

Notebook – OPT-125M com treinamento em português

Leandro Carísio

Problemas e soluções no desenvolvimento

Problema: como prever a próxima palavra do texto?



Como eu uso esse modelo para prever a próxima palavra de um texto?



Para usar o modelo treinado para prever a próxima palavra de um texto, você pode seguir os seguintes passos:

1. Tokenizar o texto de entrada usando o tokenizador do modelo. O tokenizador irá dividir o texto em uma sequência de tokens que podem ser usados como entrada para o modelo.

```
python
text = "Eu gosto de comer"
input_ids = tokenizer.encode(text, return_tensors='pt')
```

2. Passar a sequência de tokens através do modelo para obter a distribuição de probabilidade condicional para o próximo token.

```
python
with torch.no_grad():
    output = model(input_ids)
    next_token_logits = output.logits[:, -1, :]
```

3. Obter a lista dos tokens com as maiores probabilidades condicionais.

```
python
next_token_probs = torch.softmax(next_token_logits, dim=-1)
top_next_token_probs, top_next_token_ids = torch.topk(next_token_probs, k=5)
```

4. Converter os IDs dos tokens em texto usando o tokenizador e imprimir as palavras mais prováveis.

```
python
for i in range(top_next_token_ids.size(1)):
    next_token = tokenizer.decode(top_next_token_ids[0, i].item())
    print(f"Próxima palavra {i+1}: {next_token}, Probabilidade: {top_next_token_probs[0, i].item()}")
```

Este código irá imprimir as próximas palavras mais prováveis com as suas respectivas probabilidades condicionais. Note que você pode alterar o valor do parâmetro `k` na etapa 3 para obter mais ou menos palavras candidatas. Além disso, tenha em mente que o modelo só pode fazer previsões com base no que aprendeu durante o treinamento, portanto, se o texto de entrada for muito diferente dos dados de treinamento, as previsões podem não ser precisas.

Regenerate response

Problemas e soluções no desenvolvimento

Resultado da sugestão do ChatGPT:

Frase:

"Era uma vez"

Primeira implementação do ChatGTP:

Era uma vez um homem bus ca í d as mais e e u ma is so que e e e e e e e e e e
e
e e e e e e e

Solução foi usar a biblioteca direto (copiado do caderno da Monique):

```
generator = pipeline('text-generation', model=model, tokenizer=tokenizer, device=device_generator)
output = generator(frase, max_length=max_length, do_sample=amostragem_estocastica)
return output[0]['generated text']
```

Chamando pipeline:

Era uma vez um homem que estava em uma caminhonete, quando foi atropelado por um caminhão. Ele foi atingido por um carro e foi socorrido. O fato aconteceu na noite de ontem (14), no bairro São João, em Campo Grande. Segundo a Polícia Militar, o homem estava em uma motocicleta e foi atropelado por um caminhão. Ele foi socorrido e foi até o Hospital Regional de Campo Grande. O motorista do caminhão, que est

Problemas e soluções no desenvolvimento

Ao pedir pro ChatGPT “abrir” a implementação do pipeline:

```
generator = pipeline('text-generation', model=model, tokenizer=tokenizer, device=device_generator)
output = generator(frase, max_length=max_length, do_sample=amostragem_estocastica)
return output[0]['generated_text']
```

```
def continuar_frase_implementacao_chatgpt_2(frase, max_length=100, amostragem_estocastica=True, temperatura=0.7):
    # Tokeniza o texto inicial
    input_ids = tokenizer.encode(frase, return_tensors='pt').to(device)

    # Gera a sequência de texto
    output = model.generate(
        input_ids=input_ids,
        max_length=max_length,
        do_sample=amostragem_estocastica,
        temperature=temperatura
    )

    # Decodifica a sequência de texto gerada
    output_text = tokenizer.decode(output[0], skip_special_tokens=True)

    return output_text
```

Resultados interessantes/inesperados

A forma como o final da frase termina muda todo o resto. Intuitivamente faz sentido se a gente pensa em ponto (final/exclamação/interrogação etc). Mas não faz muito sentido se a frase termina sem pontuação ou com algum(ns) espaço(s).

Frase:

"Era uma vez."

[illegible]

Frase:

"Era uma vez"

Era uma vez um homem que estava em uma caminhonete, quando foi atropelado por um caminhão. Ele foi atingido por um carro e foi socorrido. O fato aconteceu na noite de ontem (14), no bairro São João, em Campo Grande. Segundo a Polícia Militar, o homem estava em uma motocicleta e foi atropelado por um caminhão. Ele foi socorrido e foi até o Hospital Regional de Campo Grande. O motorista do caminhão, que est

Frase:

"Era uma vez "

Era uma vez irmã de um homem que se encontrava em uma cidadezinha do interior de São Paulo. Ele estava com um amigo, um homem de origem indígena, e ao lado dele, um homem de origem alemã. O homem era um homem de origem alemã, que se encontrava em uma cidadezinha do interior de São Paulo. Ele estava com um amigo, um homem de origem alemã, e ao lado dele, um homem de origem alemã. O homem era um homem de orig

Frase:

"Era uma vez "

[illegible]

Obs.: Sem amostragem estocástica e com temperatura = 0.

Um tópico avançado(*) para discutirmos

Considerando os métodos de decode disponíveis [1] (greedy search, beam Search, top-k, top-p), quando usar cada um deles?

Algun oferece resultado claramente melhor que todos os outros ou depende da aplicação?

(*) ou dúvida básica, depende de quem está lendo

[1] <https://huggingface.co/blog/how-to-generate>

Obrigado

Leandro Carísio
carisio@gmail.com