

# Facebook/OPT-125M-Portuguese LLM



Open in Colab

# 1. Questões

1. **Explicação de conceitos importantes do exercício feito**
2. ~~Técnicas para garantir que a implementação está correta~~
3. ~~Truques de código que funcionaram~~
4. **Problemas e soluções no desenvolvimento**
5. ~~Resultados interessantes/inesperados~~
6. **Uma dúvida "básica" que você ou os colegas possam ter**
7. ~~Um tópico "avançado" para discutirmos~~

## 2. Conceitos

A **perplexidade** está intimamente relacionada à entropia cruzada que é diretamente minimizada durante o treinamento (intrínseca). Outras métricas são o **BLEU** e o **ROUGE** que estão mais relacionadas, por exemplo, à precisão da classificação (extrínseca). **BLEU** é uma pontuação semelhante à precisão para avaliar a qualidade de um texto traduzido. **ROUGE** é uma pontuação semelhante ao recall para avaliar o texto resumido.

- Uma alta perplexidade indica maior dificuldade do modelo em prever a próxima palavra na sequência.
- Ou seja, se o seu modelo de linguagem tem perplexidade de  $N$ , significa que ele é igual a um classificador aleatório de  $N$  classes

“ A perplexidade é apenas a exponenciação da entropia! [artigo](#)

Entropia é o número médio de bits necessários para codificar a informação contida em uma variável aleatória, portanto, a exponenciação da entropia deve ser a quantidade total de todas as informações possíveis, ou mais precisamente, o número médio ponderado de escolhas que uma variável aleatória possui.

”


## 2.1 Conceitos

- por que usar PPL ao invés de entropia cruzada? <https://towardsdatascience.com/perplexity-intuition-and-derivation-105dd481c8f3> ver no chatgpt tbm
- Diferença entre logits e probabilidades
  - logits são valores numéricos não normalizados brutos produzidos pelo modelo na saída da sua última camada
  - probabilidades são valores normalizados dos logits que indicam a probabilidade relativa de cada palavra na sequência ou outro tipo de tarefa, caracterizando uma distribuição de probabilidade
  - Ambas as medidas podem ser usadas para avaliar a qualidade de um modelo de linguagem e para gerar previsões de palavras futuras em uma sequência.

### 3. Problemas e soluções no desenvolvimento

- primeira vez treinando/finetuning um LLM
- **Dataset** do HF é muito bom. Customização deles pra fazer preprocessing e outras coisas
- notebooks/apresentações interessantes
  - **Mirelle** e outros colegas
    - dataset de teste para overfitting
  - **Marcos Vieira**
    - truques de código
    - tempo gasto artigo + notebook (meu caso foi por parecido também, com ~20h de estudo na semana)
  - **Eduardo Oliveira**
    - treino com length de 2048
    - batchs em disco

# 3.1 Problemas e soluções no desenvolvimento

- calcular perplexidade inicial do dataset = número de palavras no vocabulário
- HF facilita demais o uso de transformers, mas é importante ficar atento a outros aspectos do modelo e tokenizador
- usei T4 16Gb no Colab Pro
  - treino demorado
  - +1 epoch = +2h-3h de treino
- no final consegui pegar um A100 pra treinar. ~~40 créditos que valem do que dinheiro~~ 
- teste de overfit com dataset pegando os 100 primeiros textos
- salvar `tokenized dataset` no disco

## 4. Resultados

split train/validation 80/20 (200.000, 50.000)

seq length	epochs	batch size	gpu	time/epoch	val_ppl
256	1	4	T4 16GB	~2h	31.67
512	1	4	T4 16GB	OOM 💣	-
512	1	8	A100 40GB	45min	10.96
512	1+1	8	A100 40GB	colab fail 🤔	-
512	2	8	A100 40GB	~2h	<b>9.67</b> 🏆
1024	1	16	A100 40GB	OOM 💣	-

epoch	train loss	val loss
1	2.571000	2.339543
2	2.327500	2.269327

sem overfit (1 epoch somente e muitos dados)

# 4.1 Resultados

[transformer\\_generation config](#)

```
max_length=50, do_sample=True
```

Era uma vez na fazenda irmã, e seu vizinho e o vizinho que comanda o salão de restaurantes com a cachoeira e a borboleta. Na frent

Era uma vez na fazenda irmão o destacando e os dias de hoje fomei meus olhos de manho que a chuva deu conta de que meu am

Era uma vez na fazenda irmão de Paulo que, no início dos anos 90, tanto na família do empresário, como na família dos comerci

Era uma vez na fazenda irmão que desafiou a vida para poder se dedicar ao povo que ele tinha. Ele, por sua vez, foi um grande

Era uma vez na fazenda ilesa do sossego Zey, um pediatra que fez a reativação dos dentes fazendo parte da rotina com as crian

Era uma vez na fazenda irmã do casal que pessoas de coração amada e muito amadora deu àquelas que gostavam de se amare

Era uma vez na fazenda ilegal a mãe de um filho que estava comercialando seus produtos. A mãe da menina sentiu-se aos dentes



## 4.2 Resultados

```
max_length=50, do_sample=True, num_beams=5, early_stopping=True, no_repeat_ngram_size=2
```

Era uma vez na fazenda Ângela, na cidade de Santo Antônio de Jesus, no Rio de Janeiro, que se encontrava com a família de um homem

Era uma vez na fazenda irmã de um grupo de amigos que estavam na cidade de Florianópolis, nos Estados Unidos. Eles se encontrav

Era uma vez na fazenda irmã de um grupo de pessoas que trabalhavam como lanchonete. Eles tinham um grande público, que

Era uma vez na fazenda Ângela, no bairro São João, que eu tinha oportunidade de conhecer um pouco mais sobre a histó

Era uma vez na fazenda irmã de um jovem de 19 anos que acabou de morar em um apartamento na cidade de São José dos Campos, no interior de

Era uma vez na fazenda irmã de um jovem de 16 anos que, após a morte de sua mãe, decidiu seguir para a cidade de

Era uma vez na fazenda irmã de um amigo que eu tinha ouvido falar sobre o assunto. Não sabia o que fazer, mas acho que quando

Era uma vez na fazenda Ângela, no bairro de São Bento, que se encontrava com um grupo de moradores que estavam próximos de

## 5. Questões/Dúvidas para discussão

- concatenação de sentenças para evitar padding e melhorar uso da gpu? discussão do slack
- não precisaria ajustar o vocab\_size da config do modelo para o novo dataset tokenizado?
- salvar processamento/tokenizer e outras coisas em disco para depois habilitar gpu e treinar - salvei o dataset tokenizado
- como calcular a perplexidade sem treino?
  - vi no notebook do **Thiago Laitz** que ele usa a loss da primeira iteração do finetuning
- como faria pra ajustar a função de custo pra usar ppl?