Parâmetros pré-treinados do Wav2Vec2 auxiliam BERT em tarefas de texto?

Talles Viana e Vitor Luquezi

Dezembro 2021

Abstract

Outros artigos já demonstraram que pré-treinos do BERT utilizando textos não necessariamente estruturados trazem benefício para o *finetune* em tarefas específicas de texto. O pré-treino em áudio (domínio descorrelato), como no Wav2Vec2, também traria algum benefício? Esse trabalho traz alguns experimentos comparando a utilização dos pesos do Wav2Vec2 no BERT versus pesos randômicos (Xavier) para *finetune* em tarefas de texto. Os experimentos propostos não trouxeram nenhuma evidência clara de transferência de conhecimento, porém é um estudo inicial que pode servir como base para futuros trabalhos.

1 Introdução

Atualmente, os modelos estado da arte em NLP como BERT [3] e o T5 [6] necessitam de um pré-treinamento utilizando dados que pertençam ao domínio de texto para serem *fine-tuned* para alguma tarefa específica no mesmo domínio. Porém, alguns estudos [2] [5] demonstram que mesmo um pré-treino utilizando dados de domínios não correlatos, como textos gerados algoritmicamente, códigos fonte ou mesmo sequências musicais podem trazer algum benefício para o modelo. Portanto, este trabalho se propõe a investigar se os parâmetros do modelo Wav2Vec2 [1], que são pré-treinados à partir de sinais de áudio (mais especificamente, de fala, sendo utilizado em tarefas de transcrição), podem de alguma forma auxiliar o BERT em tarefas específicas de texto.

A priori, para uma investigação inicial, o dataset de classificação de sentimento do IMDB[1] foi utilizado de forma a realizar comparações entre os modelos.

A hipótese levantada por este trabalho é verificar se e quanto uma inicialização (total ou parcial) de um modelo BERT utilizando pesos do modelo Wav2Vec2 auxilia no treinamento (ou fine-tuning) deste modelo em uma tarefa específica relacionada à texto, comparado com outros tipos de inicialização existentes. Este estudo visa contribuir em áreas como investigação da

[1] http://files.fast.ai/data/aclImdb.tgz

interpretabilidade de modelos de Processamento de Linguagem Natural, explorando a possibilidade de que há conhecimentos que podem ser transponíveis entre o domínio de áudio e texto.

2 Metodologia

A metodologia utilizada consiste em partir de uma arquitetura inicial, como BERT *base* e realizar modificações como a substituição de pesos do *encoder* por pesos de camadas correspondentes Wav2Vec2; esta substituição pode ocorrer em camadas específicas ou no *encoder* inteiro. Após esta substituição, realizar o fine-tuning do modelo em uma tarefa de texto específica. Além da substituição pelos pesos do Wav2Vec2, realizar reinicialização dos pesos usando estratégias como inicialização aleatória Xavier, ou de outros modelos baseados em BERT e pré-treinados em outros domínios (como VisionTransformer [4]), para se obter uma base de comparação. Como *baseline*, é utilizado um modelo de BERT pré-treinado em texto em inglês (mesmo domínio da tarefa proposta).

A partir dos resultados obtidos para cada inicialização, é possível realizar comparações como: acompanhamento da evolução do valor de *loss* nos dados treino e validação, o que pode indicar o quão rápido o treino do modelo converge, além de métricas alvo calculadas sobre o conjunto de dados de validação e especialmente de teste, comparando com os resultados obtidos com o modelo *baseline*.

3 Dados

O dataset de classificação de sentimentos do IMDB[2] foi escolhido como ponto de partida por ser um conjunto de dados não tão grande, que possibilita rápidas iterações e avaliações dos modelos. O conjunto foi dividido em 20000 amostras para treinamento, 5000 amostras para validação e 25000 amostras para teste.

Durante o último período desse trabalho, sugeriu-se a utilização e avaliação dos modelos na tarefa de MNLI do GLUE (*General Language Understanding Evaluation*), porém, devido ao tempo limitado, não foi possível tal avaliação, ficando assim para um futuro trabalho.

