



Treino Modelo de Linguagem Língua Portuguesa

Monique Monteiro –
moniquelouise@gmail.com

Conceitos Importantes

- *Causal Language Modeling / Modelagem de Linguagem Causal:*
 - *“...a technique used in natural language processing and machine learning to generate text by predicting the next word in a sequence of words, given the previous words in the sequence. (...)”*
- *Perplexity / Perplexidade*
 - *“...a quantitative measure used to evaluate the performance of language models, **particularly those trained for predicting the next word in a sequence of words**. It is a measure of how well the language model is able to predict the next word in a given sequence, based on the probability distribution of possible next words. (...)”*

Experimentos

- Treinamento de 3 modelos de linguagem com diferentes combinações entre tamanho do bloco e tamanho do batch
- Hiperparâmetros:
 - 3 épocas
 - LR = $5e-5$ (*linear scheduler*)
Valores default HuggingFace (caberia *grid search* idealmente)
 - Uso completo dos textos
- Perplexidade do checkpoint: 23.65

Block Size	Batch Size	Perplexity
1024	4	7.94
512	8	8.33
256	16	8.91

Técnicas para garantir corretude

-
- Avaliação qualitativa do texto gerado
 - Exemplos (prompt “*O Ministério Público Federal de Santa Catarina investiga a construção* ”)
 - *O Ministério Público Federal em Santa Catarina investiga a construção ilegal de construções particulares na Zona Leste de Florianópolis.*
 - *O Ministério Público Federal em Santa Catarina investiga a construção ilegal de casas particulares em áreas de risco e que faz uso commercial*
 - *O Ministério Público Federal em Santa Catarina investiga a construção ilegal de quatro escolas da rede básica de ensino. A obra, que custava quase*
 - *O Ministério Público Federal em Santa Catarina investiga a construção ilegal de moradias na Ilha da Fronteira, na região Noroeste do Estado,*
 - *O Ministério Público Federal em Santa Catarina investiga a construção ilegal de moradia na Área de Preservação Permanente (APP) do Povoado*

Técnicas para garantir corretude

-
- Avaliação qualitativa do texto gerado
 - Exemplos (prompt “*Era uma vez*”)
 - *Era uma vez uma família grande. Tudo corria tranquilo quase na rua em que estávamos numa das casas, os dois piscando em direção*
 - *Era uma vez o "bicho que voa". Atravessou de maneira singular essa paisagem. Depois de alguns minutos, como mãe, uma sé*
 - *Era uma vez uma viúva de fé que com o seu filho estuda em Colnago, no sul de Macedónia, e que era uma criança em uma cidade*
 - *Era uma vez por semana, o que é pior. Fazer o ponto final naquele dia, seis horas e uma noite de chuva, e o resultado era esper*
 - *Era uma vez no interior de uma cidade em áreas de terra muito povoadas e com abundância de animais abatidos sob a alegação*
 - *Era uma vez na praia, quando um pinguim me atirou, eu estava a fazer a segunda aula quando o menino se ajoelhou na cadeira e*

Resultados inesperados

-
- Uma época a mais aumentou a perplexidade em 0.01
 - Pode não ter significância estatística
 - Erro de treinamento reduziu em apenas 0.02
 - Limitações de tempo impediram investigações adicionais
 - 1 época a mais → + 5 horas de treinamento
 - Ao “prosseguir” treinamento:
 - Definir explicitamente taxa de aprendizado inicial como igual à final do treinamento anterior
 - Outros resultados:
 - Uso completo dos textos aumenta em mais de 3 vezes o tempo de treinamento
 - Código mais “conciso” pode induzir a ignorar aspectos importantes (ex.: truncagem, tokenização com/sem “overflowing” etc.) e pré-processamento