RELATÓRIO FINAL – Fine-Tuning e Destilação do BERT-tiny para Português (BERT-tiny-PT)

Abstract
Este projeto investiga a possibilidade de adaptar o modelo BERT-tiny para o português brasileiro utilizando técnicas de destilação de conhecimento. O objetivo central é produzir um modelo leve, eficiente e capaz de compreender português, mantendo parte da capacidade representacional de um professor muito maior, o BERTimbau-base. Para isso, foram exploradas duas abordagens: a destilação parcial, preservando o tokenizador original do aluno, e a destilação total, que substitui completamente o vocabulário do aluno pelo do professor. Os experimentos demonstram que o modelo resultante, denominado BERT-tiny-PT, aprende de fato os padrões básicos da língua portuguesa, com melhora substancial em Masked Language Modeling. Contudo, o desempenho em tarefas downstream de natureza semântica ainda fica abaixo do esperado, revelando limitações da técnica quando aplicada sem um pré-treinamento mais profundo ou fine-tuning especializado.

1 Introduction

Modelos de linguagem de grande porte apresentam desempenho excepcional em tarefas de compreensão textual, porém com altos custos computacionais, tornando difícil sua utilização em cenários restritos em memória, latência e recursos. Uma solução bastante adotada é a destilação de conhecimento, na qual um modelo grande (teacher) transfere parte de sua capacidade para um modelo menor (student), que se torna mais rápido e eficiente, preservando parte do poder expressivo.

Neste projeto, buscamos adaptar o modelo BERT-tiny, originalmente treinado apenas em inglês,para o português brasileiro. Para isso, utilizamos o modelo BERTimbau-base como professor, devido à sua robusta capacidade e ao seu amplo vocabulário treinado em textos do português. A pergunta que guiou todo o desenvolvimento foi se seria possível construir um modelo extremamente pequeno, rápido e economicamente viável, mas com capacidade de compreender textos em português de forma útil. Ao longo do projeto, construímos dois caminhos experimentais: a destilação parcial, que mantém o vocabulário do aluno, e a destilação total, que faz o aluno adotar integralmente o tokenizador e o vocabulário do professor.

2 Methodology

A metodologia utilizada baseou-se nos princípios clássicos de destilação, com rápidas iterações experimentais. Inicialmente, foram consideradas duas estratégias distintas. Na destilação parcial, mantivemos o tokenizador original do BERT-tiny, que é baseado em inglês. O professor, BERTimbau-base, fornecia os embeddings de referência, e o aluno aprendia por meio de dois componentes de perda: Masked Language Modeling e similaridade de cosseno entre embeddings do professor e do aluno.

Na destilação total, adotamos uma abordagem mais fiel ao que é recomendado na literatura. O aluno foi reconstruído do zero, preservando apenas a arquitetura compacta do BERT-tiny, mas substituindo completamente seu vocabulário pelo do professor. O tokenizador passou a ser o mesmo do BERTimbau. Como a dimensão do embedding do professor (768) e a do aluno (128) são diferentes, adicionamos ao aluno uma camada linear de projeção para que as representações do professor pudessem ser aproximadas pelo espaço vetorial reduzido do aluno. Essa técnica possibilita que o aluno aprenda de forma muito mais compatível com o professor.

Após definir a arquitetura, combinamos as três perdas (MLM, KD e Cosine) com pesos específicos e treinamos o modelo por três épocas completas no corpus limpo. O treinamento foi conduzido utilizando GPUs de alto desempenho, permitindo que milhões de sentenças fossem processadas.

3 Dataset

Para realizar a destilação, utilizamos o corpus brWaC, composto por mais de 3,5 milhões de textos de páginas brasileiras da web. Por ser um dataset ruidoso, realizamos uma etapa rigorosa de pré-processamento. Removemos HTML, URLs, metadados, repetições e entradas extremamente curtas ou extremamente longas. Após essa limpeza, o conjunto foi reduzido para aproximadamente 3,25 milhões de textos, preservando apenas aqueles com conteúdo linguístico relevante.

Esse corpus foi então dividido em 90% para treinamento e 10% para teste, formando a base para avaliar a capacidade do modelo em prever tokens mascarados em sentenças nunca vistas. A diversidade temática e a quantidade massiva de dados tornaram o brWaC ideal para treinar um modelo destinado a compreender português de forma geral.

