Análise de Performance dos Modelos Gerais de Aprendizado de Máquina Pré-Treinados: BERT vs DistilBERT

Rafael Silva Barbon[1], **Ademar Takeo Akabane**[1]

[1] Faculdade de Engenharia de Computação
Pontifícia Universidade Católica de Campinas (PUC-Campinas) – Campinas, SP – Brazil

rafael.sb2@puccampinas.edu.br, ademar.akabane@puc-campinas.edu.br

Resumo. *Modelos de aprendizado de máquina (AM) vêm sendo amplamente utilizados devido à elevada quantidade de dados produzidos diariamente. Dentre eles, destaca-se os modelos pré-treinados devido a sua eficácia, porém estes normalmente demandam um elevado custo computacional na execução de sua tarefa. A fim de contornar esse problema, técnicas de compressão de redes neurais vem sendo aplicadas para produzir modelos pré-treinados menores sem comprometer a acurácia. Com isso, neste trabalho foram utilizados dois diferentes modelos pré-treinados de AM: BERT e DistilBERT na classificação de texto. Os resultados apontam que modelos menores apresentam bons resultados quando comparados com seus equivalentes maiores.*

Abstract. *Machine learning (ML) models have been widely used due to the high amount of data produced daily. Among them, the pre-trained models stand out due to their effectiveness, but normally these demand a high computational cost in the execution of their tasks. To circumvent this problem, neural network compression techniques have been applied to produce smaller pre-trained models without compromising accuracy. Therefore, in this work two different pre-trained models of ML were used: BERT and DistilBERT in text classification. The results show that smaller models present good results when compared to their larger counterparts.*

1. Introdução

Sabe-se que a adesão dos sistemas computacionais está em diversas áreas do conhecimento seja ela nas áreas das humanas, das exatas e das biológicas. Consequentemente, isso contribui com o aumento acelerado na geração, consumo e na transmissão dos dados na rede global. De acordo com o estudo realizado pela Statista Research Department [Statista Research Department 2018], em 2018, a quantidade total de dados criados, capturados e consumidos no mundo foi de 33 zettabytes (ZB) – equivalente a 33 trilhões de gigabytes. Já em 2020 cresceu para 59 ZB e está previsto para atingir 175 ZB até 2025.

Nesse cenário, tarefas ligadas a análise automática de dados textuais utilizando modelos de AM pré-treinados vêm ganhando grande importância. Entre elas destaca-se a classificação de texto, a filtragem de mensagens e a classificação automática de documentos, entre outras aplicações. Um dos modelos que se pode destacar é o BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2018]. O BERT foi proposto pelos pesquisadores do Google AI Language e se baseia em um modelo de representação de linguagem projetado para pré-treinar representações bidirecionais profundas de texto não rotulado. No entanto, este modelo possui um elevado custo computacional, principalmente em tempo de processamento e memória, muitas vezes imposibilitando sua implementação em sistemas com baixo poder de processamento e

memória [Sanh et al. 2019]. Com intuito de lidar com esse problema, outros modelos utilizam as técnicas de compressão de redes neurais convolucionais, por exemplo o DistilBERT (*distilled version of BERT*) [Sanh et al. 2019]. Em poucas palavras, DistilBERT aplica a técnica de compressão em que um pequeno modelo é treinado para reproduzir um comportamento do modelo equivalente maior.

O objetivo deste trabalho é avaliar o desempenho dos classificadores de texto produzidos a partir da compressão de redes neurais que representam modelos gerais pré-treinados de língua natural (distilBERT). Essa avaliação foi feita comparando os modelos comprimidos com o modelos original (BERT). Como resultado dos experimentos, pode-se destacar que modelo destilado apresentou resultados significativos: *(i)* redução do tempo de processamento em cerca de 50 %; e *(ii)* acurácia bem semelhante aos seus modelos maiores.

O restante deste artigo está organizado da seguinte maneira. Na Seção 2 é apresentada o referencial teórico necessário para a compreensão do trabalho desenvolvido. As avaliações dos resultados dos experimentos são descritas na Seção 3. Por fim, na Seção 4 é apresentada a conclusão e os trabalhos futuros.

