September 20

1 Deep Learning and Transformers

Today's focus was on advanced deep learning architectures, particularly the Transformer model.

1.1 Transformer Architecture

The Transformer [?] revolutionized NLP with its attention mechanism:

Self-Attention Formula:

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

where:

- Q = Queries matrix
- K = Keys matrix
- V = Values matrix
- d_k = Dimension of key vectors

1.2 Multi-Head Attention

Instead of single attention, use h parallel attention heads:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

1.3 Positional Encoding

Since Transformers have no recurrence, positional information is added:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

1.4 Training Insights

1. **Warm-up**: Learning rate increases linearly for first steps
2. **Layer Normalization**: Applied before each sub-layer
3. **Residual Connections**: Help with gradient flow in deep networks
4. **Label Smoothing**: Prevents overconfidence, $\epsilon = 0.1$ typically

Key Advantage: Parallelizable training unlike RNNs, enabling large-scale models like GPT and BERT.