PROCESSAMENTO DE LINGUAGEM NATURAL

EPISODE VI : TRANSFORMERS

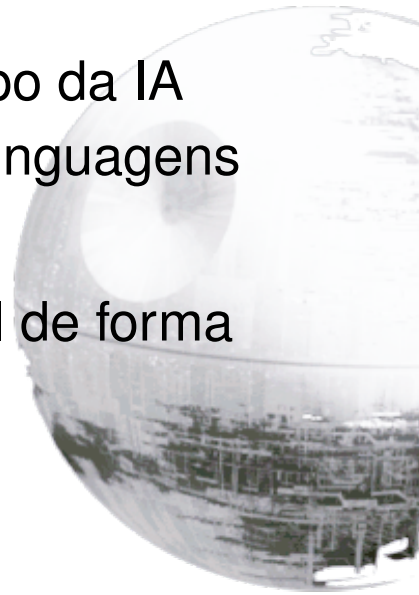
Alex Marino

14 de novembro de 2024

Conteúdo

O que é NLP?

- Processamento de Linguagem Natural (NLP) é o campo da IA que se concentra na interação entre computadores e linguagens humanas.
- NLP envolve a análise e geração de linguagem natural de forma automática.

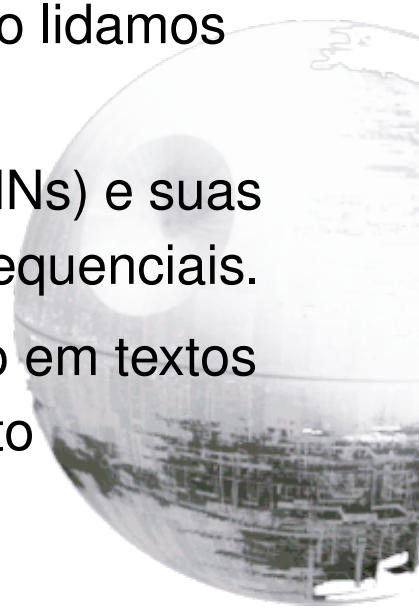


O que são Transformers?

- Introduzido por Vaswani et al. em 2017 [1].
- Substituiu RNNs/LSTMs em muitas tarefas de NLP.
- Baseado em mecanismos de atenção (self-attention).

Revolução com Transformers no NLP

- Os transformers mudaram drasticamente a forma como lidamos com tarefas de NLP.
- Antes dos transformers, redes neurais recorrentes (RNNs) e suas variantes, como LSTMs, eram o padrão para tarefas sequenciais.
- RNNs/LSTMs enfrentavam problemas de desempenho em textos longos, além de serem lentos devido ao processamento sequencial.

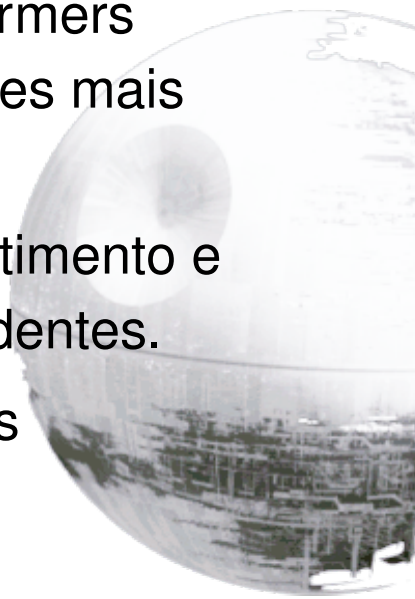


Vantagens dos Transformers

- **Paralelização**: Os transformers processam as palavras de uma sentença simultaneamente, aproveitando o poder computacional dos GPUs.
- **Autoatenção**: Captura dependências longas e contextos complexos de maneira eficiente.
- **Escalabilidade**: Modelos como BERT e GPT podem ser escalados para bilhões de parâmetros, aproveitando dados em grande escala.

