PLANO

Mecanismos de Atenção em NLP: Fundamentos e Aplicações

Slide 1: Introdução

- **O que é Atenção em NLP?**

  Um mecanismo que permite que modelos foquem nas partes mais relevantes da entrada ao gerar uma saída.

- **Importância**

  Superou limitações de modelos sequenciais tradicionais (RNNs, LSTMs).

Slide 2: Motivação

- **Problema nas RNNs/LSTMs**

  Dificuldade em lidar com longas dependências.

- **Solução**

  Mecanismos de atenção capturam relações entre todas as palavras diretamente.

Slide 3: Conceito Básico de Atenção

- Cada palavra da entrada recebe um **peso** de importância.
- A saída é uma **combinação ponderada** das entradas.
- **Foco seletivo**: o modelo aprende **onde prestar atenção**.

Slide 4: Atenção Escalar Simples

- Fórmula base:

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

- Onde:
  - Q = Queries
  - K = Keys
  - V = Values

Slide 5: Self-Attention

- Cada palavra **atende a todas as outras** no mesmo input.
- Importante para:
  - Entender o **contexto global**.
  - Capturar relações **longas e curtas** entre tokens.

Slide 6: Multi-Head Attention

- Várias "cabeças" de atenção operam em paralelo.
- Cada cabeça aprende **relações diferentes**.
- Fórmula:

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

- Benefício: **riqueza contextual**.

Slide 7: Atenção na Prática: Transformer

- Base do modelo Transformer (Vaswani et al., 2017).
- Arquitetura elimina RNNs: **pura atenção**.
- Utiliza:
  - Multi-Head Attention
  - Feed-Forward Networks
  - Positional Encoding

Slide 8: Variações de Atenção

- **Masked Attention** (Ex: GPT)

  Atenção apenas ao passado (evitar olhar o futuro em geração).

- **Cross-Attention** (Ex: T5)

  Atende a outra sequência (ex: input → output).

Slide 9: Aplicações em NLP

- Tradução automática (Ex: Transformer, T5)
- Resumo automático de textos
- Pergunta e resposta (Ex: BERT QA)
- Geração de texto (Ex: GPT)

Slide 10: Conclusão

- **Atenção revolucionou NLP** ao permitir modelagem eficiente de dependências complexas.
- **Modelos modernos** são construídos sobre variantes de atenção.
- **Futuro**: Pesquisas em atenção mais eficiente (Ex: Linformer, Longformer).