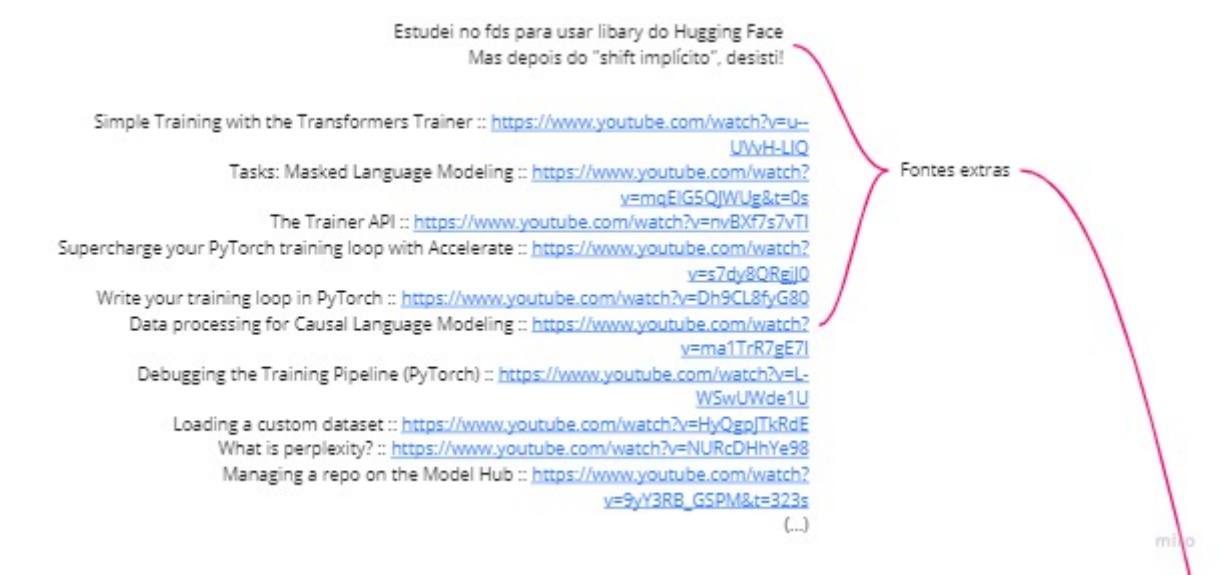
Code Presentation
IA368DD_2023S1: Deep Learning aplicado a Sistemas de Buscas
Student: Marcus Vinícius Borela de Castro