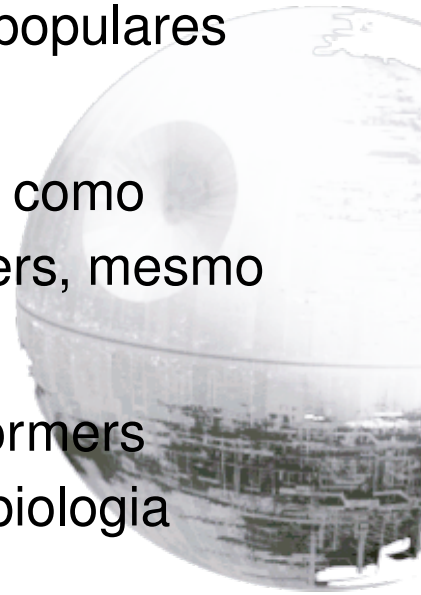
Tarefas Resolvidas com Transformers

- **Tradução Automática**: Modelos baseados em transformers superaram os métodos anteriores, entregando traduções mais precisas e fluídas.
- **Classificação de Texto**: Aplicado para análise de sentimento e classificação de tópicos com uma precisão sem precedentes.
- **Sumarização de Texto**: Capacidade de gerar resumos coerentes a partir de grandes textos.



Impacto dos Transformers no NLP

- **Melhorias em Benchmarks**: Modelos como BERT e GPT-3 elevaram significativamente o padrão de benchmarks populares de NLP, como GLUE e SuperGLUE.
- **Democratização de Modelos Grandes**: Ferramentas como Hugging Face facilitam o acesso e o uso de transformers, mesmo sem grandes recursos computacionais.
- **Influência em Outras Áreas**: Além de NLP, os transformers começaram a ser aplicados em visão computacional, biologia computacional e mais.

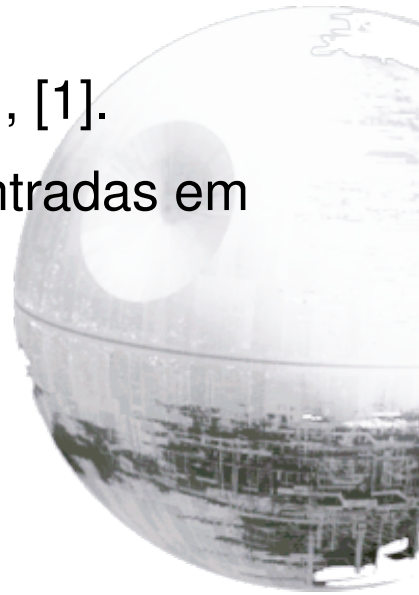


Arquitetura Transformer

- Composto por blocos de Encoder e Decoder.
- Usa Multi-head Self-Attention para processar a entrada em paralelo.
- Elimina a necessidade de processar dados sequencialmente (como em RNNs).

