Compressão Consciente por Poda Seguida de Quantização Iterativa em Redes Convolucionais: Um Estudo no CIFAR-10

Vitor Y. F. Freitas*†, Sérgio N. Silva*†§ e Marcelo A. C. Fernandes*†‡¶
*InovAI Lab, nPITI/IMD, UFRN, Natal, RN, Brasil
†Leading Advanced Technologies Center of Excellence (LANCE), nPITI/IMD, UFRN, Natal, RN, Brasil
‡Departamento de Engenharia Elétrica (DEE), UFCG, Campina Grande, PB, Basil
§Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, RN, Basil
Email:vitor.freitas.123@ufrn.edu.br, sergionatan@dee.ufcg.edu.br, mfernandes@dca.ufrn.br

***Resumo*—Este trabalho investiga a aplicação da compressão consciente por poda seguida de quantização iterativa em redes convolucionais, com foco na tarefa de classificação de imagens no conjunto de dados CIFAR-10. A técnica emprega poda adaptativa baseada em estatísticas por camada, seguida de quantização dinâmica aplicada durante o treinamento. O objetivo é reduzir o tamanho do modelo e sua complexidade computacional, mantendo desempenho preditivo competitivo. Os resultados obtidos demonstram que a técnica é capaz de manter acurácia superior a 84% com menos de 9% da densidade do modelo original e alcançar esparsidade de até 94,55%, com perda inferior a 4,2 pontos percentuais na acurácia. A métrica de Pontuação de Eficiência revelou que a combinação de quantização de 4 bits e poda intensa oferece o melhor equilíbrio entre compressão e desempenho, validando a aplicabilidade da abordagem em cenários com restrições de recursos.**

***Index Terms*—Poda, quantização, compressão consciente, aprendizado profundo, CIFAR-10, redes convolucionais**

I. Introdução

Na última década, o aprendizado profundo revolucionou várias áreas da computação, proporcionando avanços sem precedentes em visão computacional, processamento de linguagem natural e sistemas de recomendação. Este progresso, no entanto, foi acompanhado por um aumento exponencial na complexidade e tamanho dos modelos neurais, criando desafios significativos para sua implementação em sistemas com recursos limitados [1], [2].

A compressão de redes neurais surge como uma abordagem fundamental para viabilizar a execução de modelos complexos em sistemas com restrições de memória, processamento e largura de banda. Técnicas como poda [3], [4] e quantização [5], [6] têm sido amplamente estudadas de forma independente, permitindo reduções expressivas no custo computacional com impacto controlado no desempenho. No contexto mais amplo da aceleração da computação em aprendizado profundo, que tem recebido atenção crescente na literatura, estratégias baseadas em algoritmo, incluindo paralelismo de dados e de modelo, compressão por poda e quantização, exploração de esparsidade em diferentes níveis e redução de complexidade por aproximação, vêm sendo desenvolvidas com o objetivo de otimizar a inferência e a implantação de modelos em ambientes restritos [3]–[11].

Parâmetros e gradientes com valores pequenos (isto é, próximos de zero) tendem a ter pouco efeito nos resultados e no desempenho do modelo. A técnica de poda remove pesos, conexões ou até mesmo camadas inteiras que apresentam impacto mínimo sobre um modelo já treinado. A taxa de compressão alcançável por meio da poda depende tanto do conjunto de dados quanto da arquitetura da rede neural profunda [3], [4], [12], [13]. Para melhores resultados, os limiares de poda devem ser cuidadosamente selecionados de modo a minimizar o impacto na acurácia, sendo recomendável que a poda seja aplicada de forma iterativa, com re-treinamento do modelo a cada etapa. No entanto, a definição adequada dos limiares é uma tarefa desafiadora, uma vez que diferentes camadas podem apresentar sensibilidades distintas, e o re-treinamento repetido exige tanto acesso ao conjunto de dados original quanto tempo computacional adicional [3], [4], [8], [14].

Em computação, números de ponto flutuante de 32 bits são úteis para representar uma ampla faixa de valores. No contexto do aprendizado profundo, os parâmetros dos modelos tendem a assumir valores pequenos, em razão de técnicas como normalização em lote (batch normalization) e decaimento de pesos (weight decay), não sendo necessário utilizar uma faixa dinâmica tão ampla. A quantização utiliza tipos de dados menores (por exemplo, inteiros de 8 bits) para representar esses valores, reduzindo o consumo de memória e o custo associado às operações realizadas. Tanto os pesos das redes neurais profundas quanto as funções de ativação podem ser quantizados [5], [8], [10], [15]. Em contraste com a poda, a quantização pode ser aplicada diretamente a modelos já treinados, sem a necessidade de re-treinamento. Em certo sentido, a quantização também pode atuar como uma forma de poda, uma vez que valores suficientemente pequenos podem ser quantizados para zero [16].