Treinamento de um modelagem de linguagem decoder-only em textos português

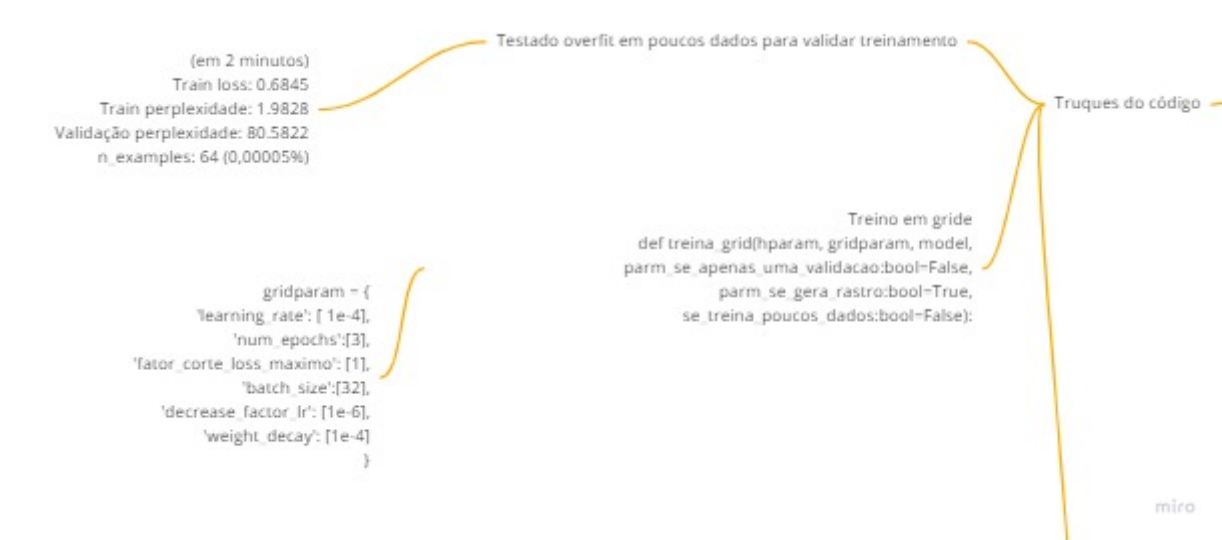

Estudei no fds para usar libary do Hugging Face
Mas depois do "shift implícito", desisti!
Simple Training with the Transformers Trainer :: https://www.youtube.com/watch?v=u--UVvH-LIQ
Tasks: Masked Language Modeling :: https://www.youtube.com/watch?v=mqElG5QJWUg&t=0s
The Trainer API :: https://www.youtube.com/watch?v=nvBXf7s7vTI
Supercharge your PyTorch training loop with Accelerate :: https://www.youtube.com/watch?v=s7dy8QRgjJ0
Write your training loop in PyTorch :: https://www.youtube.com/watch?v=Dh9CL8fyG80
Data processing for Causal Language Modeling :: https://www.youtube.com/watch?v=ma1TrR7gE7I
Debugging the Training Pipeline (PyTorch) :: https://www.youtube.com/watch?v=L-WSwUWde1U
Loading a custom dataset :: https://www.youtube.com/watch?v=HyQgpJTkRdE
What is perplexity? :: https://www.youtube.com/watch?v=NURcDHhYe98
Managing a repo on the Model Hub :: https://www.youtube.com/watch?v=9yY3RB_GSPM&t=323s
(...)
Fontes extras

Concatenando os textos no dataset conforme sugerido em https://huggingface.co/course/chapter7/6?fw=pt

"A more efficient way to prepare the data is to join all the tokenized samples in a batch with an eos_token_id token in between, and then perform the chunking on the concatenated sequences. "

Conceitos

```
class MyDataset():
def __init__(self, texts: List[str], tokenizer,
max_seq_length: int):
(...)
def __getitem__(self, idx):
return self.data_tensor[idx][:-1],
self.data_tensor[idx][1:]
```


Testado overfit em poucos dados para validar treinamento
(em 2 minutos)
Train loss: 0.6845
Train perplexidade: 1.9828
Validação perplexidade: 80.5822
n_examples: 64 (0,00005%)
Truques do código
Treino em gride
def treina_grid(hparam, gridparam, model,
parm_se_apenas_uma_validacao:bool=False,
parm_se_gera_rastro:bool=True,
se_treina_poucos_dados:bool=False):
gridparam = {
'learning_rate': [1e-4],
'num_epochs':[3],
'fator_corte_loss_maximo': [1],
'batch_size':[32],
'decrease_factor_lr': [1e-6],
'weight_decay': [1e-4]
}
miro

Salvando rastro no neptune.ai

https://app.neptune.ai/marcusborela/IA386DD/e/IAD-29

assert saida['logits'].shape[1] == hparam['max_seq_length'], "Saida[1] deveria ser do tamanho de max_seq_length"
assert saida['logits'].shape[2] == hparam['vocab_size'], "Saida[2] deveria ser do tamanho do vocabulário do tokenizador"
Asserts
Bom prompt: tem alguma incorreção nesse código ...
Consultoria do ChatGPT
Técnicas para buscar correção

Problemas encontrados e soluções

- O treino estava muito devagar...
- Solução: aumento do batch_size (limite gpu)