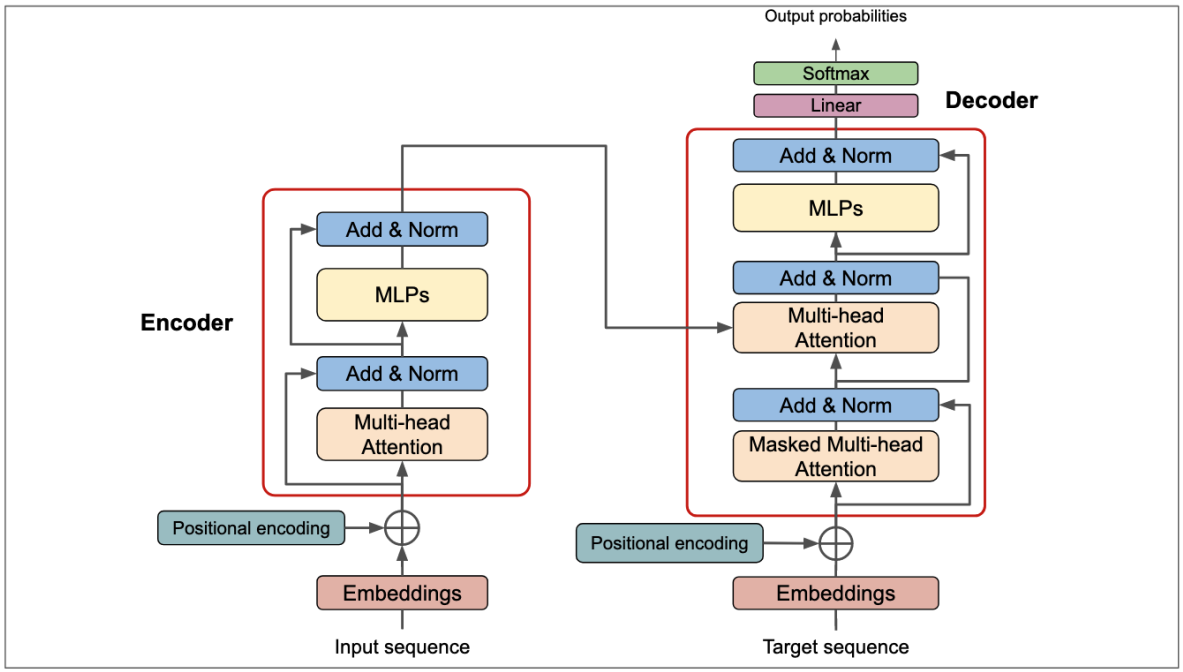
Introdução à Arquitetura Transformer

- O Transformer foi introduzido por Vaswani et al. (2017), [1].
- Usa o mecanismo de **self-attention** para processar entradas em paralelo.
- Composto por uma arquitetura **Encoder-Decoder**.



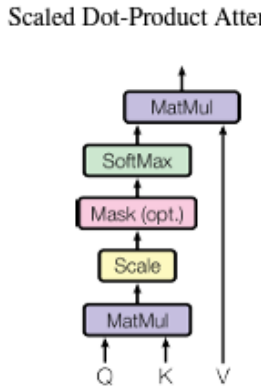
O Encoder no Transformer

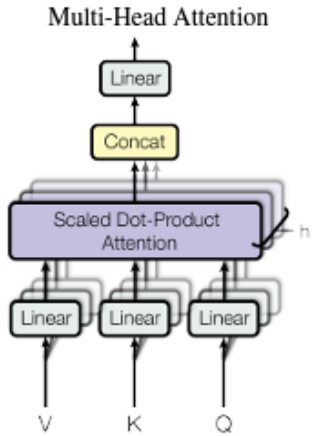
- O Encoder é composto por **N camadas**, cada uma com:
  - Um bloco de **self-attention**.
  - Uma rede **feed-forward** conectada.
- Cada bloco é rodeado por **Layer Normalization**.



Mecanismo de Self-Attention

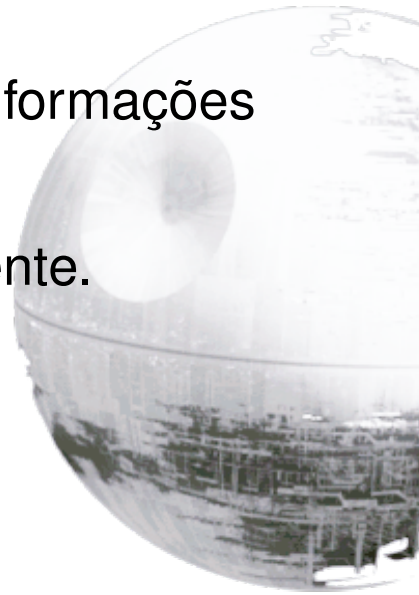
- **Self-attention** analisa o relacionamento entre todas as palavras da sentença.
- Calcula a **pontuação de atenção** para cada palavra em relação às demais.
- Usa **Keys, Queries, Values** para operar.





Multi-head Attention

- O **multi-head attention** permite ao modelo capturar informações de várias representações.
- Cada cabeça de atenção trabalha de forma independente.
- As saídas são concatenadas e projetadas.



Codificação Posicional no Transformer

- O Transformer usa **codificação posicional** para capturar a ordem das palavras.
- Adicionada aos embeddings das palavras.
- Funções seno e cosseno são usadas para gerar as codificações.

Codificação Posicional

- Como o Transformer processa entradas em paralelo, é necessário codificar a posição das palavras.
- A codificação posicional é somada aos embeddings das palavras para fornecer essa informação ao modelo.

