Proposta de Projeto TPF-01: Análise e Reprodução do Artigo "DistilBERT, a distilled version of BERT"

Henrique Alves
Disciplina de Redes Neurais
Universidade Federal de Viçosa - Campus Florestal
Florestal, MG
henrique.a.campos@ufv.br

1 Introdução

O presente projeto tem como objetivo analisar e reproduzir os experimentos de *fine-tuning* descritos no artigo "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter" (1). O trabalho original aborda um dos principais desafios dos modelos baseados em Transformers, como o BERT: o elevado custo computacional e de memória exigido para seu treinamento e inferência.

A motivação central do artigo está na busca por arquiteturas mais leves e eficientes, capazes de operar em dispositivos com recursos limitados — como smartphones — ou em aplicações em larga escala de maneira mais econômica. Nesse contexto, surge o **DistilBERT**, um modelo 40% menor e aproximadamente 60% mais rápido que o BERT, mantendo mais de 97% de sua performance original. O método proposto utiliza o conceito de *knowledge distillation* (destilação de conhecimento) durante o pré-treinamento, no qual o modelo aluno DistilBERT aprende a reproduzir as representações internas e as previsões do modelo professor BERT.

Para demonstrar sua eficiência, o artigo realiza o *fine-tuning* do DistilBERT em diversas tarefas do conjunto de *benchmark* **GLUE** (2). No presente projeto, será conduzida a replicação de um desses experimentos, especificamente na tarefa **SST-2 (Stanford Sentiment Treebank v2)**, que consiste em uma classificação binária de sentimentos (positivo ou negativo) em críticas de filmes.

A implementação utilizará o modelo pré-treinado distilbert-base-uncased e o conjunto de dados glue/sst2, ambos disponibilizados pela biblioteca **Hugging Face Transformers**[1] (3).

2 Cronograma

O cronograma para a execução deste projeto está organizado em quatro semanas, conforme detalhado na Tabela 1. O planejamento visa desde a configuração inicial e entrega desta proposta até a análise final dos resultados obtidos.

[1] Repositório do modelo: `https://github.com/huggingface/transformers`

Dataset (GLUE/SST-2): `https://huggingface.co/datasets/glue`

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Cronograma de execução do projeto.

Semana	Tarefas Planejadas
Semana 1 (27/10 - 05/11)	Finalização e entrega da Proposta (TPF-01). Configuração do ambiente no Google Colab. Instalação das bibliotecas.
Semana 2 (06/11 - 12/11)	Estudo do tutorial de *fine-tuning* do DistilBERT. Download do modelo 'distilbert-base-uncased' e seu Tokenizer. Download, pré-processamento e tokenização do dataset GLUE/SST-2.
Semana 3 (13/11 - 19/11)	Execução do script de *fine-tuning* no Google Colab GPU T4. Monitoramento das métricas de treinamento. Execução do script de avaliação no conjunto de validação do SST-2.
Semana 4 (20/11 - 02/12)	Coleta da acurácia final e organização dos resultados. Comparação da acurácia obtida com os resultados do *benchmark* GLUE (2). Preparação do relatório final do projeto e da apresentação.

References

[1] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv:1910.01108.

[2] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). *GLUE: A Multi-Task Natural Language Understanding Benchmark*. arXiv:1804.07461.

[3] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv:1910.03771.