4 Experimentos

Todos experimentos utilizaram a versão Pro do Google Colab[3], assim como o dataset do IMDB previamente apresentado. Para fins de comparação da performance dos modelos na tarefa de classificação de sentimento, a métrica utilizada é a acurácia, ou seja, qual a taxa de predições corretas do modelo. Como mencionado na seção anterior, o modelo utilizado como *baseline* é um

[2] http://files.fast.ai/data/aclImdb.tgz
[3] https://colab.research.google.com/

BERT *base* ou *large*, dependendo do experimento, pré-treinado no domínio do texto [4] [5] e *fine-tuned* no dataset do IMDB, ou seja, pré-treino com mesmo domínio. Nos diversos experimentos, pequenas alterações são feitas nos pesos das camadas ou no método de treinamento, e após cada alteração, o modelo é avaliado no conjunto separado de testes.

4.1 Pesos de uma camada

Primeiramente, como teste rápido, para verificarmos grosseiramente a influência de alterações dos pesos no BERT, realizamos o treinamento e *finetune* do *baseline*. Depois, partindo desse *baseline finetuned*, apenas uma *head* do Transformer dele é alterada, onde carregamos os parâmetros do Wav2Vec2 e de inicializações randômicas, de forma a verificar quanto a acurácia do modelo será afetada. Nesse caso, após a alteração, não há re-treino, apenas a avaliação no conjunto de testes. Esperava-se um redução mínima ao se utilizar os parâmetros do Wav2Vec2 e uma redução mais significante ao se utilizar qualquer inicialização randômica. As métricas avaliadas após cada alteração pode ser vista na Tabela 1, onde vemos que a inicialização Xavier foi a que trouxe o menor perda de performance para o modelo.

Modelo	Alteração	Acurácia
Baseline	n/a	92,02%
-	Pesos do Wav2Vec2-base na *head* 12	82,74%
-	Pesos do Wav2Vec2-base-960h na *head* 12	82,78%
-	Inicialização Xavier na *head* 12	89,30%
-	Inicialização Uniforme na *head* 12	46,93%
-	Inicialização Normal na *head* 12	51,89%

Table 1: Comparação da acurácia do modelo *baseline* com o mesmo modelo diferentes alterações nos pesos da *head* 12 do Transformer do modelo

4.2 Pesos do *Transformer*

Para este experimento, fora realizada a substituição de todos os pesos do *encoder* do BERT base por pesos das arquiteturas de Wav2Vec2-base[6] e Wav2Vec2-base-960h[7], além da adição de camadas densas no final da rede para a tarefa de classificação. Fora realizado o *fine-tuning* do modelo para a tarefa de classificação em questão, com parâmetros de *learning rate* 5.10^{-5}, *batch size* de 16 com acúmulo em 4 *batches*, tamanho da sequência truncado em 256 caracteres, otimizador AdaBelief, duração de 10 épocas (desconsiderando *early stop*). Durante o processo foram realizados testes alterando os parâmetros citados para verificar sua

[4] https://huggingface.co/bert-base-uncased
[5] https://huggingface.co/bert-large-uncased
[6] https://huggingface.co/facebook/wav2vec2-base
[7] https://huggingface.co/facebook/wav2vec2-base-960h

influência na convergência do treinamento, entretanto, para os resultados reportados, foram mantidos os mesmos parâmetros e *seed* para fins de comparação. Assim como no experimento anterior, fora realizada também a reinicialização dos pesos com o métodos Xavier.
Os resultados da acurácia de cada modelo após *fine-tuning* são apresentados na Tabela 2, onde observamos que a utilização dos pesos do Wav2Vec2 não trouxe benefício direto para o *fine-tune* na tarefa de classificação. Observamos que a inicialização randômica utilizando Xavier trouxe mais benefício do que os pesos do Wav2Vec2, possivelmente indicando um grande desalinhamento entre os pesos do Wav2Vec2 pré-treinado em voz, em relação aos pesos do BERT necessários para o domínio do texto.