Modelos Baseados em Transformers

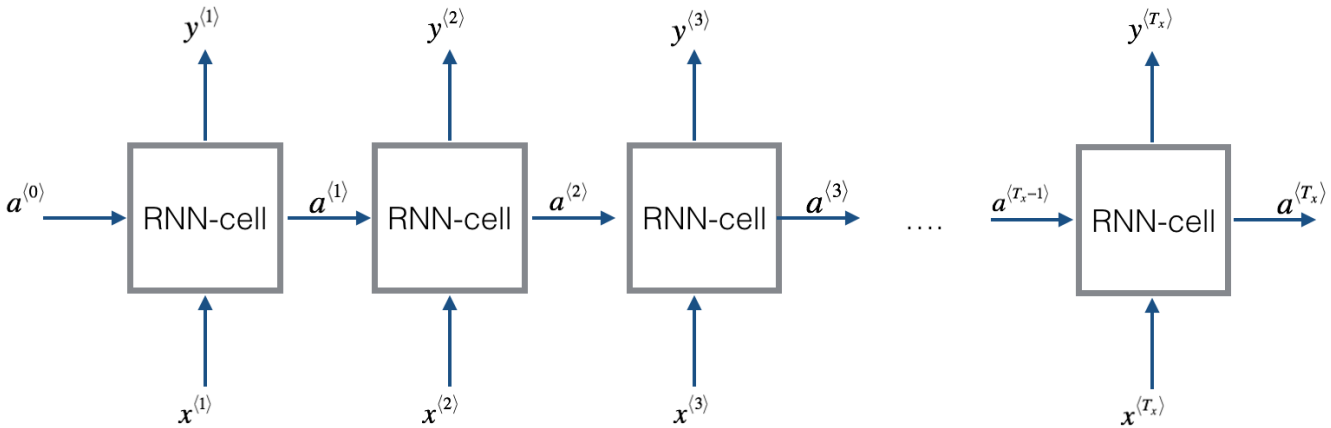
- BERT: Bidirectional Encoder Representations from Transformers.
- GPT: Generative Pre-trained Transformer.
- Outras variantes: T5, RoBERTa, XLNet.

Treinamento de Transformers

- Pré-treinamento em grandes volumes de dados.
- Fine-tuning para tarefas específicas, como classificação ou tradução.