A compressão convencional por poda é abordada em [3] e [4], onde a remoção de pesos ocorre apenas uma vez por época durante o treinamento. De forma análoga, a quantização

tradicional, discutida em [5], [6], aplica diferentes níveis de quantização por camada, também restritos a uma única aplicação por época. Este trabalho adota um novo esquema de compressão consciente do modelo, originalmente proposto em [17] para a tarefa de classificação de dados genômicos e posteriormente validado na classificação de modulações digitais em [18]. Esse esquema realiza, a cada iteração de treinamento, a aplicação sequencial de poda seguida de quantização ($P \rightarrow Q$), buscando minimizar simultaneamente as perdas associadas às duas técnicas.

A compressão iterativa por $P \rightarrow Q$ consiste em um algoritmo que combina poda e quantização de forma coordenada e contínua ao longo do treinamento. O principal desafio desse processo é reduzir o tamanho do modelo com o mínimo de perda de acurácia. Enquanto a poda remove parâmetros e reduz o número de operações matemáticas, a quantização reduz a precisão da representação binária dos pesos, diminuindo a complexidade do modelo.

Neste artigo, a compressão iterativa por $P \rightarrow Q$ é aplicada à tarefa de classificação de imagens utilizando o conjunto de dados CIFAR-10. O objetivo é avaliar a eficácia do esquema de compressão, anteriormente validado em domínios como genômica e comunicações digitais, agora em um cenário de visão computacional. A abordagem visa reduzir significativamente o tamanho do modelo e sua complexidade computacional, mantendo níveis competitivos de acurácia. Os resultados obtidos demonstram que a técnica é capaz de manter acurácia superior a 84% com menos de 9% da densidade do modelo original, e alcançar esparsidade de até 94,55% com perda inferior a 4,2 pontos percentuais na acurácia, evidenciando sua aplicabilidade em cenários com restrições de recursos.

II. Compressão Consciente de Modelos por Poda seguida de Quantização ($P \rightarrow Q$) Iterativa

O processo de treinamento, ilustrado na Figura 1, integra a compressão em cada iteração, permitindo que o modelo se ajuste progressivamente às restrições impostas pela poda e quantização ao longo do aprendizado. A técnica de compressão consciente utilizada adota uma abordagem sequencial, na qual a poda é aplicada antes da quantização em cada etapa do treinamento. Essa metodologia difere das abordagens tradicionais por possibilitar que o modelo aprenda representações mais robustas e otimizadas, ajustadas à aplicação iterativa e coordenada das duas operações de compressão.

O processo inicia com a propagação direta (Forward) da entrada $\mathbf{X}(n)$, resultando na saída $\mathbf{Y}(n)$. A saída é então comparada com a referência $Y_{\text{ref}}(n)$ por meio de uma função de erro (Error Func.), gerando o erro $\mathbf{E}(n)$. Esse erro é utilizado para calcular o gradiente $G(n)$, que orienta a atualização dos pesos do modelo. Após a atualização, os pesos $\mathbf{W}(n)$ passam por uma etapa de poda, reduzindo os parâmetros menos relevantes com base em uma função $P(\mathbf{W}(n), \beta)$, e em seguida são quantizados, produzindo os coeficientes comprimidos $\mathbf{C}(n)$. Esses coeficientes quantizados são então utilizados na próxima iteração, permitindo que o modelo se adapte iterativamente às restrições de compressão impostas ao longo do treinamento.

O princípio central desse esquema reside na aplicação da poda como etapa inicial, resultando em uma estrutura de pesos esparsa que pode ser posteriormente quantizada de forma mais eficaz. O modelo passa a adaptar dinamicamente suas distribuições de pesos ao processo sequencial de compressão, favorecendo a preservação de informações relevantes ao longo do pipeline. O primeiro estágio da técnica de compressão consciente consiste na aplicação de poda adaptativa, baseada nas características estatísticas específicas de cada camada. Para l-ésima camada do modelo, o limiar de poda β_l é calculado dinamicamente, conforme expresso por

$$\beta_l = \gamma \cdot \text{std}(\mathbf{W}_l) \tag{1}$$

em que $\text{std}(\mathbf{W}_l)$ representa o desvio padrão dos pesos da l-ésima camada e γ é um parâmetro que regula a intensidade da poda. Essa abordagem adaptativa permite que os limiares de poda sejam automaticamente ajustados de acordo com as propriedades estatísticas locais, preparando a estrutura de pesos de maneira eficiente para a quantização subsequente.

