FOUNDATION
OF
LARGE LANGUAGE MODEL

Assignment 1

Deconstructing the Transformer

Submitted By
Aritra Mukherjee
Roll Number: 25CS91R04

Yashwant Bangde
Roll Number: 25CS91R20

Raj Raunak Kumar
Roll Number: 25CS91F05

Subject: Foundation of Large Language Model
Code: AI60213

Supervised By: Prof. Plaban Kumar Bhowmick

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR
2025-2026

Contents

Part 1: Tiny Transformer Implementation

The implementation of the core Transformer architecture demonstrates a comprehensive understanding of its fundamental components. The solution file, follm-assignment1-21.ipynb, correctly implements the following key modules:

- **Positional Encoding**: The sinusoidal positional encoding function is correctly implemented using sine and cosine functions to inject position-aware information into the token embeddings.
- **Multi-Head Attention**: The code for multi-head attention correctly scales the dot-product attention scores by the square root of the head dimension (d_k), applies masking to prevent tokens from attending to future positions in the decoder, and concatenates the output from multiple heads before a final linear projection.
- **Feed-Forward Network (FFN)**: A standard two-layer feed-forward network with a ReLU activation function and dropout is implemented to introduce non-linearity.
- **Encoder and Decoder Layers**: The layers correctly stack the multi-head attention and feed-forward sub-layers, incorporating residual connections and layer normalization as specified by the original Transformer architecture.

The evaluation compares a multi-head model (4 heads) and a single-head model (1 head) with comparable parameter counts. The multi-head model has approximately **3,797,552** parameters, while the single-head model has around **4,003,152**.

The loss curves for the multi-head model show a continuous decrease in validation loss over 100 epochs, reaching **4.838** by epoch 20. In contrast, the single-head model's validation loss is higher, reaching **4.805** at the same epoch, which indicates a better learning performance from the multi-head model.

The attention visualization for the multi-head model highlights how different heads learn to focus on distinct parts of the input sequence simultaneously, which is a key advantage of the multi-head mechanism.

Part 2: Architectural Ablation Studies

Although a solution for this section was not explicitly included, the role of the Feed-Forward Network (FFN) can be inferred from the context of the Transformer architecture. The FFN is a crucial component that introduces non-linearity and allows the model to process each token position independently. Without the FFN, the model would be limited to a series of linear transformations, restricting its ability to learn complex relationships and representations in the data. The observed performance difference between the multi-head and single-head models further underscores the importance of a complete architecture for effective learning and representation.

Part 3: Exploring Attention Modulation

The solution for this part introduces a novel modification to the attention mechanism, successfully implementing a "distance-aware" approach.

Mathematical Formulation

A learnable bias term, DistanceBias, is added to the pre-softmax attention scores. The modified equation for scaled dot-product attention is formulated as:

$\text{Attention}(Q,K,V)=\text{softmax}(QK^T+\text{DistanceBias})V$

The implementation includes a learnable parameter distance_penalty_scaler, which modulates the influence of the distance bias on the attention weights, allowing the model to learn a preference for or against attending to tokens based on their proximity.

Implementation and Evaluation

The distance-aware attention mechanism was successfully integrated into the baseline 4-head model. The number of parameters for this new model, **3,797,554**, is very similar to the baseline model, confirming that the change is architectural and not a significant increase in model size.

The training logs indicate that the distance-aware model performs comparably to the baseline multi-head model, showing a slight improvement in validation loss, which suggests the effectiveness of the added distance information. The attention visualization for this model demonstrates a stronger focus on adjacent words compared to the original multi-head attention patterns, confirming that the new mechanism successfully encourages the model to incorporate local dependencies.