Modelo	Alteração no *encoder*	Acurácia
Baseline	n/a	92,02%
-	Utilização dos pesos do Wav2Vec2-base	80,92%
-	Utilização dos pesos do Wav2Vec2-base-960h	83,16%
-	Pesos inicializados usando Xavier	85,64%

Table 2: Comparação da acurácia do modelo *baseline* com modelos com diferentes substituições de pesos no *encoder* do BERT

4.3 Congelamento

Como medida encontrada na literatura para a melhoria do aproveitamento do conhecimento em tarefas de *transfer learning*, foi implementado o congelamento de pesos do *encoder* durante uma fase inicial do treinamento. Este congelamento visa forçar um alinhamento de domínios entre os vetores de *embeddings* e o próprio *encoder* antes de alterar seus pesos durante o treino. Foram testados congelamentos de pesos por 300 e 600 passos no treino, sendo que 300 passos equivale a aproximadamente uma época.
Os resultados podem ser encontrados nas Tabelas 3 e 4, onde pelas observações nenhum ganho foi observado ao se utilizar a técnica de congelamento nas passos iniciais do treinamento.

Alteração	Acurácia
Sem congelamento	84,94%
Congelamento de 300 passos	84,41%
Congelamento de 600 passos	83,78%

Table 3: Comparação da acurácia do modelo BERT com substituição pelos pesos de Wav2Vec2, para diferentes durações de congelamento do *encoder*

Modelo	Alteração no *encoder*	Acurácia
Baseline	n/a	91,34%
-	Utilização dos pesos do Wav2Vec2-base	83,45%
-	Pesos inicializados usando Xavier	84,68%

Table 4: Comparação da acurácia do modelo *baseline* com modelos com diferentes substituições de pesos no *encoder* do BERT, aplicando o congelamento nos 600 passos iniciais

4.4 Variações na taxa de aprendizado

Esse experimentou visou, assim como no congelamento, tentar adotar alguma estratégia com o *learning rate* que pudesse de alguma forma trazer algum benefício para utilização dos pesos do Wav2Vec2 no BERT. Assim, fizemos a implementação do treinamento utilizando a) *warm-up* da *learning rate*, ou seja, durante a primeira época, ir incrementando gradualmente a *lr* até atingir o valor de 5×10^{-5}. E após a primeira época, b) o decaimento cossenoidal da *lr*. Como se pode observar na Tabela 5, tais estratégias também não beneficiaram o modelo que utiliza os pesos do Wav2Vec2 se comparado com a inicialização randômica utilizando Xavier.

Modelo	Alteração no *encoder*	Acurácia
Baseline	n/a	90,95%
-	Utilização dos pesos do Wav2Vec2-base	84,92%
-	Pesos inicializados usando Xavier	85,68%

Table 5: Comparação da acurácia do modelo *baseline* com modelos com diferentes substituições de pesos no *encoder* do BERT, aplicando *warm up* da *learning rate*

4.5 Outros modelos baseados em BERT

Para ampliar a base de comparação, foram selecionadas duas outras arquiteturas que contém BERT em sua constituição para a realização da substituição dos pesos. Primeiramente, utilizou-se o VisionTransformer, ou ViT, um modelo treinado para tarefas com processamento de imagens, como origem dos pesos a serem transferidos ao BERT *base*, porém, após diversas tentativas (inclusive utilizando as diversas versões disponíveis), não foi possível concluir um treinamento com estabilidade, ou seja, sem que houvesse perda dos gradientes e convergência do modelo. Com isso, especula-se que o pré-treino em imagem leva a uma configuração muito distante da otimização para um problema em texto, possivelmente devido a alta discrepância entre os domínios imagem e texto.