2. Referencial Teórico

2.1. BERT e a Arquitetura Transformers

O BERT (*Bidirectional Encoder Representations from Transformers*) foi apresentado em [Devlin et al. 2018] por pesquisadores da Google AI Language, tendo causado grande impacto na comunidade de AM pelos resultados obtidos, definindo um novo patamar para o estado da arte. Modelos pré-treinados como o BERT, já passaram por um processo de treinamento não supervisionado com uma grande corpora de dados. Dessa forma, estes modelos necessitam apenas de um ajuste fino (*fine tunning*) utilizando conjuntos de dados de um domínio de interesse para produzir um classificador para uma determinada tarefa. Ao invés de um intenso processo de treinamento, além de não necessitar de um conjunto de dados grande e robusto.

A principal inovação do BERT é a aplicação do treinamento bidirecional da arquitetura Transformer, que utiliza um mecanismo de atenção para a modelagem de língua natural. Tal mecanismo aprende as relações contextuais entre palavras em um texto, por meio das análises das palavras adjacentes em ambas as direções. Assim, esse mecanismo permite uma maior compreensão do contexto de uma palavra e também do texto como um todo [Luong et al. 2015]. Em sua forma original, o Transformer inclui dois mecanismos separados: *(i)* um codificador que lê a entrada de texto; e *(ii)* um decodificador que produz uma previsão para a tarefa. Vale destacar que o objetivo do BERT é gerar um modelo da língua, assim apenas o mecanismo do codificador é necessário [Devlin et al. 2018].

Essa bidirecionalidade permite obter um desempenho elevado mesmo utilizando amostras pequenas de texto [Devlin et al. 2018]. Ao contrário dos modelos unidirecionais que necessitam de uma quantidade maior de amostras para obter resultados satisfatórios pelo fato de analisar somente as palavras à esquerda ou à direita [Luong et al. 2015]. O codificador Transformer lê toda a sequência de palavras de uma vez. Portanto, é considerado bidirecional, embora seja mais preciso dizer que é não direcional. Esta característica permite que o modelo aprenda o contexto de uma palavra com base em todo o seu entorno (esquerda e direita da palavra).

2.2. Compressão de Redes Neurais Profundas

A técnica de compressão de redes neurais profundas utilizadas no desenvolvimento de modelos destilados, como o DestilBERT [Sanh et al. 2019], utiliza destilação de conhe-

cimento. A destilação do conhecimento é uma técnica de compressão na qual um modelo compacto – o aluno – é treinado para reproduzir o comportamento de um modelo maior – o professor – ou um conjunto de modelos [Hinton et al. 2015]. Esta técnica tem como objetivo criar um modelo menor que deve reproduzir as decisões de um modelo maior, como no caso um modelo BERT.

A principal ideia que envolve a compressão de um modelo é aproximá-lo da função gerada por um modelo mais robusto, lento e maior, mas com uma melhor performance [Sanh et al. 2019]. Diferente de uma função desconhecida, a função aprendida pelo modelo maior é fornecida e pode ser usada para classificar uma grande quantidade de pseudo dados, que mostram independentemente o valor de cada atributo dentro de sua distribuição. Um modelo rápido e compacto treinado com pseudo dados não correrá o risco de apresentar *overfitting* e irá se aproximar da função aprendida pelo modelo maior [Buciluǎ et al. 2006].

Redes Neurais tipicamente produzem probabilidade de classes utilizando uma camada de saída *softmax* que converte o $logit$, z_i, calculado para cada classe em uma probabilidade, q_i, comparando-o com os outros $logits$, conforme a Equação 1:

$$q_i = \frac{exp(\frac{z_i}{T})}{\sum_j exp(\frac{z_j}{T})} \tag{1}$$

onde T é a temperatura que normalmente utiliza o valor 1, porém utuilizando um valor maior de T, obtemos uma distribuição mais suave (*soft-targets*) sobre as classes.

Na forma mais simples de destilação, o conhecimento é transferido para o modelo destilado treinando-o com um conjunto de transferência. Além disso, utiliza-se uma distribuição suave para cada caso do conjunto de transferência que é produzido pelo modelo maior com um alto valor de T na sua *softmax*. O mesmo T com um valor elevado é utilizado para treinar o modelo destilado, mas após ser treinado é usada a temperatura 1 [Hinton et al. 2015]. Em temperaturas baixas, a destilação presta muito menos atenção na combinação entre os resultados da função $logit$ que são muito mais negativos que a média. Dessa forma, utilizando temperaturas maiores que 1, o modelo destilado extraí mais informações relevantes do conjunto de dados de treinamento [Hinton et al. 2015].