A operação de poda é formalmente definida pela função $P(\mathbf{W}_l(n), \beta_l)$, que atua diretamente sobre os pesos $\mathbf{W}_l(n)$ da camada l na iteração n, conforme é expressa como

$$\mathbf{H}(n) = P\left(\mathbf{W}_l(n), \beta_l\right) = \begin{cases} \mathbf{W}_l(n) & \text{se } |\mathbf{W}_l(n)| \geq \beta_l \\ 0 & \text{se } |\mathbf{W}_l(n)| < \beta_l \end{cases} \tag{2}$$

Essa operação estabelece que todos os elementos dos pesos com magnitude inferior ao limiar β_l sejam substituídos por zero, enquanto os demais são preservados. O resultado é um tensor esparso de pesos, reduzindo a complexidade computacional do modelo e preparando-o para a etapa subsequente de quantização.

O segundo estágio do processo aplica quantização dinâmica aos pesos podados provenientes da etapa anterior. A quantização é realizada sobre a estrutura de pesos esparsa $\mathbf{H}_l$, resultante da poda, o que permite uma codificação mais eficiente dos pesos remanescentes de maior significância. Para cada l-ésima camada, o passo de quantização q'_k é calculado da seguinte forma

$$q'_k = \frac{\max(|\mathbf{H}_l|) - \beta_l}{2^{(b-1)} - 1} \tag{3}$$

Em seguida, os pesos são quantizados por meio da expressão

$$\mathbf{C}_l = q'_k \cdot \left\lfloor \frac{\mathbf{H}_l}{q'_k} \right\rceil \tag{4}$$

Essa aplicação sequencial assegura que a quantização seja realizada sobre um conjunto de pesos já esparso, o que contribui para uma compressão mais eficiente e para uma melhor preservação dos valores mais relevantes, reduzindo a perda de informação durante o processo.

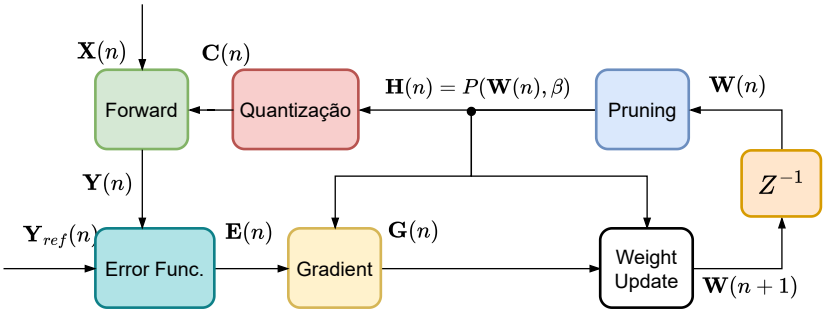

Figura 1. Loop de treinamento com compressão consciente: aplicação sequencial de poda seguida de quantização.

III. Metodologia

A. Métrica de Pontuação de Eficiência

Para avaliar de forma abrangente os compromissos entre compressão e acurácia, é adotada a métrica denominada Pontuação de Eficiência (PE), que quantifica o equilíbrio entre o grau de compressão do modelo e a preservação do desempenho preditivo

$$PE = \left(\frac{\text{Acurácia}_{\text{comprimida}}}{\text{Acurácia}_{\text{linha de base}}} \right)^{p} \times \frac{1}{RC_{\text{total}}} \tag{5}$$

A razão de compressão total é definida como

$$RC_{\text{total}} = \rho \times \frac{b_{\text{comprimido}}}{b_{\text{original}}} \tag{6}$$

Nessa expressão, ρ representa a densidade do modelo (porcentagem de pesos remanescentes após a poda), b denota o número de bits utilizados na representação dos pesos, e p é um parâmetro que pondera a importância da preservação da acurácia ($p \geq 1$).

A métrica PE fornece as seguintes interpretações:

- $PE > 1$: Indica compressão eficiente com perdas de desempenho consideradas aceitáveis;
- $PE < 1$: Sinaliza degradação significativa da acurácia em relação aos ganhos obtidos com a compressão;
- Valores elevados de p: Priorizam configurações que mantêm a acurácia mais próxima da obtida pelo modelo de linha de base.

A acurácia de linha de base é determinada automaticamente a partir da melhor configuração não comprimida (sem poda e com quantização de 32 bits), assegurando consistência na avaliação de todas as variantes comprimidas.