Outra possibilidade explorada foi a utilização de pesos do BERTimbau[7], um modelo do BERT pré-treinado com textos em português (ou seja, não é um

domínio totalmente não relacionado pois trata-se de linguagem em texto), para a transferência de pesos. É possível observar que os resultados obtidos foram equivalentes, sendo que é possível que um treinamento com mais épocas ou com outros hiperparâmetros possa diferenciar melhor as abordagens.

Modelo	Alteração no *encoder*	Acurácia
Baseline	n/a	92,02%
-	Utilização dos pesos do BERTimbau-base	84,54%
-	Pesos inicializados usando Xavier	84,68%

Table 6: Comparação da acurácia do modelo *baseline* com modelos

5 Conclusão

A partir dos resultados dos experimentos realizados não foi possível observar evidências de que ocorreu transferência de conhecimento ao se utilizar os pesos do Wav2Vec2, pré-treinado em áudio, na aplicação de tarefas com texto. A performance atingida no final de cada experimento é um pouco menor ou equivalente à obtida com o uso de inicialização com o método Xavier. Em geral, observou-se que modelos inicializados com o método Xavier convergem (no sentido de atingir determinado valor de *loss* no treinamento) mais rapidamente do que modelos com pesos do Wav2Vec2.

A expectativa inicial do estudo era que ao se treinar modelos com a transferência de pesos fosse possível superar os métodos de inicialização randômica existentes, no caso, o método Xavier, porém este se mostrou relativamente robusto. Um dos principais desafios enfrentados ao tentar realizar o *fine-tuning* de modelos com pesos oriundos de domínios não relacionados é a convergência, pois observou-se em muitos casos a ocorrência de gradientes explodindo ou sumindo (os quais são problemas conhecidos de inicializações "ruins"). Alguns dos recursos utilizados para tentar garantir a convergência dos treinamentos e superar tais problemas foram: uso de *gradient clipping* (não teve efeito significativo), mudanças nos *batches* como aumento do tamanho do *batch* ou alteração da *seed* para outros dados (isso contribuiu para um treinamento levemente mais "controlado", porém não garantiu convergência), estratégias de alteração dinâmica da *learning rate*, entre outras; no entanto, esses métodos em geral não foram bem sucedidos e os resultados destes treinos permaneceram inexpressivos. Uma vantagem clara da inicialização com método Xavier é que ela foi desenvolvida para evitar os problemas com gradiente citados.

6 Trabalhos futuros

A realização desse trabalho utilizando o dataset do IMDB não trouxe evidências de transferência de conhecimento, logo, uma continuação iminente do trabalho

consiste em executar a mesma metodologia porém em diferentes tarefas de Processamento de Linguagem Natural, por exemplo, tarefas do GLUE. Tentando buscar alguma evidência de uma possível transferência de conhecimento em algura outra tarefa.

Outra continuação proposta é explorar as mesmas transferências de pesos porém em arquiteturas maiores, como a BERT *large*. A expectativa é que as diferenças entre modelos inicializados randomicamente e com pesos substituídos (total ou parcialmente) seja mais perceptíveis. Este trabalho foi iniciado, porém devido a limitações de *hardware* do Google Colab, não pode ser concluído como esperado. Treinos com a arquitetura BERT *large* demoram muito e eram derrubados inesperadamente (inclusive em diferentes momentos da execução de um mesmo código) pela plataforma devido consumo de memória RAM exigido, apesar do código otimizado.

Um outro problema sugerido para exploração seria a transferência de conhecimento entre vetores de *embeddings* de domínios não relacionados, especialmente entre texto e áudio (por exemplo, o mapeamento entre *tokens* de texto e fonemas no domínio do áudio), considerando que os *embeddings* trazem informações essenciais para um treinamento bem sucedido de um modelo, além de abrir novas possibilidades de treino de modelos multi domínio.

References

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[2] Cheng-Han Chiang and Hung-yi Lee. Pre-training a language model without human language. *arXiv preprint arXiv:2012.11995*, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Isabel Papadimitriou and Dan Jurafsky. Learning music helps you read: Using transfer to study linguistic structure in language models. *arXiv preprint arXiv:2004.14601*, 2020.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[7] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.