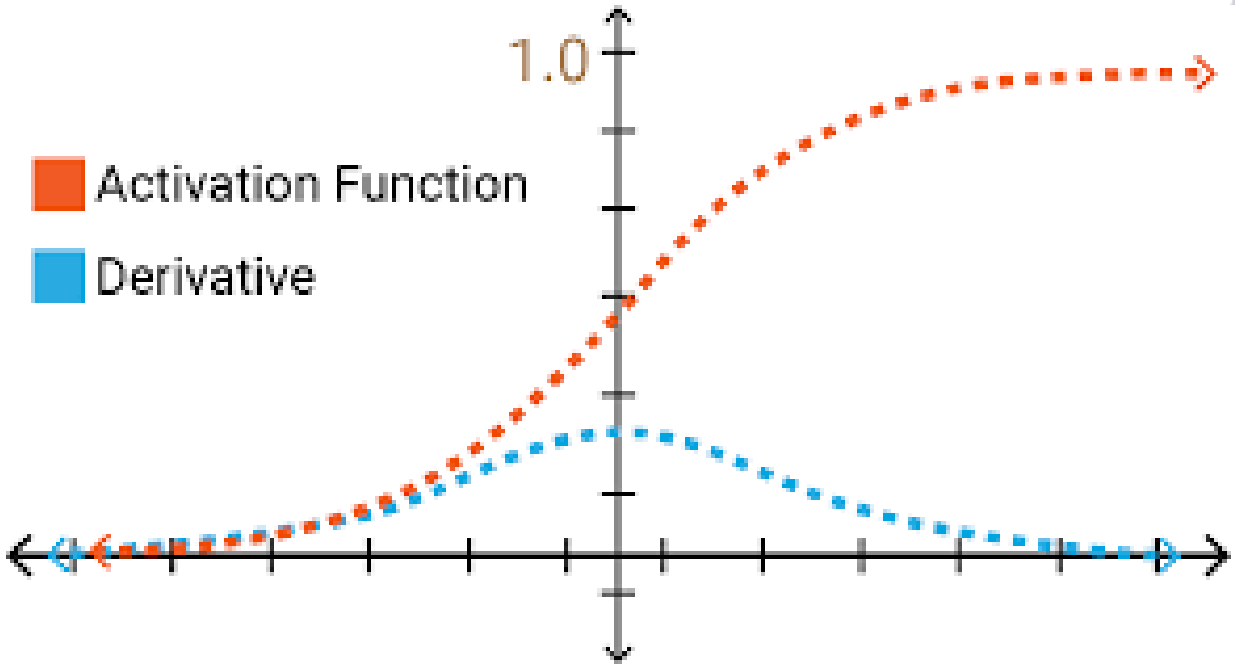
Introdução às RNNs e LSTMs

- **RNNs** processam sequências de dados de forma sequencial.
- **LSTMs** são uma melhoria das RNNs, com "células de memória"para capturar dependências longas.
- Aplicações: tradução automática, reconhecimento de fala.



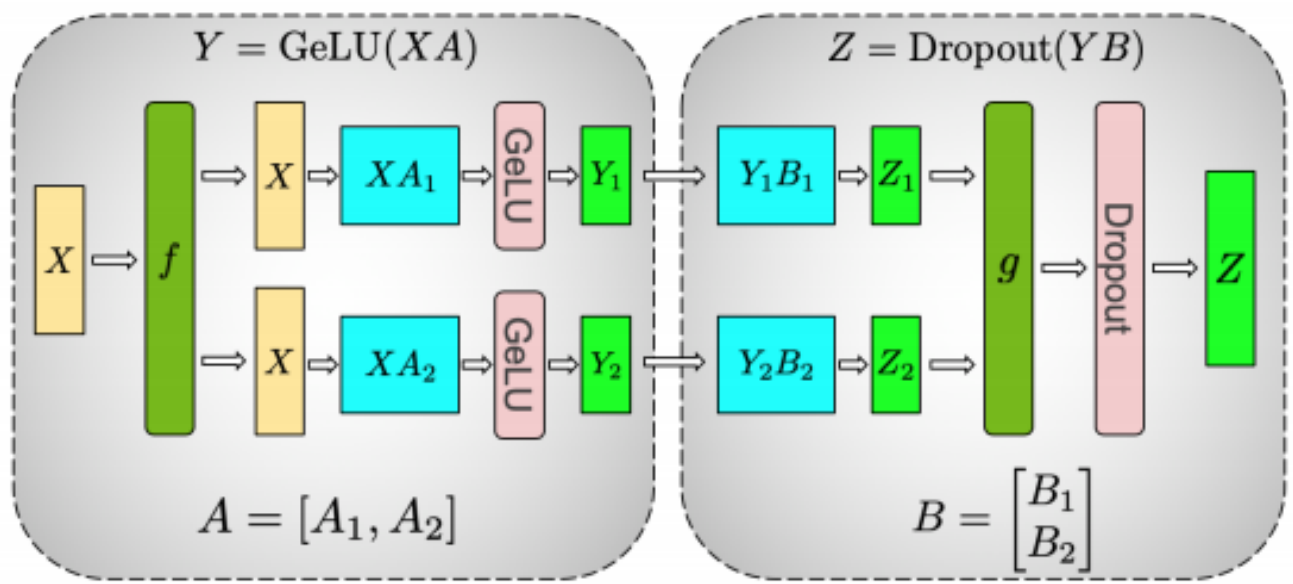
Problemas com RNNs/LSTMs

- **Dependência Sequencial**: Processam dados uma palavra de cada vez.
- **Desvanecimento de Gradientes**: Gradientes diminuem exponencialmente em longas sequências.
- **Contexto Local Limitado**: Difícil capturar dependências de longo prazo.