B. Arquitetura de Rede Neural

Os experimentos foram realizados com uma arquitetura de rede neural convolucional (CNN) desenvolvida para tarefas de classificação multiclasse em imagens. A rede é composta por cinco camadas convolucionais com normalização em lote, duas camadas de max-pooling e quatro camadas totalmente conectadas, com regularização por dropout. A arquitetura recebe imagens de entrada com dimensão $32 \times 32 \times 3$ e gera previsões para 10 classes, totalizando aproximadamente 2,1 milhões de parâmetros no modelo de linha de base sem compressão.

C. Configuração Experimental

Os experimentos foram conduzidos no conjunto de dados CIFAR-10, utilizando uma busca em grade exaustiva para explorar diferentes combinações de parâmetros na aplicação da técnica de compressão consciente. As variações consideradas incluíram:

- Intensidade da poda (γ): {0,0; 0,125; 0,250; 0,375; 0,500; 0,625; 0,750};
- Número de bits para quantização (b): {3, 4, 8, 16, 32};
- Taxas de aprendizado (*Learning Rate* - RT): {0,0001; 0,001; 0,005; 0,01; 0,05; 0,1}.

Cada experimento foi executado por até 200 épocas, com parada antecipada baseada na acurácia de validação. A melhor época de treinamento foi definida como aquela que obteve a maior acurácia de teste, e as métricas correspondentes (acurácia de treinamento, densidade do modelo) foram registradas nessa época, a fim de garantir uma comparação justa entre as diferentes configurações avaliadas.

A análise da Pontuação de Eficiência foi conduzida utilizando pesos de preservação da acurácia $p \in \{1, 2, 3\}$, permitindo avaliar os compromissos entre compressão e desempenho sob diferentes cenários de sensibilidade à acurácia. A acurácia de linha de base adotada foi de $87{,}21\%$, obtida com a melhor configuração não comprimida ($\gamma = 0$, quantização de 32 bits, taxa de aprendizado = 0.005).

IV. Resultados e Discussão

A. Eficácia da Compressão

A Figura 2 apresenta os resultados de acurácia de teste obtidos com a aplicação da técnica de compressão consciente para diferentes níveis de quantização. Observa-se que a aplicação sequencial de poda seguida de quantização permite a manutenção da acurácia do modelo mesmo sob condições de compressão com baixo número de bits.

A Figura 3 mostra a variação da densidade do modelo em função do número de bits de quantização, para diferentes intensidades de poda. Os resultados indicam que combinações intermediárias de intensidade de poda ($\gamma = 0{,}250 - 0{,}500$) com

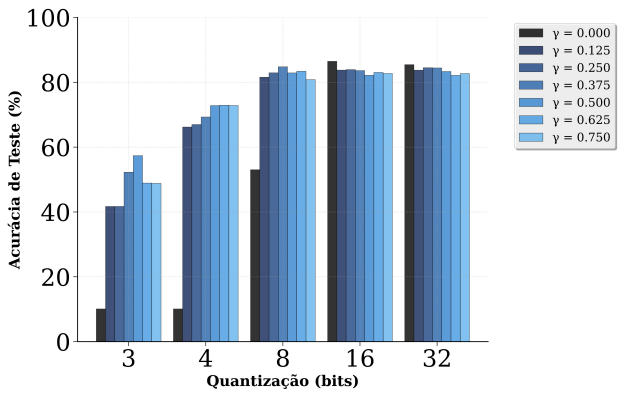

Figura 2. Acurácia de teste vs número de bits de quantização para diferentes intensidades de poda (LR = 0,01).

níveis de quantização entre 8 e 16 bits resultam em relações equilibradas entre compressão e densidade do modelo.

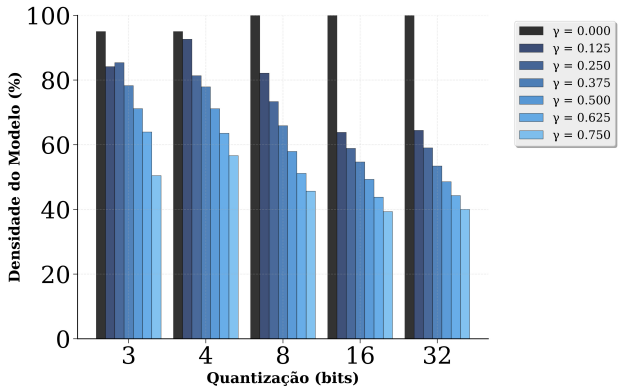

Figura 3. Densidade do modelo vs número de bits de quantização para diferentes intensidades de poda (LR = 0,001).

