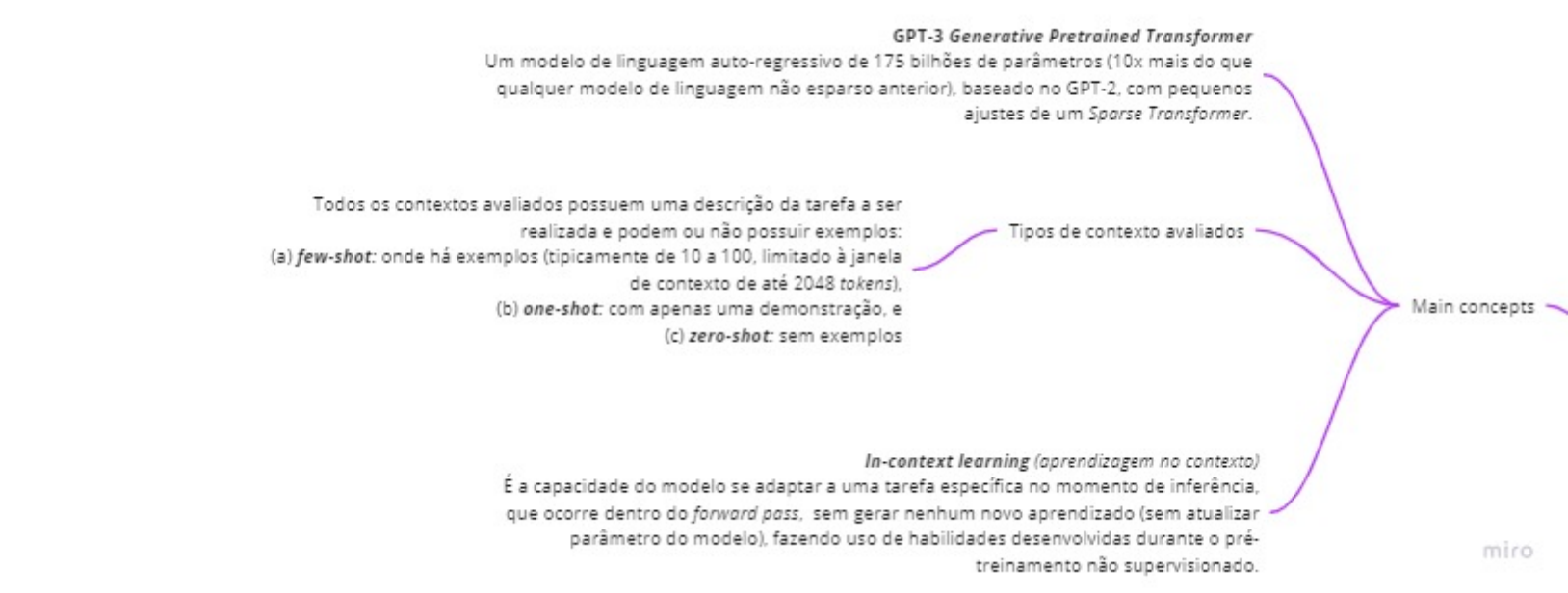
Article Presentation
IA368DD_2023S1: Deep Learning aplicado a Sistemas de Buscas
Student: Marcus Vinícius Borela de Castro

Language Models are Few-Shot Learners

By: Tom B. Brown *et al.* (OpenAi). 2020


GPT-3 *Generative Pretrained Transformer*
Um modelo de linguagem auto-regressivo de 175 bilhões de parâmetros (10x mais do que qualquer modelo de linguagem não esparso anterior), baseado no GPT-2, com pequenos ajustes de um *Sparse Transformer*.
Todos os contextos avaliados possuem uma descrição da tarefa a ser realizada e podem ou não possuir exemplos:
(a) ***few-shot:*** onde há exemplos (tipicamente de 10 a 100, limitado à janela de contexto de até 2048 *tokens*),
(b) ***one-shot:*** com apenas uma demonstração, e
(c) ***zero-shot:*** sem exemplos
Tipos de contexto avaliados
Main concepts
In-context learning *(aprendizagem no contexto)*
É a capacidade do modelo se adaptar a uma tarefa específica no momento de inferência, que ocorre dentro do *forward pass*, sem gerar nenhum novo aprendizado (sem atualizar parâmetro do modelo), fazendo uso de habilidades desenvolvidas durante o pré-treinamento não supervisionado.
miro

Com o GPT-3 se tenta repetir capacidade humana de poder realizar uma nova tarefa de linguagem a partir de apenas alguns exemplos ou de instruções simples.

A diferença entre o desempenho de zero, um e poucos exemplos geralmente cresce com a capacidade do modelo

A importância do pré-treinamento em larga escala para o desempenho do modelo GPT-3 (by ChatGPT)

Modelos maiores viabilizam "in-context-aprendizes"

Article contribution

Apresenta uma análise detalhada das habilidades e limitações do modelo em diferentes tarefas de linguagem natural. (by ChatGPT)

Basic doubts that may arise

Foi dito que modelos maiores podem usar *batch size* maior, mas com *learning rate* menor

Por quê um LR menor?

O LR é menor por causa do batch size maior?

O artigo diz que fine-tunning no GPT-3 pode ser uma direção promissora para trabalhos futuros. Os modelos LLM open ai podem ser alvo de fine-tunning.

Quantos gb ocupa o GPT-3?

Como são armazenados os LLM pré-treinados por um cliente?

Completamente ou só parcialmente (diferenças do base model)?

Está no texto: "Tarefas sintéticas, como embaralhar palavras ou definir palavras sem sentido, parecem ser especialmente prováveis de serem aprendidas de novo, enquanto a tradução claramente deve ser aprendida durante o pré-treinamento, embora possivelmente de dados que são muito diferentes em organização e estilo dos dados de teste. Em última análise, nem mesmo está claro o que os humanos aprendem do zero em relação a demonstrações anteriores."

