

INF721

2023/2



Aprendizado em Redes Neurais Profundas

A22: Estudo de casos de Transformers

Logística

Avisos

- ▶ Entrega e apresentação do Projeto Final nessa segunda-feira (04/12)!
- ▶ Entrega (pós-graduação): slides, código e relatório
- ▶ Entrega (graduação): slides e código + relatório (ponto extra)

Última aula

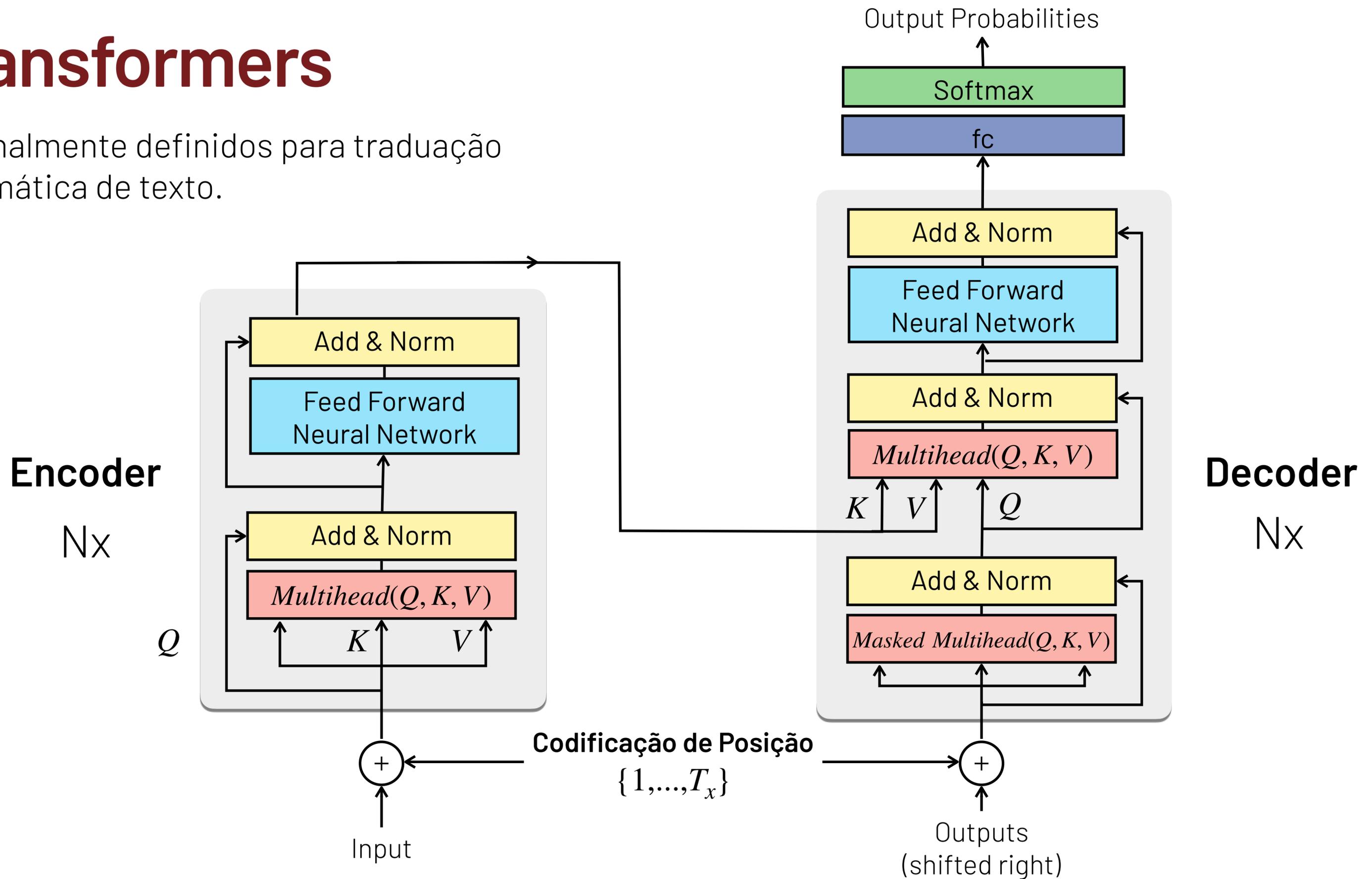
- ▶ GPT

Plano de Aula

- ▶ General Pre-Training (GPT)
- ▶ Bidirectional Encoder Representations from Transformers (BERT)
- ▶ Pré-treinamento (*Pre-training*)
- ▶ Transferência de Aprendizagem com Ajuste Fino

Transformers

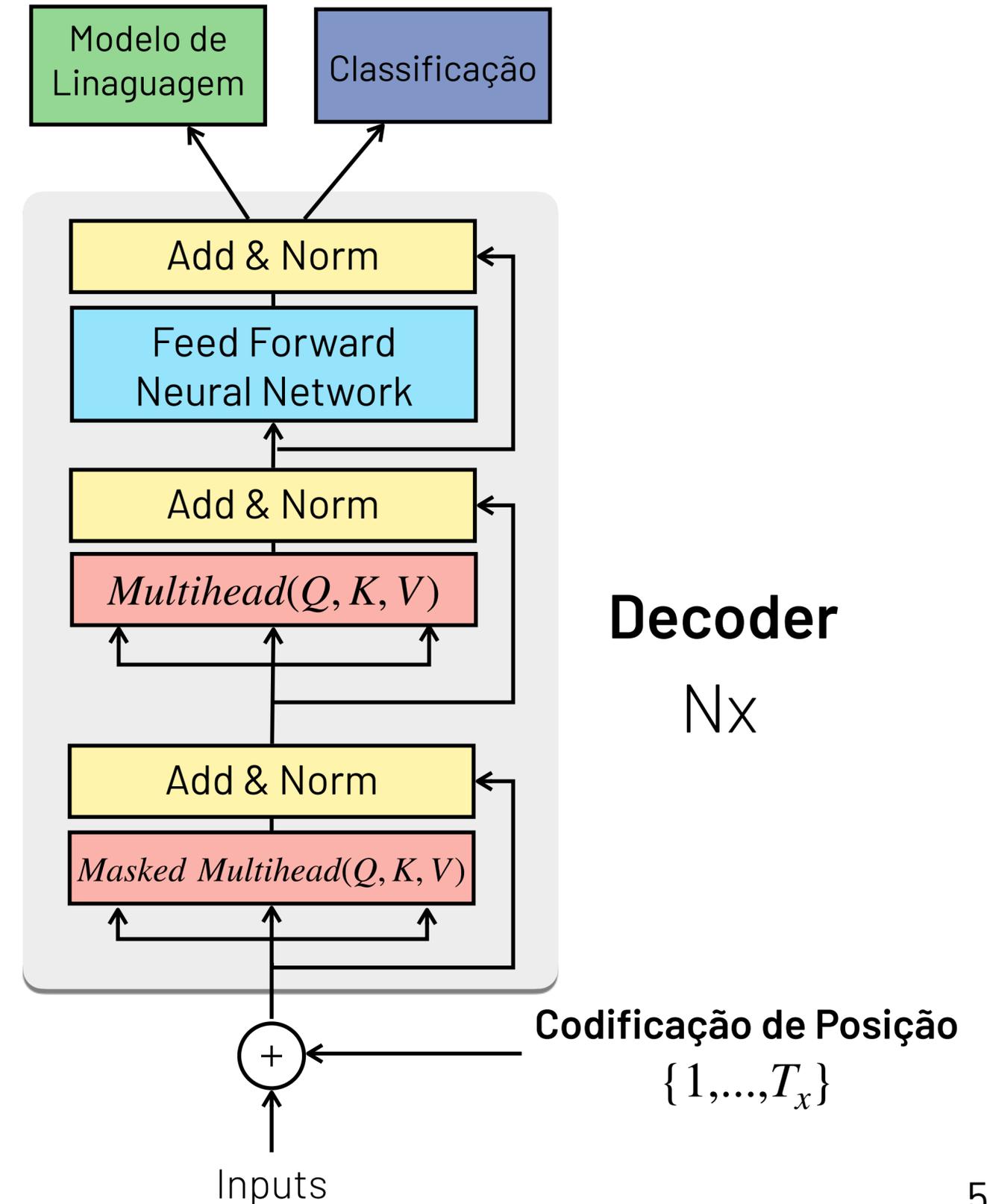
Originalmente definidos para tradução automática de texto.