3. Avaliação Experimental

Os modelos BERT e seu modelo comprimido DistilBERT, utilizados para avaliação experimental, foram retirados da plataforma Hugging Face[1] em duas versões; *(i) cased*, que faz diferenciação de letras maiúsculas e minúsculas; e *(i) uncased*, que não fazem esta diferenciação. Todos os testes de execução foram executados na plataforma Google Colab[2].

Além disso foi utilizado o método de *K-fold cross validation*, que consiste em dividir o conjunto de dados de treinamento em n-pastas. Neste caso 5 pastas (5-*fold cross validation*), sendo 4 delas (80%) utilizadas para o processo de ajuste fino e 1 delas (20%) para validação. O processo é repetido, descartando o modelo treinado em cada uma das vezes, fazendo com que todas as partes do conjunto de dados seja utilizada tanto para treinamento como para validação do modelo, comprovando a sua capacidade de generalização [Refaeilzadeh et al. 2009].

[1]https://huggingface.co/docs/transformers/index
[2]https://colab.research.google.com

A fim de comprovar o elevado custo computacional requerido para o processamento dos modelos, foi utilizado, além do Google Colab em sua versão gratuita, sua versão paga (Google Colab PRO). A versão PRO dispõe de uma GPU com maior poder de processamento e com uma maior disponibilidade de memória RAM. Dessa forma foi possível realizar a comparação dos tempos de processamento nas duas versões.

3.1. Conjunto de dados

Para treinamento e avaliação foram utilizados três conjuntos de dados. O primeiro conjunto utlizado foi o *brexit blog corpus* [Simaki et al. 2020], composto de 1682 frases retiradas de *blogs* associados ao Brexit divididos em 9 classes, apresentado na Tabela 1.

Tabela 1. Características do conjunto de dados *brexit blog corpus*.

ID	Classes	Número de exemplos
0	agreement/disagreement	50
1	certainly	84
2	contrariety	352
3	hypothetically	171
4	necessity	204
5	prediction	252
6	source of knowledge	287
7	tact/rudeness	44
8	uncertainty	196
9	volition	42
	Total	**1682**

O segundo, denominado de *bbc-text*, retirado da plataforma Kaggle[3] e origidado do BBC News [Greene and Cunningham 2006], contém 2225 comentários classificados em 5 classes apresentadas na Tabela 2.

Tabela 2. Características do conjunto de dados *bbc-text*.

ID	Classes	Número de exemplos
0	business	510
1	entertainment	386
2	politics	417
3	sport	511
4	tech	401
	Total	**2225**

E o terceiro conjunto de dados foi o *amazon alexa reviews*, também retirado da plataforma Kaggle, possui 3150 comentários avaliativos da assistente virtual da Amazon: Alexa, distribuídos de acordo com a Tabela 3.

3.2. Resultados

A Tabela 4 apresenta os resultado obtidos em termos de acurácia, que indica uma performance geral do modelo utilizando o número de predições corretas divido pelo número total de predições (Equação 2), com os diferentes modelos e conjunto de dados utilizados.

[3]https://www.kaggle.com/

Tabela 3. Características do conjunto de *amazon alexa reviews*.

ID	Classes	Número de exemplos
0	positive	2893
1	negative	257
	Total	**3150**

$$Acuracia = \frac{\sum PredicoesCorretas}{\sum Predicoes} \tag{2}$$

Tabela 4. Acurácia de cada modelo para os diferentes conjuntos de dados.

MODELOS	AMAZON ALEXA REVIEWS	BBC-TEXT	BREXIT CORPUS
BERT-Cased	0,956	0,974	0,412
BERT-Uncased	0,955	0,976	0,391
DistilBERT-Cased	0,955	0,977	0,385
DistilBERT-Uncased	0,950	0,981	0,366
RNC	0,905	0,677	0,357

Como pode ser observado, para o conjunto de dados *brexit blog corpus*, os modelos *cased* obtiveram acurácias superiores aos modelos *uncased*. Em ambas as versões (*cased* e *uncased*) o BERT obteve um *score* mais alto que o DistilBERT. O DistilBERT-Uncased apresentou um desempenho 11% menor que o BERT-Cased. Além disso, é possivel observar que a avaliação com o conjunto de dados *brexit corpus* obteve uma acurácia modesta em relação aos outro modelos. Este fato pode ser explicado devido ao conjunto de dados ser pequeno e possuir muitas classes, apresentando poucos exemplos de cada classe.

