Otimizando Aplicações de Aprendizado Profundo na Cloud *

Thiago Araújo[1], **Cristiano Künas**[1], **Philippe Navaux**[1], **Mateus dos Reis**[3], **Beatriz Schaan**[12]

[1] Instituto de Informática – UFRGS, [2] HCPA, [3] Feevale

{tsaraujo, cakunas, navaux}@inf.ufrgs.br, bschaan@hcpa.edu.br

mateusaugustodosreis@gmail.com

Resumo. *Aprendizagem profunda (DL) demonstrou êxito na detecção de padrões em imagens e conjuntos tabulares, mas seu processo de treinamento é computacionalmente dispendioso e demorado. A computação em nuvem surge como uma alternativa econômica, aproveitando aceleradores como GPUs e TPUs. Este estudo aborda o treinamento de um modelo para predizer o encaminhamento de casos graves de retinopatia diabética a especialistas, utilizando dados de pacientes brasileiros e realizando experimentos na nuvem. A avaliação de desempenho indicou uma redução de 36% no tempo de execução com o uso de TPUs, ressaltando a atratividade da computação em nuvem em aplicações de DL, especialmente com hardware dedicado.*

1. Introdução e motivação

O treinamento de modelos, cada vez mais robustos e complexos, está transformando as plataformas computacionais de alto desempenho em ferramentas cruciais. Nesse contexto, os aceleradores de hardware emergiram como elementos essenciais para acelerar o treinamento de modelos de *Deep Learning* (DL), uma vez que esta é uma das tarefas mais intensivas em recursos computacionais e energéticos. Sem o suporte adequado de hardware, o treinamento pode se estender por vários dias. Contudo, a crescente demanda por esses aceleradores está elevando seus custos, tornando-os financeiramente inviáveis para muitos pesquisadores.

A computação em nuvem surge como uma alternativa viável aos custos elevados de hardware associados ao treinamento de modelos de DL. Com esse tipo de computação, você paga apenas pelo uso [Künas et al. 2023] e tem acesso a uma ampla gama de recursos e serviços computacionais. Além disso, os provedores de nuvem realizam atualizações regulares em seus recursos, incluindo GPUs e TPUs. Esses dois aceleradores são empregados no treinamento de modelos de inteligência artificial, porém, a TPU se destaca por ser otimizada especialmente para operações como multiplicação de matrizes e outras tarefas fundamentais em modelos de redes neurais profundas.

Neste trabalho, propomos utilizar a nuvem para acelerar o treinamento de um modelo de DL para predição do encaminhamento de pacientes para o médico especialista em casos de retinopatia diabética com maior gravidade. A avaliação de desempenho envolve o treinamento utilizando GPU e TPU.

2. Implementação e resultados

A predição do encaminhamento de casos graves de retinopatia diabética a especialistas já foi realizada utilizando a base de dados EyePACS, que possui imagens de pacientes indianos [Moreira et al. 2020]. Este trabalho utiliza dados anonimizados do hospital terciário de

*O presente trabalho foi apoiado por CAPES - Código de Financiamento 001, CNPq/MCTI/FNDCT - Universal 18/2021 sob número 406182/2021-3, pelos projetos CIARS RITEs/FAPERGS e CI-IA FAPESP-MCTIC-CGI-BR.

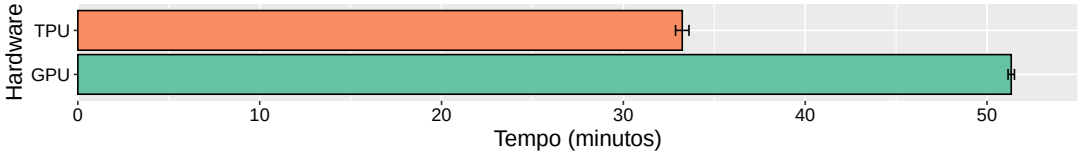

Figura 1. Valor médio dos tempos de treinamentos na cloud.

Porto Alegre, com o objetivo de obter métricas de desempenho melhores. Todas as imagens foram processadas conforme foi relatado na reprodução de Voets[Voets et al. 2019], localizando o centro e o raio do fundo do olho e redimensionando cada imagem para 299x299 pixels. O conjunto de dados de imagens foi convertido para o formato TFRecord, formato de dados personalizado do TensorFlow. O conjunto foi dividido em dois subconjuntos, treino e teste, em uma proporção de 70%/30% de maneira estratificada. O objetivo da rede neural foi produzir uma previsão binária para cada uma das imagens. Foram utilizados *callbacks* para coletar as métricas do modelo, reduzir a taxa de aprendizado e realizar pontos de verificação para salvar o melhor modelo. O algoritmo foi treinado por 200 épocas com batch de tamanho 8.

Os experimentos conduzidos na GPU foram realizados no Kaggle, onde foram empregadas 2 GPUs Nvidia T4. Por outro lado, a execução na TPU ocorreu no Google Colab. A Figura 1 apresenta o tempo médio de 10 treinamentos e o desvio padrão nesses ambientes. Destaca-se um substancial ganho ao utilizar a TPU, resultando em uma redução de tempo de 36%, diminuindo de $51,33$ minutos com as GPUs para $33,24$ minutos. É notável que o desvio padrão do tempo de execução na TPU foi o dobro do observado na GPU, uma diferença que pode ser atribuída ao compartilhamento de recursos na nuvem, especialmente ao considerarmos o uso de recursos gratuitos que podem ter sido compartilhados com outros usuários.

3. Conclusões e trabalhos futuros

A utilização de computação em nuvem revelou-se altamente promissora para o treinamento de modelos de aprendizado profundo. Destaca-se que o emprego de hardware projetado especificamente para acelerar aplicações de inteligência artificial, como as TPUs, apresentou desempenho significativamente superior em comparação ao uso de GPUs. Isso ocorre porque o modelo faz uso de imagens, e a TPU, por ser especialmente otimizada para operações matriciais, tende a apresentar um desempenho mais eficiente no treinamento com imagens em comparação com a GPU. Trabalhos futuros expandirão a avaliação de desempenho, explorando hardware mais recente, como a TPUv4.

Referências

Künas, C. A., Serpa, M. S., and Navaux, P. O. (2023). Exploiting Hardware Accelerators in Clouds. In *High Performance Computing in Clouds: Moving HPC Applications to a Scalable and Cost-Effective Environment*, pages 127–144. Springer.

Moreira, F., Schaan, B., Schneiders, J., Reis, M., Serpa, M., and Navaux, P. (2020). Impacto da resoluçao na detecçao de retinopatia diabética com uso de deep learning. In *Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 494–499. SBC.

Voets, M., Møllersen, K., and Bongo, L. A. (2019). Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PloS one*, 14(6):e0217541.