Introdução aos Transformers

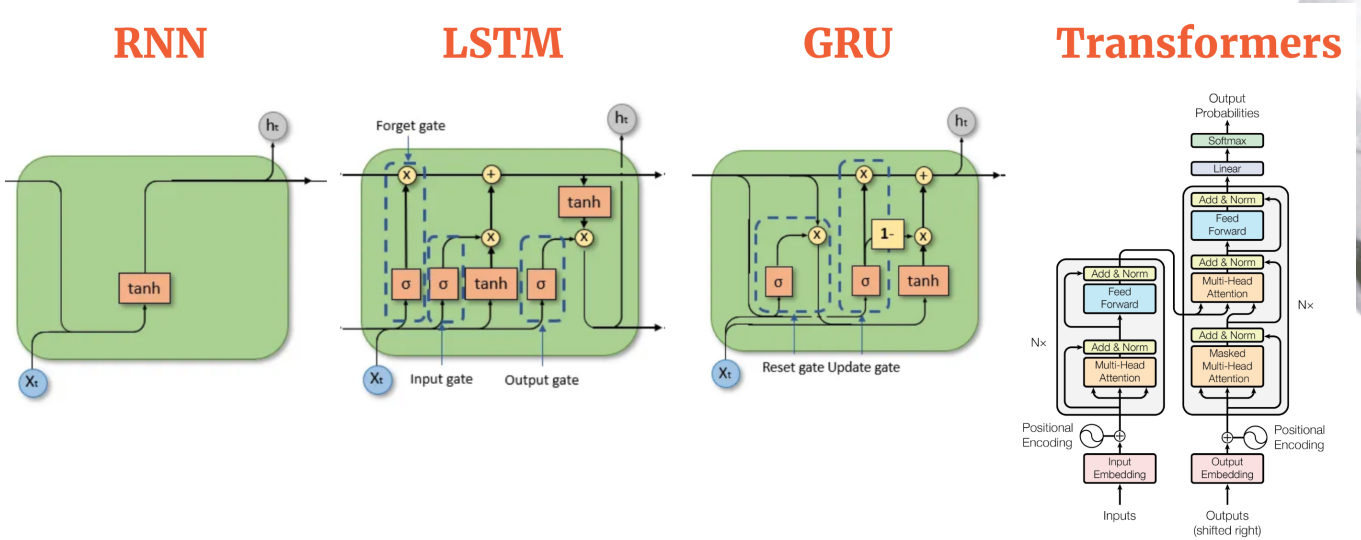
- **Transformers** resolvem a dependência sequencial e capturam dependências longas.
- Utilizam o mecanismo de **self-attention** para processar todas as palavras de forma simultânea.
- Processo paralelo, diferentemente das RNNs.



(a) MLP

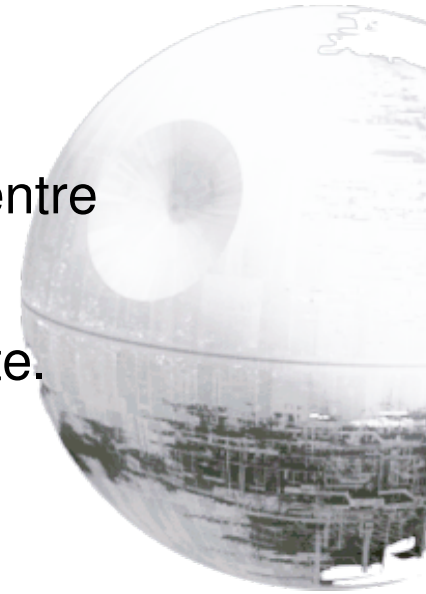
Diferenças Estruturais: RNNs/LSTMs vs. Transformers

- **RNNs/LSTMs**: Processamento sequencial, dependem da ordem das palavras.
- **Transformers**: Processamento paralelo, capturam relações globais através de self-attention.