A Tabela I apresenta as melhores configurações obtidas em todas as taxas de aprendizado avaliadas, considerando a aplicação sequencial de poda seguida de quantização. Os valores reportam a acurácia de teste, com a densidade do modelo indicada entre parênteses. Valores em negrito destacam: a melhor acurácia não comprimida (87,21%), a melhor acurácia comprimida (84,78%) e a menor densidade observada (5,45%).

A técnica atingiu 94,55% de esparsidade (5,45% de densidade) na configuração com $\gamma = 0{,}625$ e quantização de 16 bits, mantendo uma acurácia de 83,04%. Essa configuração resultou em uma redução de 4,17 pontos percentuais em relação à acurácia de referência do modelo não comprimido (87,21%). Com $\gamma = 0{,}375$ e quantização de 8 bits, foi obtida uma acurácia de 84,78% com 8,68% de densidade, o que corresponde a 91,32% de esparsidade e uma diferença de 2,43 pontos percentuais em relação à acurácia da linha de base. Diversas configurações apresentaram acurácia superior a 80% com densidade inferior a 15%, indicando que a técnica pode ser aplicada sob diferentes combinações de poda e quantização mantendo desempenho competitivo mesmo com modelos altamente compactados.

B. Análise da pontuação de eficiência

A métrica de pontuação de eficiência (PE) permite quantificar os compromissos entre compressão e desempenho do modelo sob diferentes configurações. A Tabela II apresenta os valores de PE calculados para distintos pesos de preservação da acurácia ($p \in \{1, 2, 3\}$). Os resultados são apresentados para os níveis de quantização mais representativos. Valores em negrito indicam configurações com $PE > 40$, caracterizando relações favoráveis entre compactação do modelo e manutenção da acurácia relativa à linha de base.

Configurações com quantização de 4 bits e intensidades elevadas de poda ($\gamma = 0{,}500-0{,}750$) apresentaram, de forma consistente, as maiores pontuações de eficiência para todos os valores de p, com $PE > 74$ em todas as combinações avaliadas. Em particular, a configuração com $\gamma = 0{,}625$ e quantização de 4 bits resultou em pontuações de eficiência de 104,90, 95,27 e 86,52 para $p = 1$, $p = 2$ e $p = 3$, respectivamente. Essa configuração obteve 79,20% de acurácia com densidade do modelo de apenas 6,93%. Configurações com quantização de 3 bits também apresentaram valores elevados de PE sob alta intensidade de poda. Para $\gamma = 0{,}750$, foram observadas pontuações superiores a 50 para todos os valores de p, indicando viabilidade da técnica em cenários com compressão acentuada.

A Figura 4 apresenta a distribuição das pontuações de eficiência para diferentes combinações de intensidade de poda e níveis de quantização, considerando $p = 2$. Observa-se que a quantização de 4 bits, em combinação com intensidades moderadas a altas de poda, resulta nas maiores pontuações de eficiência entre as configurações analisadas.

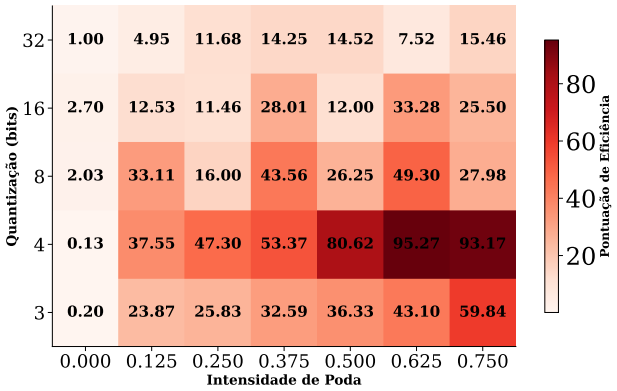

Figura 4. Mapa de calor da pontuação de eficiência para peso de preservação da acurácia $p = 2$.

A Figura 5 mostra a variação da pontuação de eficiência em função do número de bits de quantização, para diferentes intensidades de poda e com $p = 2$. Os resultados indicam que a quantização com 4 bits representa o ponto de maior eficiência no equilíbrio entre compressão e manutenção do desempenho.

