DEEP COMPRESSION: COMPRESSING DEEP NEURAL NETWORKS WITH PRUNING, TRAINED QUANTIZATION AND HUFFMAN CODING

Total de pontos 0/0

✓ No contexto do artigo, por que a compressão de redes neurais prioriza a execução em SRAM em vez de DRAM?

- ○ Porque a SRAM possui maior densidade de armazenamento e menor custo por bit
- ◉ Porque a SRAM não requer refresh e reduz significativamente o consumo de energia por acesso. ✓
- ○ Porque a DRAM não é compatível com operações de ponto flutuante.
- ○ Porque a SRAM permite armazenamento não volátil dos parâmetros da rede.

✓ Em representações de matrizes esparsas utilizadas para aceleração de inferência em redes neurais, uma estratégia consiste em armazenar a diferença entre os índices das posições não nulas, em vez dos próprios índices absolutos. Qual a principal vantagem dessa abordagem?

- ○ Reduz o tempo de busca por elementos não nulos, mas aumenta a redundância na representação.
- ○ Elimina completamente a necessidade de armazenar índices, tornando a matriz densa implicitamente.
- ◉ Diminui o número de bits necessários para representar os deslocamentos entre elementos, otimizando o uso de memória. ✓
- ○ Aumenta a precisão dos cálculos de multiplicação, já que diferenças são mais estáveis numericamente.

Considere uma camada densa de rede neural que possui 8 pesos, cada um armazenado originalmente em 32 bits (precisão simples).

Para reduzir o custo de memória, aplica-se a seguinte técnica:

Compartilhamento de Pesos (Weight Sharing): os 8 pesos distintos são aproximados por 2 valores centróides (clusters).

Quantização: em vez de armazenar os 32 bits de cada peso, armazena-se:

Os 2 valores centroides, cada um em 32 bits.

Um índice de 1 bit por peso para indicar a qual centróide ele pertence.

Com isso, os pesos originais não são mais armazenados diretamente, mas apenas os índices + centroids.

Qual é a razão de compressão ?

- 2,12
- 3,56 ✓
- 4,56
- 8,33

✓ No processo de quantização de pesos em redes neurais, diferentes estratégias podem ser utilizadas para a inicialização dos centróides. Considere três abordagens: (i) inicialização linear, (ii) baseada em densidade de frequência dos valores e (iii) inicialização randômica.
Qual das opções descreve corretamente a vantagem da abordagem linear no contexto do trabalho?

- (●) Distribui uniformemente os centróides ao longo do intervalo de valores, garantindo que pesos de maior representatividade sejam contemplados mesmo que não apareçam com grande frequência, evitando vieses de amostragem ✓
- ( ) Concentra os centróides apenas nas regiões mais populosas da distribuição de pesos, garantindo maior precisão local.
- ( ) Tende a ignorar valores extremos que poderiam impactar fortemente a reconstrução da rede após a quantização
- ( ) Nenhuma das opções

✓ No contexto do trabalho, qual a principal justificativa para a aplicação do algoritmo de Huffman sobre os índices quantizados dos pesos?

- ( ) O Huffman assegura que os centróides mais representativos sejam preservados com maior precisão, mesmo quando apresentam baixa frequência.
- (●) Huffman reduz a entropia da distribuição dos índices, gerando códigos menores para os valores mais frequentes e aumentando a eficiência de compressão do modelo. ✓
- ( ) O Huffman permite reorganizar os pesos no espaço contínuo, reduzindo o erro de reconstrução após a descompressão.
- ( ) O Huffman aumenta a capacidade de generalização da rede ao eliminar centróides de baixa probabilidade, funcionando como uma forma de regularização implícita.