General Pre-Training (GPT)

Utiliza apenas o Decoder do Transformer

1. **Pré-treinar** um modelo de linguagem em textos não-rotulados
2. **Ajuste-fino** para classificação
 - ▶ Mantém os pesos da rede após o pré-treinamento
 - ▶ Substituindo a camada do modelo de linguagem por uma camada de classificação e
 - ▶ Treinamento com uma conjunto de dados rotulado



GPT-1 (2018)

- ▶ **Número de blocos decoder:** 12
- ▶ **Número de parâmetros:** 117 milhões
- ▶ **Pré-treinamento:** BooksCorpus dataset
 - ▶ 7,000 livros não publicados de diferentes gêneros
- ▶ **Ajuste-fino**
 - ▶ Natural Language Inference
 - ▶ Question answering
 - ▶ Commonsense Reasoning
 - ▶ Semantic Similarity
 - ▶ Classification

Improving Language Understanding by Generative Pre-Training

Alec Radford Karthik Narasimhan Tim Salimans Ilya Sutskever
OpenAI OpenAI OpenAI OpenAI
alec@openai.com karthikn@openai.com tim@openai.com ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

1 Introduction

The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Most deep learning methods require substantial amounts of manually labeled data, which restricts their applicability in many domains that suffer from a dearth of annotated resources [61]. In these situations, models that can leverage linguistic information from unlabeled data provide a valuable alternative to gathering more annotation, which can be time-consuming and expensive. Further, even in cases where considerable supervision is available, learning good representations in an unsupervised fashion can provide a significant performance boost. The most compelling evidence for this so far has been the extensive use of pre-trained word embeddings [10, 39, 42] to improve performance on a range of NLP tasks [8, 11, 26, 45].

Leveraging more than word-level information from unlabeled text, however, is challenging for two main reasons. First, it is unclear what type of optimization objectives are most effective at learning text representations that are useful for transfer. Recent research has looked at various objectives such as language modeling [44], machine translation [38], and discourse coherence [22], with each method outperforming the others on different tasks.¹ Second, there is no consensus on the most effective way to transfer these learned representations to the target task. Existing techniques involve a combination of making task-specific changes to the model architecture [43, 44], using intricate learning schemes [21] and adding auxiliary learning objectives [50]. These uncertainties have made it difficult to develop effective semi-supervised learning approaches for language processing.

¹<https://gluebenchmark.com/leaderboard>

GPT-2 (2019)

- ▶ **Número de blocos decoder:** 48
- ▶ **Número de parâmetros:** 1542 milhões (1.5 bilhões)
- ▶ **Pré-treinamento:** WebText dataset
 - ▶ 45 milhões de páginas web selecionadas pela OpenAI de acordo com uma métrica de qualidade
- ▶ **Ajuste-fino**
 - ▶ Reading Comprehension
 - ▶ Summarization
 - ▶ Translation
 - ▶ Question Answering
 - ▶ Métricas de modelos de linguagem

Language Models are Unsupervised Multitask Learners

Alec Radford^{*1} Jeffrey Wu^{*1} Rewon Child¹ David Luan¹ Dario Amodei^{**1} Ilya Sutskever^{**1}

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

1. Introduction

Machine learning systems now excel (in expectation) at tasks they are trained for by using a combination of large datasets, high-capacity models, and supervised learning (Krizhevsky et al., 2012) (Sutskever et al., 2014) (Amodei et al., 2016). Yet these systems are brittle and sensitive to slight changes in the data distribution (Recht et al., 2018) and task specification (Kirkpatrick et al., 2017). Current systems are better characterized as narrow experts rather than

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.

