Self-Attention

什麼是 Self-Attention？

Self-Attention 是自然語言處理 (NLP) 和深度學習中常用的一種機制。它用於計算序列內每個詞與其他詞的相關性，幫助模型聚焦於重要的上下文資訊。

工作原理

1. **輸入向量化**
    - 將句子中的每個詞轉換為向量表示 (embeddings)。
2. **計算 Query、Key 和 Value**
    - **Query (查詢):** 表示當前詞需要從上下文中獲取什麼資訊。
    - **Key (鍵):** 提供句子中每個詞的相關特徵。
    - **Value (值):** 提供每個詞的實際資訊。
3. **計算注意力權重**
    - Query 和 Key 計算點積以衡量相關性。
    - 使用 Softmax 函數將相關性轉化為權重。
4. **加權求和**
    - 將權重與 Value 相乘並加總，得到新的詞表示。

優點

- **全局視角:** 能考慮句子內所有詞的關係。
- **動態:** 自動調整權重，不依賴於固定的窗口大小。
- **高效:** 適合並行計算。

應用範疇

Self-Attention 是 Transformer 模型（如 GPT、BERT）的核心，用於文本生成、分類和翻譯等任務。