**Mecanismo de Atenção nos Transformers**

O mecanismo de atenção é um componente essencial nos transformers, projetado para superar as limitações das redes neurais recorrentes (RNNs), como a dificuldade de memorizar dependências longas e o desafio de paralelizar o processamento. Com base no material fornecido, aqui está uma explicação detalhada:

**Atenção por Produto Escalar Escalonado**

* Este é o núcleo do funcionamento do mecanismo de atenção nos transformers. Ele atribui pesos aos tokens do texto de entrada, permitindo ao modelo focar nas partes mais relevantes.
* **Processo**:
  1. Cada token é representado por três vetores: **Query (Q)**, **Key (K)** e **Value (V)**.
  2. O modelo calcula a similaridade entre os vetores **Query** e **Key** usando o produto escalar.
  3. Os valores resultantes são escalonados dividindo-os pela raiz quadrada da dimensão dos vetores Key (dk\sqrt{d\_k}dk​​). Essa normalização estabiliza os gradientes durante o treinamento.
  4. A normalização é feita com a função softmax, transformando os valores em probabilidades que indicam a importância relativa de cada token.
  5. O modelo usa essas probabilidades para calcular uma soma ponderada dos vetores **Value**, gerando a saída codificada.

**Codificação do Contexto**

* Cada token é codificado não apenas com base em sua própria representação, mas também considerando seu contexto no texto. Isso permite capturar relações entre palavras distantes.

**Atenção Multi-Cabeças (Multi-Head Attention)**

* Em vez de realizar uma única operação de atenção, os transformers utilizam várias "cabeças" de atenção.
* Cada cabeça opera em um subespaço diferente dos vetores de entrada, permitindo ao modelo aprender múltiplas relações semânticas e estruturais ao mesmo tempo.
* Por exemplo, uma cabeça pode capturar a relação entre sujeito e verbo, enquanto outra foca em adjetivos e substantivos.

**Codificação Posicional**

* Como os transformers não processam os tokens de maneira sequencial, é necessário adicionar informações sobre a posição relativa ou absoluta dos tokens.
* Para isso, são usadas funções senoidais e cossenoidais de diferentes frequências. Essa abordagem permite que o modelo generalize bem para sequências de comprimentos variados.

**Vantagens do Mecanismo de Atenção**

1. **Memória de Longo Alcance**: Ao considerar toda a sequência simultaneamente, os transformers conseguem capturar dependências de longo prazo entre tokens.
2. **Eficiência**: O processamento paralelo permite treinar e inferir rapidamente, em contraste com o processamento sequencial das RNNs.
3. **Escalabilidade**: O design modular do mecanismo de atenção permite que os transformers sejam expandidos e adaptados para diversas tarefas.

**Conclusão**

O mecanismo de atenção nos transformers, especialmente na forma de Atenção por Produto Escalar Escalonado e Atenção Multi-Cabeças, é o principal responsável pelo sucesso desse modelo em tarefas complexas de processamento de linguagem natural. Ele fornece uma forma poderosa e eficiente de capturar padrões em dados sequenciais, superando as limitações dos métodos anteriores.