Estudos de Técnicas de Quantização e Poda em Redes Neurais sob Domínio Complexo

Palavras-Chave: Redes Neurais, Poda em Redes Neurais, Quantização

Autores/as:

Kairé Pereira Giovanetti – IC, UNICAMP

Prof. Dr. Dalton Soares Arantes – FEEC, UNICAMP

INTRODUÇÃO:

O desenvolvimento das grandes redes neurais permitiu o avanço na resolução de diversos problemas computacionais. Essas redes, no entanto, devido à sua complexidade, passam a exigir grandes quantidades de processamento computacional e tempo para se sustentarem, nesse sentido, se torna interessante encontrar jeitos para diminuir seus custos, o que leva ao desenvolvimento das técnicas de quantização e de poda de redes. Dessa forma, torna-se interessante ampliar o uso dessas técnicas para diferentes tipos de redes, nesse caso, as redes de domínio complexo.

A poda de redes neurais consiste no processo de remover pesos do cálculo, com o objetivo de tornar os cálculos da rede mais rápidos, mantendo a sua acurácia dentro de um limite seguro. Para realizar esse processo, podem ser utilizadas diversas táticas e métodos no domínio dos números reais. A partir disso, pode-se extrapolar esses métodos para o domínio complexo, permitindo a aplicação da poda em CVNN's.

Já o processo de quantização consiste em estabelecer uma precisão menor para os valores dos pesos da rede de modo a diminuir o espaço gasto, e também, os tempos de treinamento e processamento da rede.

METODOLOGIA:

Uma das técnicas mais simples para a realização do processo de poda é a poda por magnitude global (global magnitude pruning), que, no caso real, consiste em descartar os pesos que apresentam menor módulo. Além da sua simplicidade conceitual, essa técnica permite

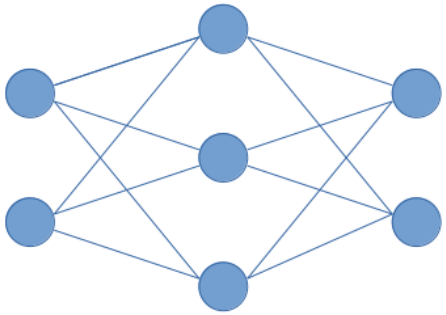

Figure 1: Exemplo de rede completamente conectada.

também a poda em redes já treinadas, não precisando de variáveis guardadas durante o processo de treinamento.

Quando consideramos sua aplicação para o caso complexo se torna necessário considerar o que fazer com o parâmetro de avaliação, já que, se tratando de números complexos, a fase pode ser tão, importante quanto o módulo. Dessa forma, é interessante considerar os efeitos na poda que surgem de considerar apenas o módulo, a fase, as partes inteiras ou reais, ou uma combinação dos dois.

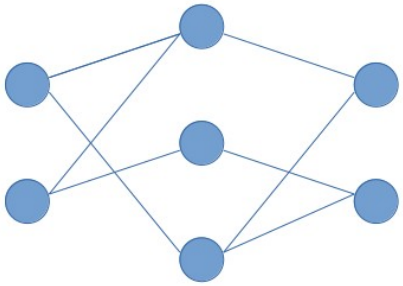

Figure 2: Exemplo da rede após o processo de poda, com alguns pesos desconectados

RESULTADOS E DISCUSSÃO

Para testes em redes de tamanho pequeno e médio, e utilizando primariamente a ideia por trás do método de poda por magnitude global, há algumas possibilidades de adaptação para o caso complexo. Nesse viés, observa-se que, ao se comparar a poda por módulo, a poda usando a fase como parâmetro, e as podas utilizando apenas a parte imaginária ou real do número, tem-se uma maior retenção da precisão da rede na poda por módulo. Ademais, as podas utilizando apenas as partes real ou imaginária também performam melhor do que a que considera apenas a fase.

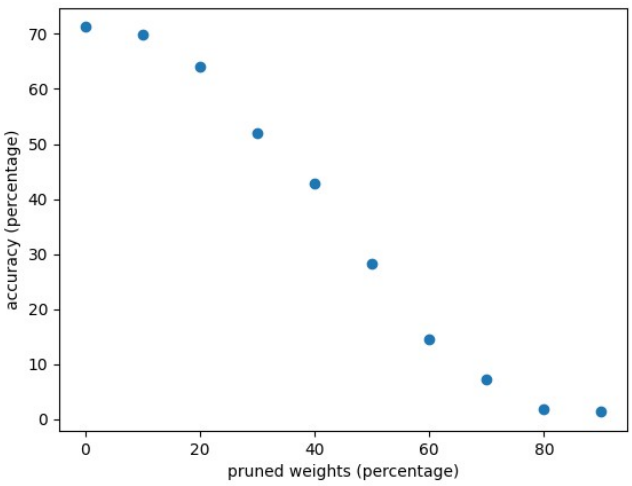

Figure 4: Gráfico relacionando a perda de acurácia com a poda de pesos com base em sua magnitude

