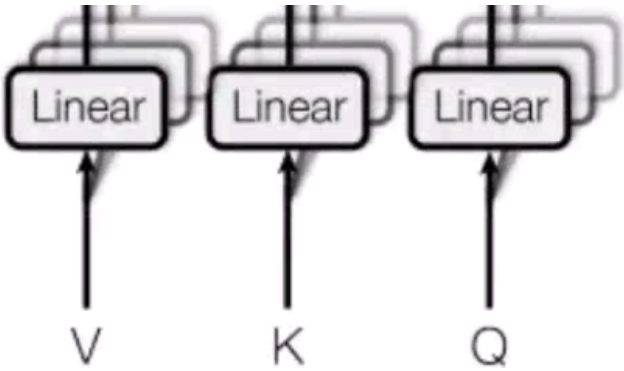
Masked Multi-Head Attention
Similar ao Multi-Head Attention. Mas o Masked significa que o modelo não pode ver tokens futuros na sequencia, apenas tokens que já foram processados. Garantindo que as predições são baseadas apenas nos tokens passados e presentes.

Linear
Transforma o output em um formato onde cada dimensão corresponde a um potencial output token.

Softmax
Converte os outputs lineares em probabilidades, destacando os tokens mais prováveis na sequencia predita.

Attention Mechanisms
Multi-Head Attention dentro da arquitetura transformers





V K Q = Values, Keys, Queries
h = número de attention heads.

Inputs > Linear
Faz uma transformação linear nos inputs para vectors que podem interagir efetivamente entre eles.

Scaled Dot-Product Attention
Calcula o attention score (mede a relevância de cada palavra em uma querie), escala eles (previne que os valores fiquem grandes demais), e depois aplica Softmax para converter os scores em probabilidades, focando nas partes do input mais relevantes. Gera um attention score.
Todo esse processo é executato em paralelo multiplas vezes. Cada vez foca em uma parte diferente da sequencia de input.

Concat
Combina todos os attention heads em um vetor calcatenado, providenciando uma comprehensive view integrando multiplas perspectivas.

Concat > Linear
Transformação do output contatenado em uma representação coesa.