Pós-Graduação Lato Sensu
Curso de Especialização em Inteligência Artificial

Introdução a Inteligência Artificial

Prof. Dr. **Lucas Dias Hiera Sampaio**

PARTE 06

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS CORNÉLIO PROCÓPIO

Inteligência Artificial em Cenários Reais

1. Apresentação

Durante as cinco primeiras semanas da disciplina abordamos diferentes aspectos relacionados à inteligência artificial - da definição mais ampla até o uso de bibliotecas e frameworks mais comuns. Nesta semana nosso objetivo é conhecer diferentes aplicações de inteligência artificial em produtos disponíveis hoje no mercado de tecnologia e relacionar estas soluções com a taxonomia de IA.

2. ChatGPT

Lançado em 30 de novembro de 2022 o ChatGPT, acrônimo de *Chat Generative Pre-Trained Transformer*, é um *chatbot* baseado em um modelo largo de linguagem (MLL) desenvolvido pela empresa OpenAI. A plataforma permite que usuários interajam por meio de texto com uma inteligência artificial capaz de responder perguntas, gerar texto, etc.

Em janeiro de 2023 de acordo com a Reuters o chatGPT tornou-se a tecnologia com o crescimento mais rápido da base de usuários da história: em menos de 2 meses a plataforma atingiu a marca de 100 milhões de usuários ativos por mês marca que demandou 9 meses da plataforma TikTok e 2 anos e meio do Instagram para ser atingida. Essa marca histórica contribuiu para que a avaliação da criadora do chatGPT, a OpenAI, superasse os 29 bilhões de dólares.

É importante compreender a diferença entre as ferramentas tecnológicas discutidas: enquanto o chatGPT é uma aplicação web na forma de *chatbot, o modelo* de inteligência artificial que é utilizado para prover o serviço da aplicação web é denominado GPT. A primeira versão deste modelo o GPT-1 foi lançado em 11 de junho de 2018, possuía uma estrutura com 117 milhões de hiperparâmetros, foi treinado com o conjunto de dados BookCorpus - 4,5GB de texto - e o treinamento da rede levou 30 dias utilizando 8 GPUS do modelo P600 que são capazes de processar aproximadamente 1 PetaFLOP/s (FLOP, Floating Point Operation ou Operações com ponto flutuante são uma medida de quantas operações matemáticas os processadores são capazes de realizar por segundo - Peta é um sufixo que indica 10 elevado a décima segunda potência ou 1 seguido de doze zeros).

O GPT-2 aumentou o número de hiperparâmetros para 1,5 bilhões e foi treinado utilizando 40GB de texto e foi lançado em 14 de fevereiro de 2019. O treinamento teve um custo estimado de 1,5 zettaFLOPs (15 seguido de 20 zeros).

O GPT-3 lançado em maio de 2020 foi o modelo utilizado inicialmente no chatGPT em seu lançamento. Este modelo elevou o número de hiperparâmetros para 175 bilhões e o conjunto de dados para treinamento foi composto pelo CommonCrawl (aproximadamente 570GB), WebText, Wikipedia e dois livros cujo conteúdo é ofuscado. O treinamento do modelo usou um total de 310 zettaFLOPs. O GPT-3 possui uma publicação aberta no arxiv.org, um artigo de mais de 60 páginas assinado pelos pesquisadores empregados pela OpenAI explicando o funcionamento dos modelos de linguagem[1].

No trabalho dos pesquisadores os mesmos avaliam a acurácia dos modelos com diferentes dimensões (quantidade de hiperparâmetros) como pode ser observado na Figura 1.