Avaliação de consumo GPU RTX-3090 (24gb)

seq_len = 100

batch_size=8 :: gpu 7807gb; tempo: 0,25% a cada 18 minutos!
batch_size=24 :: gpu 11043gb; tempo: 0,25% a cada 8 minutos!
batch_size=72 :: gpu 18235gb; tempo: 0,25% a cada 4 minutos!
batch_size=96 :: gpu 19191gb; tempo: 0,25% a cada 3:30 minutos!

seq_len = 250

batch_size=32 :: gpu 18471gb; tempo: 0,25% a cada 4 minutos!

Treino 1
max_seq_length=250,
senteces treino = 249800
teste e validação = 100

Dados:
- training examples: 4328981
- valid examples: 2036
- test examples: 1143
- training examples por época: 4328981
- training examples total: 12986943

Parametros:
- batch size: 96
- num_epochs: 3
- eval_every_steps: 338 (0,25%)
- early_stop: 3380 (steps)

Contexto:
- tempo total: 10.3 horas
- best_step: 54756
- treino por step: 0.627s
- n_examples: 5581056 (42,97%)

Perplexidade

- treino: 13.69
- teste: 12.90
- validação: 15.14

Teste geração frase:

{'max_length': 120, 'top_p ': 1, 'temperature': 0, 'top_k': 50, 'repetition_penalty': 1.5}

'[Praticar esportes é] ingerir bebida alcoólica e fazer exercícios de maneira saudável. A praticidade do atletismo pode ser feita com o objetivo principal, seja na formatação dos músculos (cortinas), no corpo humano...\n'

Treino 2
max_seq_length=250,
senteces treino = 249800
teste e validação = 100

Dados
- training examples: 1740410
- valid examples: 818
- test examples: 459
- training examples por época: 1740410
- training examples total: 5221230

Parametros:
- batch size: 32
- num_epochs: 3
- eval_every_steps: 407 (0,25%)
- early_stop: 4070 steps

Step: 45991 Amostras:1471712 de um total de 5221230 (28.187%)

Momento: [2023-Mar-29 20:25:47] lr: 9.56031e-05 Train loss: 2.6194 perplexidade: 12.9693
Validação perplexidade: 14.2466 novo best valid 14.246589894152141

Testando modelo com perplexidade menor. Frase gerada: ('Praticar esportes é ilegal e não pode ser punido. Ainda que seja um crime, ou uma pena de morte por dano moral ao menos no primeiro ano do curso da faculdade (que acontece em outubro deste ano), ele tem direito à liberdades individuais comuns na prática dos atletas brasilienses: "Acho legal isto", disse José Maria Lopes Filho, presidente da Federação'
{'max_length': 120, 'top_p ': 1, 'temperature': 0, 'top_k': 50, 'repetition_penalty': 1.5})

(em andamento)
https://app.neptune.ai/marcusborela/IA386DD/e/IAD-29/metadata

Dúvidas básicas

Dúvida gerada pelo código abaixo (coloquei no slack)

```
===CODE=====
MODEL_NAME = 'facebook/opt-125m'
tokenizer = AutoTokenizer.from_pretrained(MODEL_NAME)
model = AutoModelForCausalLM.from_pretrained(MODEL_NAME)
print(f'Por quê model.config.vocab_size ({model.config.vocab_size}) != tokenizer.vocab_size
({tokenizer.vocab_size}). Não deveriam ser iguais?')
===END=====
===OUTPUT=====
Por quê model.config.vocab_size (50272) != tokenizer.vocab_size (50265). Não deveriam ser
iguais?
```

Tópicos avançados
Bom ter controle sobre a rotina de treinamento e saber o que acontece "behind the scenes"
(aproveitei muito de um exercício do curso que fiz IA021 com os Professores Lotufo e Rodrigo no 1o semestre de 2022)
Mas é um trabalho hercúleo...
Será que vale a pena?