Deep Learning: A Brief Overview

A demonstration of NovaType for VS Code

1. Introduction

This article is made using NovaType

Deep learning has revolutionized artificial intelligence in recent years. The transformer architecture [1] has become the foundation of modern language models.

2. Mathematical Foundations

2.1. The Softmax Function

The softmax function normalizes a vector into a probability distribution:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{1}$$

As shown in Equation 1, the output values sum to 1.

2.2. Attention Mechanism

The scaled dot-product attention is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{\text{model}}}}\right)V \tag{2}$$

The attention mechanism (Equation 2) allows the model to focus on relevant parts of the input.

3. Results

Model	Parameters	Accuracy
GPT-2	1.5B	94.2%
BERT	340M	93.1%
T5	11B	96.8%

Table 1: Comparison of transformer models

Table 1 shows the performance comparison. For more details on BERT, see [2].

4. Conclusion

The transformer architecture has enabled significant advances in NLP, as demonstrated by [3].

Bibliography

[1] A. Vaswani *et al.*, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[3] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.