Tabela I
MELHORES RESULTADOS PARA COMPRESSÃO CONSCIENTE: PODA SEGUIDA DE QUANTIZAÇÃO

γ	32 bits	16 bits	8 bits	4 bits	3 bits
0,000	**87,21%** (100,00%)	86,84% (73,55%)	62,15% (99,93%)	10,95% (100,00%)	12,05% (100,00%)
0,125	83,74% (18,64%)	83,76% (14,72%)	84,04% (11,22%)	83,85% (19,69%)	74,95% (33,00%)
0,250	84,49% (8,04%)	84,48% (16,38%)	83,03% (22,66%)	81,89% (14,91%)	76,43% (31,71%)
0,375	84,42% (6,58%)	83,61% (6,56%)	**84,78%** (8,68%)	79,75% (12,53%)	76,29% (25,05%)
0,500	83,33% (6,29%)	83,34% (15,22%)	83,58% (13,99%)	80,74% (8,50%)	76,00% (22,29%)
0,625	83,20% (12,09%)	83,04% **(5,45%)**	83,41% (7,42%)	79,20% (6,93%)	75,60% (18,60%)
0,750	82,68% (5,81%)	82,72% (7,06%)	82,47% (12,78%)	79,34% (7,11%)	75,69% (13,43%)

Tabela II
PONTUAÇÕES DE EFICIÊNCIA PARA DIFERENTES PESOS DE PRESERVAÇÃO DA ACURÁCIA.

γ	p = 1			p = 2			p = 3		
	8 bits	4 bits	3 bits	8 bits	4 bits	3 bits	8 bits	4 bits	3 bits
0,125	34,36	39,06	27,78	33,11	37,55	23,87	31,90	36,10	20,52
0,250	16,81	50,37	29,48	16,00	47,30	25,83	15,23	44,41	22,64
0,375	44,81	58,37	37,25	43,56	53,37	32,59	42,35	48,80	28,50
0,500	27,39	**87,09**	41,70	26,25	**80,62**	36,33	25,16	**74,63**	31,66
0,625	51,55	**104,90**	49,72	49,30	**95,27**	43,10	47,15	**86,52**	37,36
0,750	29,59	**102,42**	**68,95**	27,98	**93,18**	**59,84**	26,45	**84,77**	**51,94**

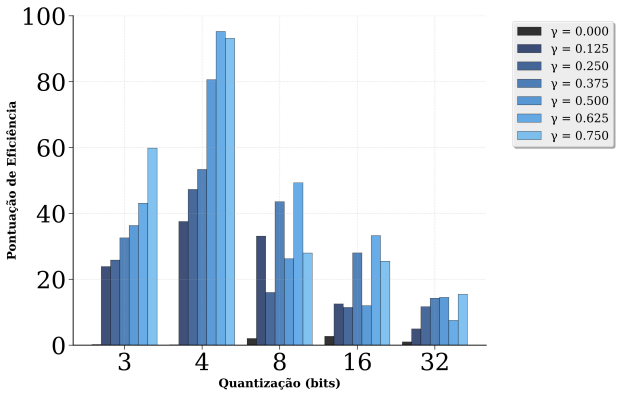

Figura 5. Pontuação de eficiência vs número de bits de quantização para diferentes intensidades de poda ($p = 2$)

C. Análise de Compressão Sequencial

Configurações com intensidades intermediárias de poda ($\gamma = 0{,}125{-}0{,}625$), combinadas com quantização de 8 a 16 bits, apresentaram os melhores resultados em termos de acurácia após compressão. A configuração com $\gamma = 0{,}375$ e quantização de 8 bits obteve a maior acurácia comprimida, alcançando 84,78%. Os resultados indicam que a quantização com 8 a 16 bits proporciona um equilíbrio eficiente entre compressão e desempenho preditivo quando aplicada após a etapa de poda. Em particular, a quantização com 8 bits demonstrou manter acurácia elevada quando precedida por poda adaptativa.

A manutenção de acurácia superior a 83% em modelos com densidade inferior a 9% ao longo do treinamento evidencia a capacidade da técnica de compressão consciente em permitir que o modelo adapte suas representações sob restrições de compactação. A Figura 6 apresenta a evolução da acurácia de teste durante o treinamento para quantização de 16 bits, considerando diferentes intensidades de poda. Os resultados mostram que configurações com poda intermediária favorecem padrões de convergência estáveis e com desempenho elevado, mesmo sob compressão progressiva.

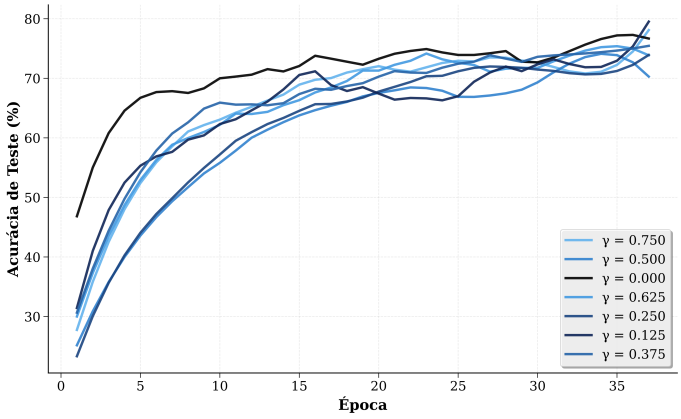

Figura 6. Evolução da acurácia de teste durante o treinamento para quantização de 16 bits com diferentes intensidades de poda.

