

### 3. Modelo Treinado

#### 3.1. Código fonte utilizado para o treinamento do modelo de linguagem com LoRA/QLoRA.

- Todo o código fonte, incluindo o modelo utilizando LoRa, está localizado no arquivo “handler.ipynb”, conforme a figura abaixo.

```
Treinando o modelo utilizando o LoRa

from transformers import AutoModelForCausalLM, AutoTokenizer

device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model_name = "pierregrillou/gpt2-small-portuguese"
model = AutoModelForCausalLM.from_pretrained(model_name)
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)

tokenizer.pad_token = tokenizer.eos_token
tokenizer.padding_side = "right"
max_seq_length = 768
trainingArgs = TrainingArguments(
    output_dir='output',
    num_train_epochs=50,
    per_device_train_batch_size=4,
    save_strategy="epoch",
    learning_rate=2e-4
)
peft_config = LoraConfig(
    lora_alpha=32,
    lora_dropout=0.1,
    r=8,
    task_type="CAUSAL_LM",
)
model.train()
trainer = SFTTrainer(
    model=model,
    train_dataset=dataset['train'],
    eval_dataset=dataset['test'],
    peft_config=peft_config,
    tokenizer=tokenizer,
    packing=True,
    formatting_func=prompt_instruction_format_to_train,
    max_seq_length=max_seq_length,
    args=trainingArgs
)
PATH_TO_MODEL_CHECKPOINT = "sample_data/checkpoint"
history = trainer.train()
model.save_pretrained(PATH_TO_MODEL_CHECKPOINT)
tokenizer.save_pretrained(PATH_TO_MODEL_CHECKPOINT)
history
```

#### 3.2. Relatório de desempenho do modelo, incluindo métricas de avaliação e análise de resultados. A participação de treino contém 900 (90%) de perguntas e respostas, enquanto a de teste ficou com 100 (10%) da base.

- O modelo foi treinado com 50 épocas, e suas métricas de avaliação foram as seguintes:

```
Avaliação do modelo na partição de treino

history.metrics

{'train_runtime': 1007.3974,
 'train_samples_per_second': 5.757,
 'train_steps_per_second': 1.439,
 'total_flos': 2281122614476800.0,
 'train_loss': 1.3643781148976293,
 'epoch': 50.0}
```

- As métricas de avaliação na partição de teste foram as seguintes:

```
Avaliação do modelo na partição de teste

[ ] trainer.evaluate()

[ ] [2/2 00:00]
{'eval_loss': 13.955838203430176,
 'eval_perplexity': 1150650.75,
 'eval_runtime': 10.5195,
 'eval_samples_per_second': 1.141,
 'eval_steps_per_second': 0.19,
 'epoch': 50.0}
```

- Em relação à análise de resultados, a figura abaixo mostra exemplos de entradas e previsões realizadas pelo modelo após o treinamento.

Testando o modelo para verificar se ele responde as perguntas

```
[27] def get_prediction(model, prompt):
    model.eval()
    inputs = tokenizer(prompt, return_tensors="pt").to(device)
    outputs = model.generate(
        input_ids=inputs["input_ids"],
        attention_mask=inputs["attention_mask"],
        max_length=max_seq_length,
        return_dict_in_generate=True
    )
    pred = tokenizer.decode(outputs.sequences[0])
    return pred.split("<|endoftext|>")[0]

for i in range(10):
    sample = get_sample()
    print(f"PERGUNTA: {sample['Instrução']}")
    print(f"RESPOSTA CORRETA: {sample['Resposta']}")
    print(f"RESPOSTA DO MODELO: {get_prediction(model, sample['Instrução'])}\n\n")
```

RESPOSTA: O PROGUNG é responsável pela coordenação do programa Residência Pedagógica.

PERGUNTA: Qual é o objetivo do Decreto nº 8.537/15?

RESPOSTA CORRETA: O Decreto nº 8.537/15 regulamenta a Lei nº 12.852/13 e a Lei nº 12.933/13, que tratam do

Setting `pad\_token\_id` to `eos\_token\_id`:50256 for open-end generation.

RESPOSTA DO MODELO: Qual é o objetivo do Decreto nº 8.537/15?

RESPOSTA: O objetivo do decreto é o aproveitamento das atividades de ensino e pesquisa para a formação de

PERGUNTA: Quando a resolução entra em vigor?

RESPOSTA CORRETA: A resolução entra em vigor na data de sua aprovação pela Câmara de Ensino de Graduação.

Setting `pad\_token\_id` to `eos\_token\_id`:50256 for open-end generation.

RESPOSTA DO MODELO: Quando a resolução entra em vigor?

RESPOSTA: O decreto-lei nº 5.636, de 30 de dezembro de 1969, que regulamenta a obrigatoriedade da matrícula

PERGUNTA: Quais são os requisitos para o professor ministrar disciplinas semipresenciais?

RESPOSTA CORRETA: Possuir capacitação específica em docência a distância (EAD) em ambiente virtual de apre

Setting `pad\_token\_id` to `eos\_token\_id`:50256 for open-end generation.

RESPOSTA DO MODELO: Quais são os requisitos para o professor ministrar disciplinas semipresenciais?

RESPOSTA: O professor deve apresentar a documentação necessária para o processo de apresentação do curso.

PERGUNTA: Como são disciplinados os casos omissos na resolução?

RESPOSTA CORRETA: Os casos omissos disciplinados nesta resolução deverão ser decididos pela Câmara de Ensi

Setting `pad\_token\_id` to `eos\_token\_id`:50256 for open-end generation.

RESPOSTA DO MODELO: Como são disciplinados os casos omissos na resolução?

RESPOSTA: Os casos omissos na resolução são considerados omissos na resolução.