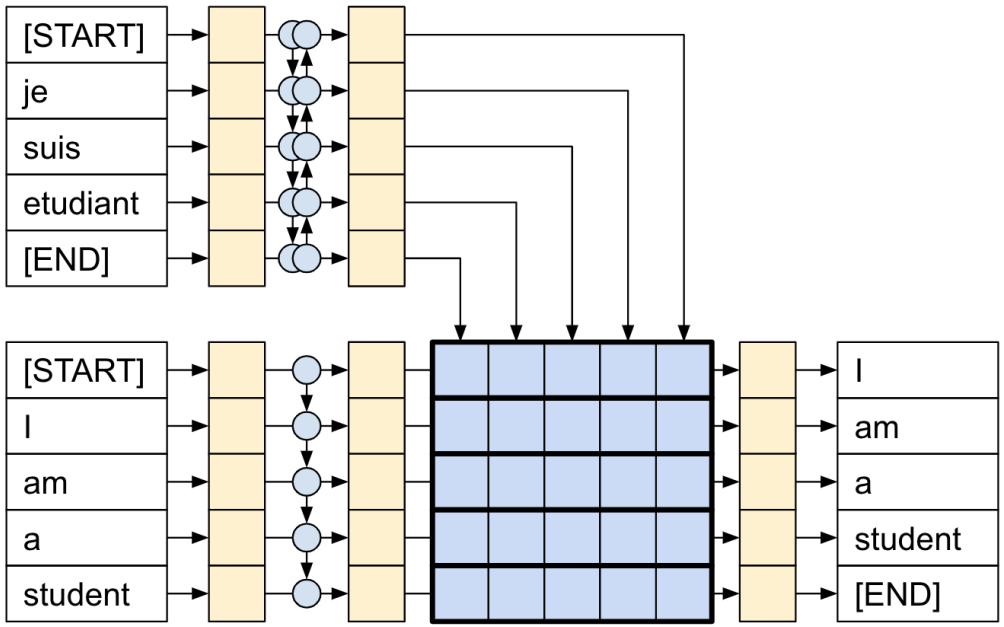
Vantagens dos Transformers

- **Eficiência Computacional**: Processamento paralelo.
- **Dependências Longas**: Capturam relacionamentos entre palavras distantes.
- **Escalabilidade**: Transformers escalam mais facilmente.



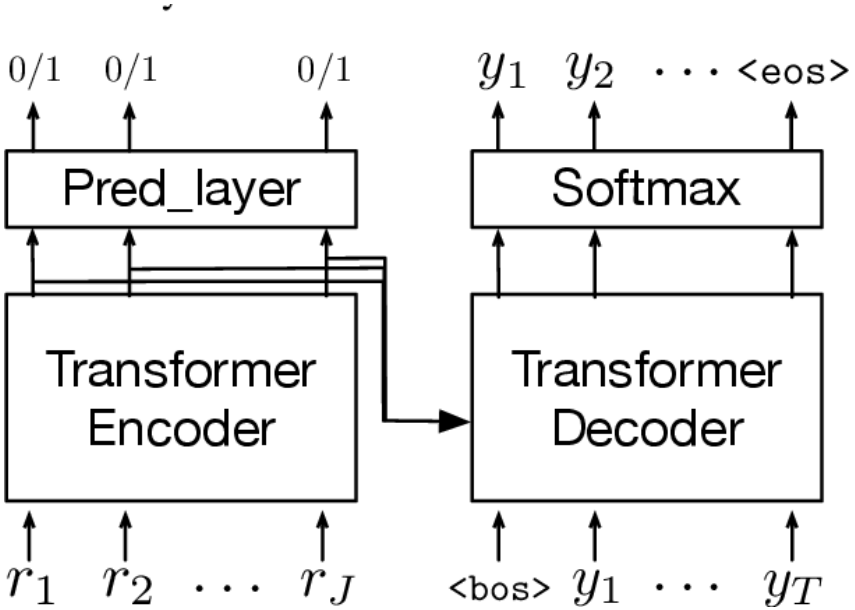
Aplicação de Transformers: Tradução Automática

- **Tradução automática** foi uma das primeiras grandes aplicações dos Transformers.
- Modelos como **Transformer** e **MarianMT** superam os modelos RNN em precisão.
- Aplicações: Google Translate, DeepL.



Aplicação de Transformers: Geração de Texto

- Modelos como **GPT** são eficazes em gerar texto a partir de prompts.
- Usado em chatbots, criação automática de conteúdo e assistentes de escrita.
- Exemplos: ChatGPT, assistentes virtuais.



Aplicação de Transformers: Classificação de Texto

- **BERT** se destaca na **classificação de texto**.
- Usado em análise de sentimentos, classificação de tópicos e categorização de e-mails.
- Exemplo: Análise de sentimentos em redes sociais.