D. Análise do impacto da taxa de aprendizado

A análise realizada por meio de uma varredura sistemática das taxas de aprendizado permite observar diferentes comportamentos do modelo sob restrições de compressão. A taxa de aprendizado de 0,01 apresentou os melhores resultados para configurações com compressão moderada, favorecendo

a adaptação do modelo às restrições impostas por poda e quantização. Já a taxa de $0{,}005$ foi responsável pelo melhor desempenho não comprimido, com acurácia de 87,21%.

Taxas de aprendizado mais elevadas ($0{,}05$–$0{,}1$) resultaram em maior desempenho para cenários com compressão intensa, como nas configurações com quantização de 3 a 4 bits. Esses resultados sugerem que dinâmicas de otimização mais acentuadas podem acelerar a adaptação do modelo a restrições severas. Observou-se que diferentes níveis de compressão requerem ajustes distintos na taxa de aprendizado, o que indica que a técnica de compressão consciente pode ser parametrizada conforme os requisitos específicos de implantação.

A Figura 7 apresenta a evolução da acurácia de treinamento ao longo do tempo para quantização de 16 bits, sob diferentes intensidades de poda. Verifica-se que a maioria das configurações atinge padrões estáveis de convergência, mesmo sob restrições crescentes de compressão.

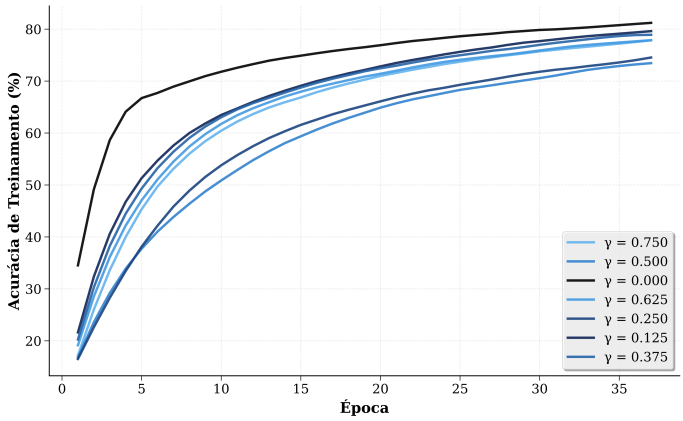

Figura 7. Evolução da acurácia de treinamento durante compressão consciente para quantização de 16 bits

E. Validação do Desempenho de Compressão

A Tabela III resume os principais resultados obtidos com a técnica de compressão consciente, apresentando diferentes combinações de acurácia, densidade e pontuação de eficiência. Os valores em negrito indicam maior acurácia não comprimida (87,21%), maior acurácia comprimida (84,78%), maior esparsidade (94,55%), menor densidade (5,45%) e maior pontuação de eficiência (95,27).

Os resultados obtidos no conjunto de dados CIFAR-10 demonstram que a aplicação sequencial de poda seguida de quantização impacta diretamente a estrutura e o desempenho do modelo, gerando diferentes perfis de compressão. A poda reduz a quantidade de pesos ao eliminar conexões com menor relevância, originando estruturas esparsas que favorecem a quantização com menor perda de informação. A execução dessas técnicas durante o treinamento, em vez de após o seu término, permite uma adaptação progressiva do modelo às restrições impostas. A análise com a métrica de pontuação de eficiência indica que a combinação de quantização com 4 bits e poda de alta intensidade apresenta a melhor relação entre tamanho do modelo e desempenho preditivo em contextos com restrições de recursos computacionais.

