NLP 101 (6): BERT vs GPT

Depois que os Transformers surgiram em 2017, a pergunta deixou de ser ”essa arquitetura funciona?” Ela passou a ser: Como devemos treiná-la? A resposta dividiu a pesquisa em…

2026-06-29



Figure 1: Capa do artigo

Depois que os Transformers surgiram em 2017, a pergunta deixou de ser “essa arquitetura funciona?” Ela passou a ser: Como devemos treiná-la? A resposta dividiu a pesquisa em dois caminhos que mudariam completamente a Inteligência Artificial. De um lado nasceu o BERT, do outro, o GPT. Embora utilizem a mesma arquitetura base, eles aprenderam a linguagem de formas completamente diferentes. Sou a professora Artemísia. Vem entender comigo.

Na thread anterior do 101, vimos que os Transformers conseguiam processar toda uma sequência simultaneamente usando Self-Attention. Mas uma arquitetura sozinha não aprende nada, era preciso definir uma tarefa para o treinamento. A primeira grande estratégia foi apresentada pelo BERT, em 2018. A ideia era esconder algumas palavras do texto e pedir que o modelo as descobrisse. Thread Imagine a frase: “O Sol nasce no [MASK].” O objetivo do modelo é prever qual palavra foi escondida.

Esse processo recebeu o nome de Masked Language Modeling. Ou MLM. Em vez de prever a próxima palavra, o modelo tenta reconstruir partes faltantes do texto.Para descobrir a palavra escondida, o BERT pode olhar tanto para a esquerda quanto para a direita. Ele utiliza todo o contexto disponível, por isso dizemos que o BERT é um modelo bidirecional. Ele aprende representações extremamente boas para compreender linguagem.

Arquiteturalmente, o BERT utiliza apenas a pilha de Encoders do Transformer.Como todos os tokens estão disponíveis simultaneamente, cada palavra pode prestar atenção em qualquer outra palavra da frase. Isso produz representações extremamente ricas para compreender linguagem, e por isso o BERT se destacou em tarefas como: • Classificação de textos • Análise de sentimentos Reconhecimento de entidades (NER) Perguntas e respostas Busca semântica Seu objetivo é entender o significado do texto.

© Artemisia Weyl 2024–2026 — Todos os direitos reservados.



Figure 2: Imagem original inserida na publicação 1/9

Mas outra linha de pesquisa fez uma pergunta diferente, e se treinássemos o modelo exatamente como fazemos quando escrevemos? Uma palavra de cada vez, sempre tentando prever o próximo token. Foi exatamente esse caminho que o GPT seguiu, ao invésvez de esconder palavras, ele recebe uma sequência incompleta e precisa prever qual token vem em seguida.

Considere a frase: "A Terra gira em torno do…" O modelo precisa prever "Sol". Depois recebe esse novo token e faz outra previsão. Esse processo é repetido bilhões de vezes durante o pré-treinamento Esse método é chamado de Next Token Prediction. Existe uma consequência importante, para prever corretamente o próximo token, o modelo não pode enxergar palavras que ainda não foram escritas.

Por isso o GPT utiliza uma Masked Self-Attention. Durante o cálculo da atenção, uma máscara impede que cada token consulte posições futuras da sequência. Em outras palavras, o modelo literalmente é impedido de "colar a resposta". Arquiteturalmente, o GPT utiliza apenas a pilha de Decoders do Transformer. Essa restrição faz com que cada previsão dependa exclusivamente dos tokens anteriores. Por isso dizemos que ele é um modelo autoregressivo.

Essa pequena diferença muda completamente o comportamento dos modelos. O BERT aprende excelentes representações para compreender linguagem. O GPT aprende excelentes representações para gerar linguagem. Curiosamente, ambos nasceram exatamente da mesma arquitetura Transformer. O que mudou foi a tarefa de treinamento e a parte da arquitetura utilizada. E essa escolha mudou muita coisa.

Mas a história não termina aí. Treinar um modelo para prever palavras ainda não era suficiente para transformá-lo em um bom assistente. Faltava ensinar esses modelos a seguir instruções, conversar e responder de forma útil. E isso nos leva ao próximo capítulo: Como nasce um LLM moderno. Gostou? Comenta.

© Artemisia Weyl 2024–2026 — Todos os direitos reservados.



Figure 3: Imagem original inserida na publicação 9/9

© Artemisia Weyl 2024–2026 — Todos os direitos reservados.