O GPT-3 é pré-treinado em uma grande quantidade de dados de texto não rotulados, o que lhe permite aprender a reconhecer e entender padrões linguísticos e semânticos. Esse pré-treinamento em larga escala permite que o modelo capture o conhecimento geral da linguagem e crie representações semânticas complexas que podem ser transferidas para novas tarefas. (by ChatGPT)

O aprendizado de contexto realmente ocorre na inferência ou simplesmente se identificam tarefas aprendidas durante o treinamento?

Como se dá essa redução de transferência de dados?

Para minimizar o tempo de treinamento, utilizou-se processamento paralelo entre GPUs com a redução da transferência de dados entre nós tanto na profundidade quanto na largura da arquitetura.

Embora o GPT-3 3B seja quase 10 vezes maior que o RoBERTa-Large (parâmetros de 355M), ambos evaram aproximadamente 50 petaflop/s-dias de computação durante o pré-treinamento

O que é um "s-day"? Por que precisa desse denominador?

Interesting/unexpected results

Habilidades de geração de texto: Na geração de artigos, dados título e subtítulo, a detecção humana, se é ou não real, alcançou 52%, quase o acaso.

Em geral, em tarefas de NLP, o GPT-3 alcançou resultados promissores nas configurações *zero-shot* e *one-shot*, e na *few-shot* às vezes é competitivo ou até supera o estado da arte (superando modelos com ajuste fino).

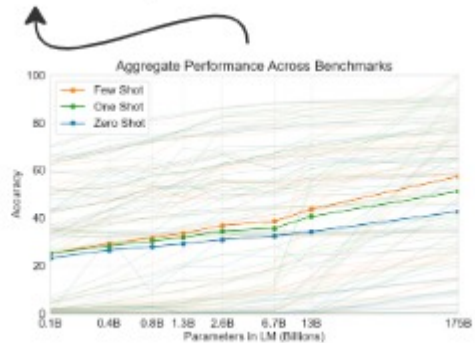
Por exemplo, GPT-3 atinge 64,3% de precisão no TriviaQA no *zero-shot*, 68,0% no *one-shot* (igual SOTA anterior) e 71,2% no *few-shot*, este SOTA

Demonstra que as habilidades de aprendizagem em contexto demonstram ganhos fortes com a escala

Para se estudar a correlação do desempenho com o tamanho do modelo, foram treinadas 8 configurações diferentes, variando em três ordens de grandeza, de 125 milhões de parâmetros a 175 bilhões de parâmetros, sendo o maior modelo referenciado como GPT-3

Todos os modelos foram treinados para um total de 300 bilhões de tokens

Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks



O GPT-3 apresenta um desempenho pior em algumas tarefas

Tarefas que se beneficiam empiricamente da bidirecionalidade (como em tarefas de preenchimento de lacunas entre textos: fill mask).

Tarefas sintéticas e qualitativas que exigem correspondência de padrões e computação não triviais e que são improváveis nos dados de treinamento.

Tarefas da "física de senso comum", com perguntas do tipo "Se eu colocar queijo na geladeira, ele derreterá?"

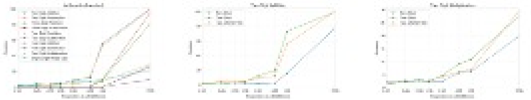
Obs.:(Isso mudou: perguntei 17/3/2023 ao WebChatGPT).

A forma como resumiu em uma tabela os resultados

H Results on All Tasks for All Model Sizes

					Zero-Shot								One-Shot								Few-Shot								
Name	Metric	Split	Fine-tune SOTA	K	Small	Med	Large	XL	2.7B	6.7B	13B	175B	Small	Med	Large	XL	2.7B	6.7B	13B	175B	Small	Med	Large	XL	2.7B	6.7B	13B	175B	175B (test server)
HellaSwag	acc	dev	85.6	20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
LAMBADA	acc	test	68.0	15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
LAMBADA	ppl	test	8.63	15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
StoryCloze	acc	test	91.8	70	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
NQs	acc	test	44.5	64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
TriviaQA	acc	dev	68.0	64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	71.2
WebQs	acc	test	45.5	64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	

E ainda tem os gráficos para cada linha da tabela



Advanced topic to discuss

- Parece ser promissor repetir em modelo bidirecional a escala do GPT-3 com meta-aprendizagem (falta de bidirecionalidade explicaria piores resultados em alguns testes)
  - Por quê não se promove esse teste?
  - Tem algum LLM (Large Language Model) bidirecional?
- Foram identificados vieses no comportamento dos modelos, que podem gerar conteúdo estereotipado ou preconceituoso, consequência dos vieses presentes nos dados de treinamento, por exemplo, quanto a sexo, raça e religião.
  - Como evitar/mitigar esses vieses?

Em algumas tarefas o problema era, provavelmente, o tokenizador.

Como em Palavras invertidas (RW) – O modelo recebe uma palavra escrita ao contrário e deve produzir a palavra original. Exemplo: stcejbo!objects.

Mas hoje o ChatGPT já resolve essa questão.

reverse this word: sucram
The reverse of the word "sucram" is "marcus".

Como ele faz isso? Ele tokeniza letra a letra? Ou ele experimenta diferentes tokenizadores durante a inferência?

Um tópico avançado que pode surgir da leitura do artigo é a discussão sobre a capacidade dos modelos de linguagem, como o GPT-3, de aprender conceitos mais abstratos e de alto nível. O artigo sugere que o modelo pode aprender conceitos mais abstratos, como a matemática avançada, programação de computadores e até mesmo habilidades sociais e emocionais. (by ChatGPT)