Aplicação de Transformers: Resposta a Perguntas

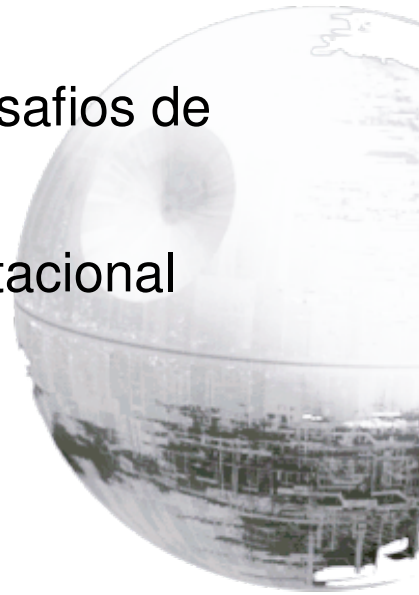
- Modelos como **BERT** e **T5** são usados para **resposta a perguntas**.
- Útil para FAQ automatizado, motores de busca, assistentes virtuais.
- Exemplo: Ferramentas de FAQ automáticas.

Aplicação de Transformers: Sumarização de Texto

- Modelos como **T5** e **BART** criam resumos curtos e coerentes de textos longos.
- Usado para resumir artigos científicos, relatórios e notícias.
- Exemplo: Ferramentas de resumo de artigos.

Transformers em Produção

- Colocar Transformers em produção exige lidar com desafios de escalabilidade.
- Questões de latência, uso de memória e custo computacional precisam ser resolvidas.

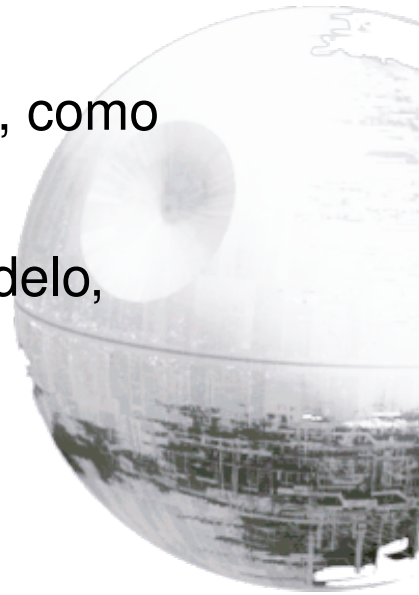


Desafios de Escalar Modelos de Transformers

- Modelos grandes, como GPT-3, exigem muitos recursos para treinamento e inferência.
- Latência é crítica em aplicações em tempo real.
- O custo computacional e de hardware é um grande desafio para implantações em larga escala.

Otimização de Transformers: Técnicas Comuns

- **Distilação**: Criação de versões menores dos modelos, como **distilBERT**.
- **Quantização**: Redução da precisão dos pesos do modelo, melhorando a eficiência.

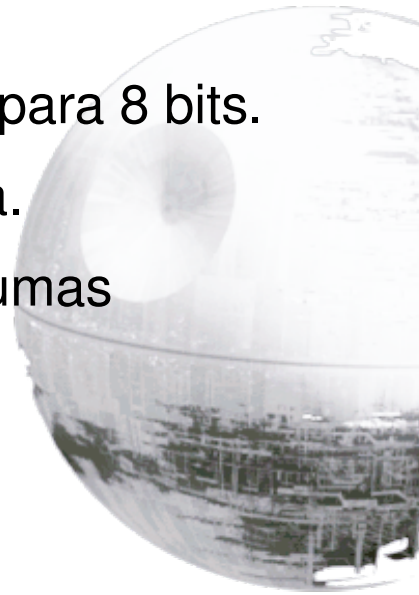


Redução de Tamanho com distilBERT

- distilBERT reduz os parâmetros em 40
- Mantém uma precisão alta com apenas pequenas perdas.

Quantização de Modelos para Produção

- A **quantização** reduz a precisão dos pesos de 32 bits para 8 bits.
- Benefícios: menor uso de memória e latência reduzida.
- Desafios: Implementação e perda de precisão em algumas tarefas.



Desafios e Limitações

- Requer grandes quantidades de dados e recursos computacionais.
- Difícil de escalar para contextos muito longos.

Transformers no Futuro do NLP

- Modelos mais eficientes (DistilBERT, Longformer).
- Aplicações emergentes em diálogos e assistentes virtuais.

Conclusão

- Transformers transformaram o campo de NLP.
- O futuro promete avanços com modelos mais eficientes e aplicáveis a novos contextos.

MAY THE

SOURCE

BE WITH

YOU

[1] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.