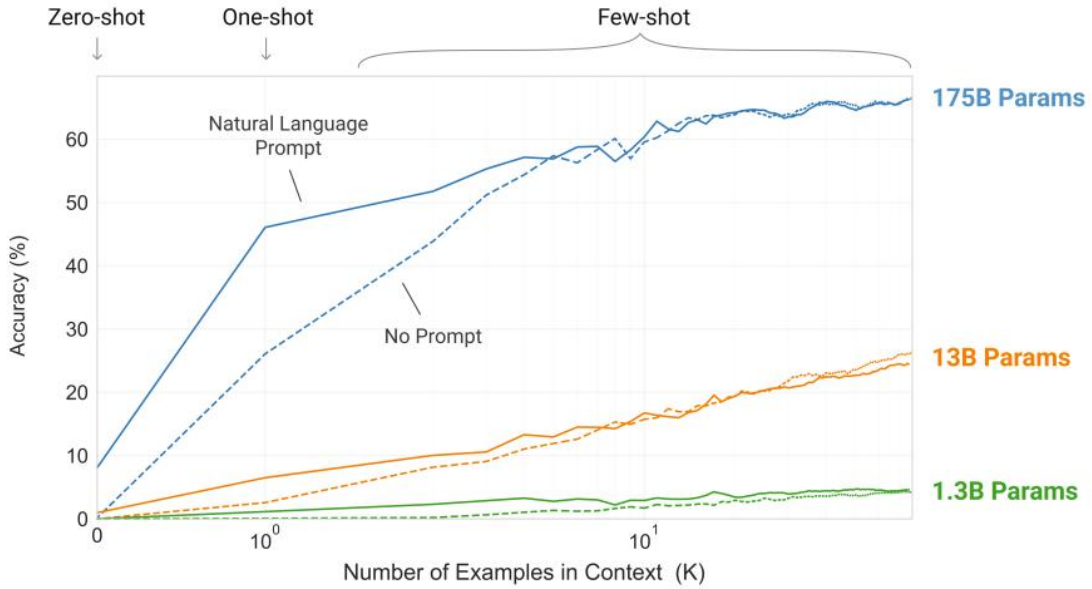

Figura 1: Acurácia dos diferentes modelos largos de linguagem (MLL) utilizando diferentes quantidades de hiperparâmetros.

É possível observar na Figura 1 que o número de hiperparâmetros (1,3 bilhões, 13 bilhões e 175 bilhões) e portanto, a dimensão do modelo, tem impacto significativo na acurácia. Enquanto os dois modelos menores não chegam a 30% de acurácia após alguns exemplos o modelo maior com 175 bilhões de hiperparâmetros atinge quase 70% de acurácia.

Embora a dimensão do modelo de inteligência artificial tenha impacto direto no sucesso da ferramenta, a tecnologia GPT não é uma criação antiga: os geradores pré-treinados (*generative pre-trained*) são modelos de *deep learning* desenvolvidos nos anos 2012. Já os transformadores (*transformers*) são uma topologia de *deep learning* proposta por pesquisadores do Google em 2017 para substituir redes neurais

[1] Language Models are Few-Shot Learners

artificiais complexas recorrentes e redes neurais convolucionais no processamento de linguagem natural[2].

Os transformadores são compostos por:

- **Tokenizers**: responsáveis por converter texto em tokens.
- **Camada de Embutir**: responsável por converter os tokens e as posições dos tokens em representação vetorial.
- **Camadas de Transformação**: executam repetidas transformações nas representações vetoriais extraindo cada vez mais informação linguística. São compostas por duas subcamadas - subcamada de atenção e subcamada feedforward.
- **Camada de Desembutir** (opcional): responsável por converter as representações vetoriais finais em distribuição de probabilidade nos diferentes tokens.

A Figura 2 apresenta a organização de um transformador como apresentado no artigo original do Google: os elementos cinza são as camadas de transformação com as subcamadas de atenção (alaranjado) e feedforward (azul). A camada de embutir está representada na cor rosa e a camada de desembutir é omitida no modelo.