Multitask learning (Caruana, 1997) is a promising framework for improving general performance. However, multitask training in NLP is still nascent. Recent work reports modest performance improvements (Yogatama et al., 2019) and the two most ambitious efforts to date have trained on a total of 10 and 17 (*dataset, objective*) pairs respectively (McCann et al., 2018) (Bowman et al., 2018). From a meta-learning perspective, each (*dataset, objective*) pair is a single training example sampled from the distribution of datasets and objectives. Current ML systems need hundreds to thousands of examples to induce functions which generalize well. This suggests that multitask training may need just as many effective training pairs to realize its promise with current approaches. It will be very difficult to continue to scale the creation of datasets and the design of objectives to the degree that may be required to brute force our way there with current techniques. This motivates exploring additional setups for performing multitask learning.

The current best performing systems on language tasks

^{*},^{**}Equal contribution ¹OpenAI, San Francisco, California, United States. Correspondence to: Alec Radford <alec@openai.com>.

GPT-3 (2020)

- ▶ **Número de blocos decoder:** 96
- ▶ **Número de parâmetros:** 175 bilhões
- ▶ **Pré-treinamento:** WebText dataset
 - ▶ CommonCrawl (410 bilhões de tokens)
 - ▶ WebText2 (19 bilhões de tokens)
 - ▶ Books1 e Books 2 (12 e 55 bilhões de tokens)
 - ▶ Wikipedia (3 bilhões)
- ▶ **Ajuste-fino**
 - ▶ Few-shot learning
 - ▶ One-shot learning
 - ▶ Zero-shot learning

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution

†Johns Hopkins University, OpenAI

Author contributions [listed at end of paper.](#)

Instruct GPT (2022)

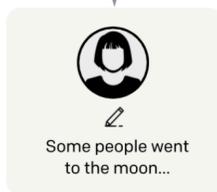
GPT-3 ajustado com Aprendizado por Reforço para melhorar a qualidade dos prompts, principalmente aumento de confiança e remoção de saídas tóxicas.

Step 1
Collect demonstration data, and train a supervised policy.

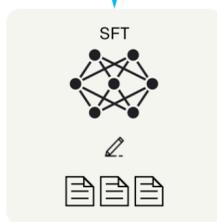
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2
Collect comparison data, and train a reward model.

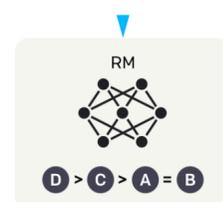
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.

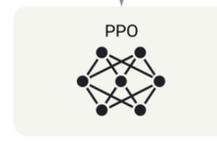


Step 3
Optimize a policy against the reward model using reinforcement learning.

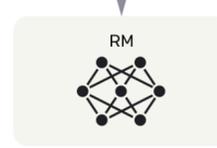
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*

Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray

John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens

Amanda Askell†

Peter Welinder

Paul Christiano*†

Jan Leike*

Ryan Lowe*

OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user’s intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models *InstructGPT*. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

1 Introduction

Large language models (LMs) can be “prompted” to perform a range of natural language processing (NLP) tasks, given some examples of the task as input. However, these models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions (Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021; Weidinger et al., 2021; Tamkin et al., 2021; Gehman et al., 2020). This is because the language modeling objective

*Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads. Corresponding author: lowe@openai.com.

†Work done while at OpenAI. Current affiliations: AA: Anthropic; PC: Alignment Research Center.

ChatGPT (2022)



You

Conte-me a sua história em uma pequena lista de eventos importantes. Inclua data para cada evento.