4 Experiments

Os experimentos foram conduzidos separadamente para as abordagens de destilação parcial e total, aplicadas sobre o corpus brWaC previamente limpo e dividido em 90% para treino e 10% para teste. O objetivo desta seção é comparar o desempenho dos modelos em tarefas de avaliação intrínseca e extrínseca, buscando entender como cada estratégia impacta o aprendizado linguístico e semântico do estudante.

Na destilação parcial, preservamos o tokenizador e o vocabulário originais do BERT-tiny em inglês, bem como seus pesos iniciais. O professor forneceu apenas representações internas, e o aluno foi treinado combinando Masked Language Modeling com similaridade de cosseno. Apesar do desalinhamento entre vocabulários, o modelo parcialmente destilado apresentou excelente desempenho em MLM, alcançando perda de **1.9433**, valor próximo ao do professor (**1.7569**) e substancialmente melhor que o tiny original (**5.7462**). Em Similaridade Textual (ASSIN2 – Pearson), o modelo atingiu **0.5763**, superando o tiny original (**0.5262**) e aproximando-se do professor (**0.6139**), indicando que parte da estrutura semântica global foi mantida. Em Análise de Sentimentos (TweetSentBR), obteve F1 de **0.2121**, levemente superior ao tiny original (**0.1905**) e inferior ao BERTimbau-base (**0.2778**). Já na tarefa de entailment (ASSIN2 – RTE), o desempenho foi notavelmente alto, com F1 de **0.4059**, superando tanto o tiny original (**0.2183**) quanto o professor (**0.1983**). Esse resultado sugere que, embora limitado no vocabulário, o modelo parcial retém estruturas

semânticas aprendidas previamente em larga escala no inglês, favorecendo tarefas que dependem de inferência lógica.

Na destilação total, reconstruímos o aluno adotando o tokenizador e o vocabulário do BERTimbau-base, adicionando uma camada de projeção para compatibilizar as dimensões de embedding. Nesse processo, utilizamos perdas de MLM, KL-divergence com temperatura 2.0 e similaridade de cosseno. O modelo totalmente destilado apresentou perda de MLM de **3.3025**, melhor que o tiny original, mas inferior ao modelo parcialmente destilado e mais distante do professor. Em Similaridade Textual, atingiu Pearson de **0.4400**, valor inferior ao tiny original e ao professor, refletindo a necessidade de reconstruir o espaço semântico a partir do zero. Em Análise de Sentimentos, o modelo alcançou F1 de **0.2000**, abaixo do tiny original (**0.3583**) e acima do professor (**0.1667**). No entailment, obteve F1 de **0.2225**, desempenho próximo ao tiny original, mas bem inferior ao modelo parcialmente destilado.

De forma geral, os resultados mostram que cada abordagem favorece aspectos diferentes do aprendizado. A destilação parcial, ao preservar parte das representações semânticas do tiny original, apresenta melhor desempenho em inferência e similaridade, além do melhor resultado entre os estudantes na tarefa de RTE. Já a destilação total produz um modelo mais alinhado ao português do ponto de vista morfossintático, mas com semântica ainda limitada devido à reconstrução completa do vocabulário. Em síntese, a abordagem parcial retém melhor o conhecimento semântico, enquanto a total é mais fiel ao idioma, embora precise de pré-treinamento adicional para alcançar profundidade semântica comparável.

5 Conclusion

O BERT-tiny-PT alcançou os objetivos essenciais do projeto: tornou-se capaz de compreender o idioma português e apresentou desempenho ao modelo original em inglês em Masked Language Modeling. Ainda assim, a transferência completa da capacidade semântica não ocorreu. As análises demonstram que a destilação, embora eficaz na transferência de conhecimento superficial, não é suficiente para garantir desempenho satisfatório em tarefas que exigem raciocínio semântico mais complexo.

O comportamento observado reforça que destilar um modelo pequeno a partir de um modelo grande exige mais do que imitar logits e embeddings; exige estratégias específicas para preservar semântica profunda e considerar as limitações arquiteturais do student.

6 Future Work

Trabalhos futuros devem explorar um pré-treinamento mais extenso, especialmente com mais épocas.