Analisando a avaliação para o *bbc-text*, há uma diferença pequena entre os modelos avaliados, porém diferentemente do observado no conjunto de dados *brexit blog corpus*, o modelo DistilBERT-Uncased obteve a maior acurácia, mostrando o potencial dos modelos destilado, que podem até obter resultados ligeiramente superiores que seus modelos originais.

Para o conjunto de dados *amazon alexa reviews*, o BERT-Cased obteve o melhor resultado, ligeiramente superior aos produzidos pelo BERT-Uncased e DistilBERT-Cased. Esse resultado pode se dar pela importância da distinção de letras maiusculas e minusculas, já que se trata de um conjunto de dados com avaliações de um produto.

Além disso, como esperado, os modelos com arquitetura Transformer apresentaram resultados expressivamente superiores em termos de acurácia quando comparados com os obtidos pelas redes neurais convolucionais (RNC). É possível também observar a eficácia da técnica de compressão de redes neurais profundas, produzindo um modelo de menores dimensões sem afetar a acurácia do modelo original.

Na Tabela 5 são apresentados os recursos computacionais demandados pelos modelos maiores, no caso o BERT, e de seu modelo destilado DistilBERT, de acordo com o seu tempo de processamento. É possível observar por meio da Tabela 5, uma redução de certa de 50% no tempo de processamento dos modelos comprimidos, em relação aos modelos maiores. E também a diferença entre o tempo de processamento dos modelos

pré-treinado com relação aos modelos baseados em rede neurais convolucionais. Modelos baseados em rede neurais convolucionais, quando comparado com modelos pré-treinados, requerem o ajuste de uma quantidade muito menor de parâmetros durante seu treinamento, demandando menos tempo de processamento.

Tabela 5. Tempo de processamento (treinamento + validação) em segundos.

MODELOS	AMAZON ALEXA REVIEWS	BBC-TEXT	BREXIT CORPUS
BERT-Cased	806,02	712,40	609,00
BERT-Uncased	765,40	655,00	634,40
DistilBERT-Cased	447,60	317,20	227,20
DistilBERT-Uncased	419,80	370,80	193,00
RNC	104,10	188,80	178,44

A fim de obter uma visualização mais exata da redução do tempo de processamento dos modelos comprimidos, a Tabela 6 apresenta a redução percentual do tempo de execução do DistilBERT com relação ao BERT.

Tabela 6. Redução percentual do tempo de execução do DistilBERT vs BERT.

	AMAZON ALEXA REVIEWS	BBC-TEXT	BREXIT CORPUS
CASED-Cased	44,5%	55,5%	62,7%
UNCASED-Uncased	45,2%	43,48%	69,6%

Além do tempo de processamento, a compressão de redes neurais profundas produz também modelos bem menores quando comparado com os modelos não comprimidos. A Tabela 7 apresenta o tamanho dos modelos gerados em *MegaBytes* (MB).

Tabela 7. Tamanho dos modelos gerados (MB).

MODELOS	AMAZON ALEXA REVIEWS	BBC-TEXT	BREXIT CORPUS
BERT-Cased	826	826	827
BERT-Uncased	835	835	835
DistilBERT-Cased	502	502	502
DistilBERT-Uncased	511	511	511

Ao comparar os modelos produzidos nota-se que os baseados no BERT, no DistilBERT e na RNC têm tamanhos muito distintos como mostrado na Tabela 8. Essa diferença em tamanho, além de acarretar distintos requisitos de memória para armazenar os classificadores, está associada também a maiores tempos de treinamento para os modelos maiores. Por exemplo, um modelo BERT tem cerca de 110 milhões de parâmetros que tem que ser ajustados durante o seu treinamento, enquanto que o DistilBERT tem cerca de 66 milhões de parâmetros [Sanh et al. 2019].