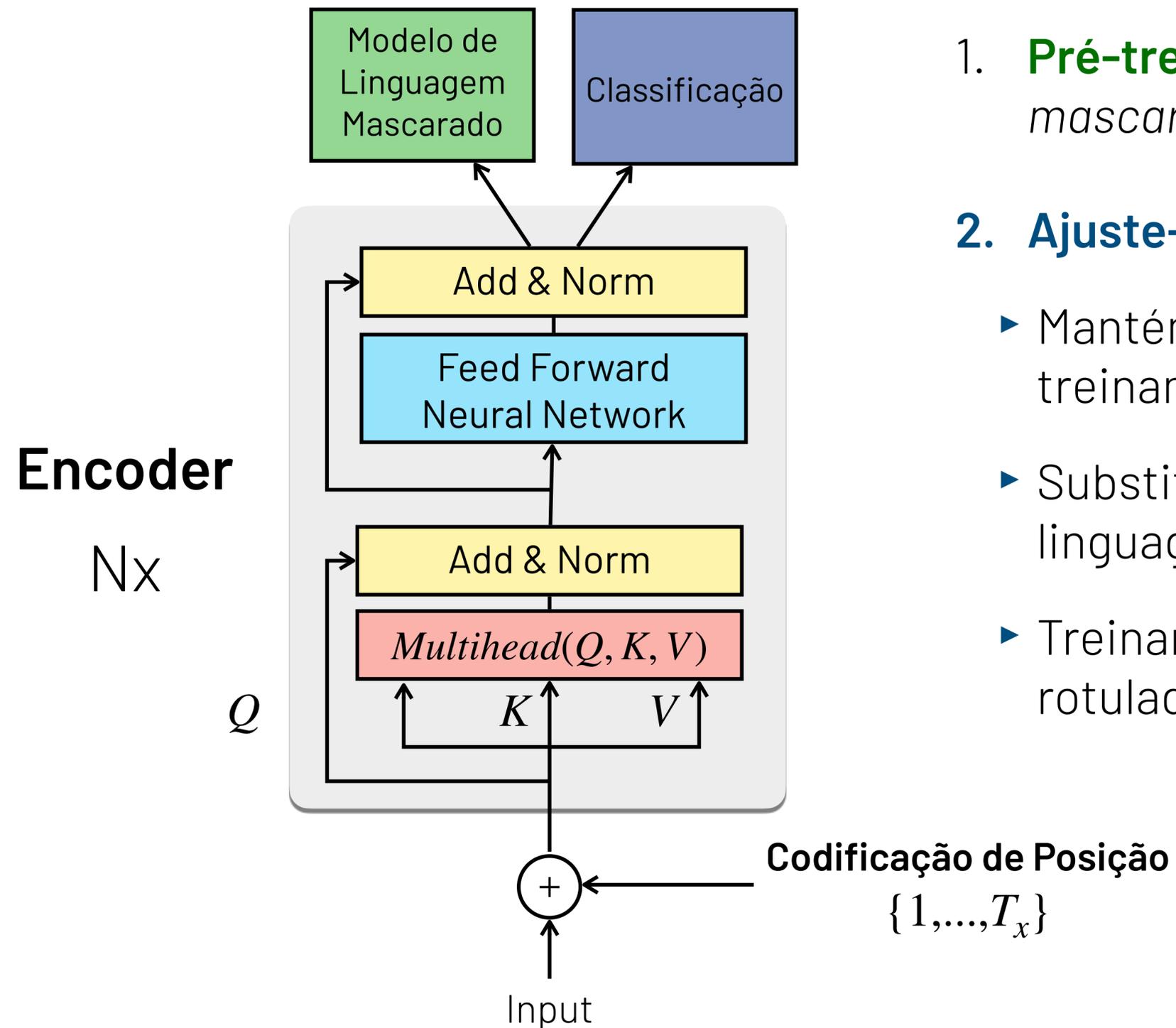
ChatGPT

1. **Nascimento da GPT-3.5:** Criada pela OpenAI, a GPT-3.5 foi lançada em junho de 2020 como a última iteração da série GPT, representando um avanço significativo em capacidade e desempenho.
2. **Treinamento e Aprimoramento:** Ao longo de 2020 e 2021, a GPT-3.5 passou por extensivos treinamentos, absorvendo uma vasta gama de conhecimentos e habilidades linguísticas a partir de diversas fontes online.
3. **Lançamento Público:** Em novembro de 2021, a GPT-3.5 foi disponibilizada para o público em geral, permitindo que desenvolvedores e usuários explorassem suas capacidades em diversas aplicações, desde assistência de linguagem natural até criação de conteúdo.
4. **Integração em Plataformas Diversas:** Ao longo de 2022, a GPT-3.5 foi integrada em várias plataformas e serviços online, ampliando sua presença e utilidade em diferentes setores.
5. **Atualizações e Aprimoramentos Contínuos:** Desde o lançamento, a GPT-3.5 recebeu atualizações regulares, refinando seus algoritmos e melhorando suas capacidades com base no feedback e nas necessidades emergentes.
6. **Contribuição para Pesquisas e Aplicações Avançadas:** A partir de 2023, a GPT-3.5 começou a ser utilizada em pesquisas mais avançadas e em aplicações especializadas, contribuindo para o desenvolvimento de tecnologias emergentes.
7. **Status Atual:** Em novembro de 2023, a GPT-3.5 continua a ser uma referência em processamento de linguagem natural, desempenhando um papel fundamental em diversas áreas, desde assistentes virtuais até projetos de inteligência artificial complexos.



BERT

Utiliza apenas o Encoder do Transformer



1. **Pré-treinar** um *modelo de linguagem mascarado* em textos não-rotulados

2. **Ajuste-fino** para classificação

- ▶ Mantém os pesos da rede após o pré-treinamento;
- ▶ Substituindo a camada do modelo de linguagem por uma camada de classificação;