V. Conclusões

Este trabalho avaliou a aplicação de uma estratégia de compressão consciente baseada em poda adaptativa seguida de quantização iterativa em redes convolucionais, com foco na tarefa de classificação de imagens utilizando o conjunto de dados CIFAR-10. A técnica proposta integra poda e quantização diretamente no ciclo de treinamento, permitindo que o modelo se adapte progressivamente às restrições impostas por cada etapa de compressão. Os resultados experimentais demonstraram que a abordagem é capaz de preservar altos níveis de acurácia, mesmo sob compressão agressiva. Foi possível manter 84,78% de acurácia com apenas 8,68% da densidade do modelo, e alcançar esparsidade de até 94,55%, com perda inferior a 4,2 pontos percentuais em relação à linha de base. A análise com a métrica de Pontuação de Eficiência mostrou que a combinação de poda intensa com quantização de 4 bits oferece o melhor compromisso entre compactação e desempenho preditivo. Esses achados indicam que a compressão consciente é uma alternativa viável para a implementação de modelos de aprendizado profundo em cenários com restrições de recursos computacionais, como dispositivos embarcados e aplicações em borda. Como trabalho futuro, pretende-se aplicar a técnica a arquiteturas mais complexas e explorar sua integração com estratégias de quantização mista e poda estruturada.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte e financiamento.

Referências

[1] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," *arXiv preprint arXiv:1710.09282*, 2017.

[2] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," *Artificial Intelligence Review*, vol. 53, pp. 5113–5155, 2020.

[3] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?" *arXiv preprint arXiv:2003.03033*, 2020.

[4] Q. Huang, K. Zhou, S. You, and U. Neumann, "Learning to prune filters in convolutional neural networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 709–718.

[5] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, "Haq: Hardware-aware automated quantization with mixed precision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 8612–8620.

[6] A. S. Rakin, J. Yi, B. Gong, and D. Fan, "Defend deep neural networks against adversarial examples via fixed and dynamic quantized activation functions," *arXiv preprint arXiv:1807.06714*, 2018.

[7] S. Jung, C. Son, S. Lee, J. Son, J.-J. Han, Y. Kwak, S. J. Hwang, and C. Choi, "Learning to quantize deep networks by optimizing quantization intervals with task loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4350–4359.

[8] F. Tung and G. Mori, "Deep neural network compression by in-parallel pruning-quantization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 568–579, 2020.

[9] W. Zhe, J. Lin, V. Chandrasekhar, and B. Girod, "Optimizing the bit allocation for compression of weights and activations of deep neural networks," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3826–3830.

Tabela III
PRINCIPAIS RESULTADOS.

Configuração	Acurácia	Densidade	Esparsidade	PE (p=2)
Não comprimido	**87,21%**	100,00%	0%	2,70
Comprimido (8 bits)	**84,78%**	8,68%	91,32%	43,56
Menor densidade	83,04%	**5,45%**	**94,55%**	33,28
Maior PE (4 bits)	79,20%	6,93%	93,07%	**95,27**
Compromisso entre métricas	84,49%	8,04%	91,96%	11,68

[10] H. Kung, B. McDanel, and S. Q. Zhang, "Term revealing: Furthering quantization at run time on quantized dnns," *arXiv preprint arXiv:2007.06389*, 2020.

[11] Z. Li, H. Li, and L. Meng, "Model compression for deep neural networks: A survey," *Computers*, vol. 12, no. 3, p. 60, 2023.

[12] J. Tmamna, E. B. Ayed, R. Fourati, A. Hussain, and M. B. Ayed, "A cnn pruning approach using constrained binary particle swarm optimization with a reduced search space for image classification," *Applied Soft Computing*, vol. 164, p. 111978, 2024.

[13] J. Tmamna, E. B. Ayed, R. Fourati, M. Gogate, T. Arslan, A. Hussain, and M. B. Ayed, "Pruning deep neural networks for green energy-efficient models: A survey," *Cognitive Computation*, vol. 16, no. 6, pp. 2931–2952, 2024.

[14] P. V. Dantas, W. Sabino da Silva Jr, L. C. Cordeiro, and C. B. Carvalho, "A comprehensive review of model compression techniques in machine learning," *Applied Intelligence*, vol. 54, no. 22, pp. 11 804–11 844, 2024.

[15] S. Naveen and M. R. Kounte, "Optimized convolutional neural network at the iot edge for image detection using pruning and quantization," *Multimedia Tools and Applications*, vol. 84, no. 9, pp. 5435–5455, 2025.

[16] K. Nakata, D. Miyashita, J. Deguchi, and R. Fujimoto, "Accelerating cnn inference with an adaptive quantization method using computational complexity-aware regularization," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 108, no. 2, pp. 149–159, 2025.

[17] M. A. C. Fernandes and H. T. Kung, "A novel training strategy for deep learning model compression applied to viral classifications," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–9.

[18] M. A. Goldbarg, M. Balza, S. N. Silva, L. M. D. Silva, and M. A. C. Fernandes, "A novel training strategy for deep learning model compression applied to automatic modulation classification," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 477–492, 2025.