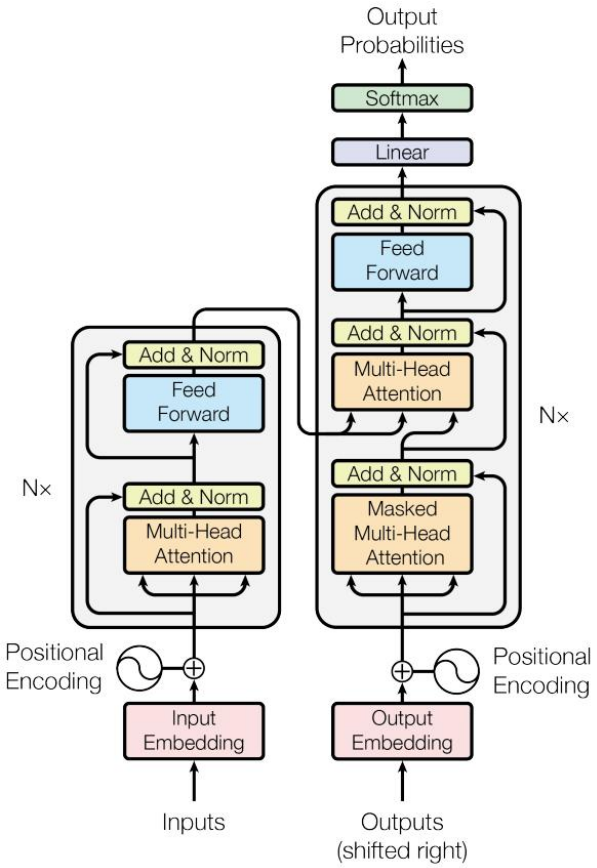

Figura 2: Arquitetura de referência para transformadores.

[2] Attention Is All You Need

3. Assistentes Pessoais

Siri, Alexa, Google Assistant e Cortana são assistentes pessoais de quatro gigantes de tecnologias: Apple, Amazon, Google e Microsoft. A função destas soluções é simples: receber comandos por voz, executar comandos e providenciar um *feedback* por meio de áudio para o usuário.

Embora seja simples, existem diferentes desafios tecnológicos na criação de um assistente pessoal virtual:

- Reconhecimento de fala: o software deve ser capaz de receber o sinal captado pelo microfone e transformar o mesmo em texto,
- Processamento dos comandos: o assistente deve com base nos comandos recebidos decidir quais ações executar,
- Gerar respostas: uma vez que o comando tenha sido executado é necessário dar uma resposta ao usuário por meio de duas etapas - tomar a decisão de qual resposta deve ser dada e gerar o áudio a ser reproduzido.

Fica evidente que múltiplas tarefas de diferentes paradigmas são utilizadas nestes produtos. É evidente que não é possível determinar com o mesmo nível de exatidão do GPT como é de fato o funcionamento dos assistentes citados uma vez que as aplicações são executadas em nuvem, isto é, nenhum dado, decisão ou geração é realizada no dispositivo que você está utilizando, pelo contrário, o dispositivo fica responsável unicamente por captar as informações (inputs) enviar a nuvem, receber a resposta e reproduzir os resultados.

4. IBM Watson

Em 2011 a IBM desenvolveu um sistema computacional destinado a responder perguntas com base no programa de TV norte-americano Jeorpardy! onde os participantes respondem perguntas de diferentes assuntos no formato de quiz. Para o projeto a IBM desenvolveu um *hardware* cuja arquitetura era composta por 2880 threads de processadores POWER7 utilizando 16 TB de memória RAM e capaz de processar 80 TeraFLOPs por segundo.

O *software* desenvolvido pela IBM abrangia diferentes linguagens como Java, C++ e Prolog e era composto por dois módulos principais: uma parte responsável por responder as perguntas e um framework Apache responsável pela interface de entrada e saída.

A solução da IBM foi utilizada contra outras duas IAs no início de 2011 numa disputa de Jeopardy! e foi vitoriosa nas 3 partidas que jogou. Após a competição o IBM Watson continuou sendo desenvolvido e expandido para outras aplicações como por

exemplo: tomada de decisão de diagnóstico clínico principalmente em oncologia e radiologia, chatbot, criação de código para programas, assistente para professores, previsão do tempo, preparação de imposto de renda e propaganda.

Em maio de 2023, a IBM lançou o Watsonx uma IA generativa e plataforma de dados em nuvem que oferece diferentes ferramentas de IA como serviço. Embora não se tenha acesso ao código fonte ou uma publicação explicando a arquitetura da solução a IBM informa que diferentes MLLs estão disponíveis na plataforma incluindo o GPT.

5. Na próxima semana.

Na próxima semana falaremos sobre aplicações de IA na pesquisa. Vamos analisar como a inteligência artificial pode auxiliar pesquisadores em diferentes áreas e perceber a diferença no nível de detalhamento quando comparamos as soluções comerciais existentes com os resultados de pesquisa de IA aplicada.