- ▶ Treinamento com um conjunto de dados rotulado.

BERT (2019)

- ▶ **Número de blocos encoder:**
 - ▶ Base: 12
 - ▶ Large: 24
- ▶ **Número de parâmetros:**
 - ▶ Base: 110 milhões
 - ▶ Large: 240 milhões
- ▶ **Pré-treinamento:**
 - ▶ BooksCorpus (800 milhões de tokens)
 - ▶ English Wikipedia (2500 milhões de tokens)
- ▶ **Ajuste-fino**
 - ▶ GLUE
 - ▶ Question Answering
 - ▶ Commonsense inference

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin, mingweichang, kentonl, kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT alleviates the previously mentioned unidirectionality constraint by using a “masked language model” (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

Lista de Modelos BERT

	Bert	RoBERTa	DistilBERT	ALBERT
Parâmetros	Base: 110M	Base: 125M	Base: 66M	Base: 12M
	Large: 340M	Large: 355M		Large: 18M
Nº de camadas/ Dimensão/ Nº de cabeças	Base: 12/768/12	Base: 12/768/12	Base: 6/768/12	Base: 12/768/12
	Large: 24/1024/16	Large: 24/1024/16		Large: 24/1024/16
Tempo de treinamento	8 x V100 x 12d	1024 x V100 x 1d	8 x V100 x 3.5d	Base: -
	280 x V100 x 1d			Large: 1.7x mais rápido
Desempenho		88.5 on GLUE		89.4 on GLUE
Dados de pré-treinamento	BookCorpus + English Wikipedia = 16 GB	BookCorpus + English Wikipedia + CCNews + OpenWebStories = 160 GB	BookCorpus + English Wikipedia = 16 GB	BookCorpus + English Wikipedia = 16 GB



BERTimbau

- ▶ **Número de blocos encoder:**
 - ▶ Base: 12
 - ▶ Large: 24
- ▶ **Número de parâmetros:**
 - ▶ Base: 110 milhões
 - ▶ Large: 240 milhões
- ▶ **Pré-treinamento:**
 - ▶ brWaC Corpus (2,68 bilhões de tokens)
- ▶ **Ajuste-fino**
 - ▶ GLUE
 - ▶ Question Answering
 - ▶ Commonsense inference



Próxima aula

A31: Conclusão

Apresentação dos projetos finais desenvolvidos pelos alunos, revisão dos tópicos vistos na disciplina e recomendações de atividades futuras.