Se, por um lado, os modelos BERT e DistilBERT requerem mais memória e tempo de processamento para serem treinados, por outro, apresentaram resultados expressivamente superiores em termos de acurácia quando comparados com os modelos baseados apenas em redes neurais convolucionais, conforme mostrado na Tabela 4.

Tabela 8. Redução percentual do tamanho dos modelos DistilBERT com relação ao BERT.

	AMAZON ALEXA REVIEWS	BBC-TEXT	BREXIT CORPUS
CASED-Cased	39,2%	39,2%	39,2%
UNCASED-Uncased	38,8%	38,8%	38,8%

Com o intuito de avaliar os recursos computacionais demandados por esses modelos, foi utilizado o Google Colab PRO, versão paga da plataforma que dispõe de GPUs com maior poder de processemento e maior memória RAM. A versão PRO utiliza a GPU T4 e P100 enquanto a versão gratuita utiliza GPU K80s. Os modelos passaram pelo mesmo processo de ajuste fino e treinamento (para RNC) com os mesmos parametros e conjunto de dados utilizados na versão gratuita. O tempo de processamento em segundos está apresentado na Tabela 9 e a redução percentual do tempo de execução do DistilBERT com relação ao BERT na Tabela 10.

Tabela 9. Tempo de processamento (treinamento + validação) em segundos (Google Colab PRO)

MODELOS	AMAZON ALEXA REVIEWS	BBC-TEXT	BREXIT CORPUS
BERT-Cased	519,20	223,00	322,60
BERT-Uncased	515,80	421,80	328,20
DistilBERT-Cased	231,20	132,40	136,20
DistilBERT-Uncased	197,00	197,60	174,20
RNC	106,06	56,57	36,60

Tabela 10. Redução percentual do tempo de execução do DistilBERT com relação ao BERT (Google Colab PRO)

	AMAZON ALEXA REVIEWS	BBC-TEXT	BREXIT CORPUS
CASED-Cased	55,5%	40,6%	57,8%
UNCASED-Uncased	61,8%	53,2%	46,9%

O que era de esperar, o tempo de execução dos modelos utilizando um *hardware* com maior poder de processamento é, via de regra, bem menor que quando comparado com um *hardware* mais simples.

4. Conclusão

Os experimentos realizados mostraram que a técnica de compressão de redes neurais denominada de destilação de conhecimento, utilizada para a produção do DistilBERT, permite produzir modelos de classificação de texto em torno de 40% menores que os produzidos pela arquitetura original. Além disso, possuem um tempo cerca de 50% menor, sem comprometer a acurácia dos modelos destilados, ou com uma ligeira redução, quando comparados aos modelos não destilados. Por outro lado, conforme esperado, modelos baseados em Redes Neurais Convolucionais requerem menos memória e tempo de execução, à custa de uma redução, em geral expressiva, da acurácia.

A técnica de compressão de redes produzem modelos que por demandarem menos poder de processamento e memória, podem ser aplicados em sistemas mais simples.

Um estudo da implementação destes modelos em sistemas mais simples podem ser estudadas em trabalhos futuros. Além disso, novos modelos pré-treinados e mais robusto está sendo amplamente estudados e desenvolvidos como o GPT-2 [Radford et al. 2019] e GPT-3 [Dale 2021], que poderão ser avaliados e comparados com os modelos citados neste artigo em trabalhos futuros. Uma aplicação muito interessante destes modelos mais modernos podem ser vistos no desenvolvimento do CODEX [Chen et al. 2021], um modelo que está em desenvolvimento e que gera código de programação através de linguagem natural.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Dale, R. (2021). Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Greene, D. and Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proc. 23rd International Conference on Machine learning (ICML'06)*, pages 377–384. ACM Press.

Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, 5:532–538.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Simaki, V., Paradis, C., Skeppstedt, M., Sahlgren, M., Kucher, K., and Kerren, A. (2020). Annotating speaker stance in discourse: the brexit blog corpus. *Corpus Linguistics and Linguistic Theory*, 16(2):215–248.

Statista Research Department (2018). Amount of data created, consumed, and stored 2010-2025. `https://www.statista.com/statistics/871513/worldwide-data-created/`, Último acesso: 08/04/2022.