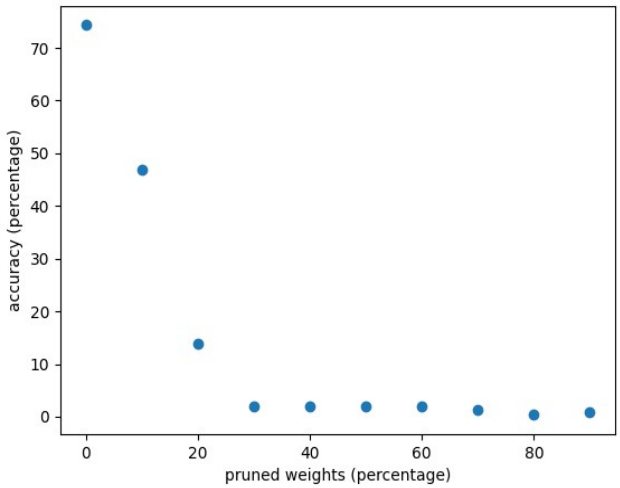

Figure 3: Gráfico relacionando a perda de acurácia com a poda de pesos com base em sua fase

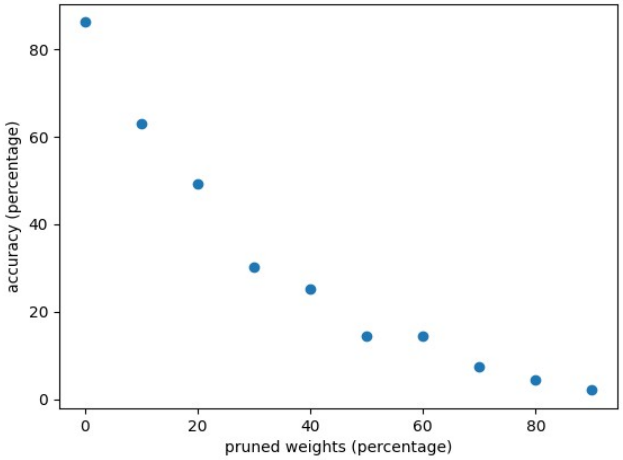

Figure 5: Gráfico relacionando a perda de acurácia com a poda de pesos com base em sua parte real.

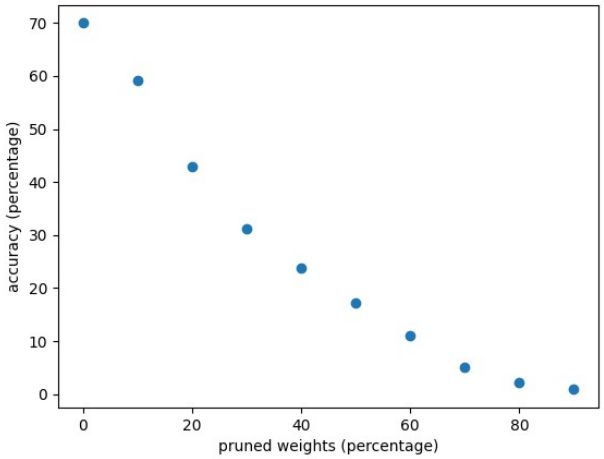

Figure 6: Gráfico relacionando a perda de acurácia com a poda de pesos com base em sua parte imaginária

CONCLUSÕES:

É possível observar, então, a partir disso, que, apesar da grande importância da fase para os números complexos, para a poda de redes seguindo algum parâmetro, o módulo permanece como fator principal, tanto pela queda de precisão menos abrupta da poda por módulo, como também pelas dos casos apenas com as parcelas reais e imaginárias dos pesos, que são as projeções do módulo nos eixos real e imaginário, respectivamente.

BIBLIOGRAFIA

CASTELLANO, G.; FANELLI, A.; PELILLO, M. **An iterative pruning algorithm for feedforward neural networks**. IEEE, 1997.

CRUZ, A. A.; MAYER, K. S.; ARANTES, D. S. **RosenPy: An open source Python framework for complex-valued neural networks.** SSRN, p. 1–18, nov. 2022a.

CRUZ, A. A.; MAYER, K. S.; ARANTES, D. S. **RosenPy. Holders: University of Campinas and Federal Institute of São Paulo.** BR 51 2022 002388 1. Deposited: Aug. 2022b.

HIROSE, A. **Complex-valued neural networks.** 2. ed. Berlin, Germany: Springer, 2012. p. 20–26.

LEE, J. et al. **Layer-adaptive sparsity for the Magnitude-based Pruning.** ICLR 2021, 2021.