Seminiãrio - MAC6931

Frederico Meletti Rappa - 15601846

1 Resumo

O seminário abordará o artigo **Dynasparse: Accelerating GNN Inference through Dynamic Sparsity Exploitation** [1] propõe um método para acelerar a inferência de Graph Neural Networks (GNNs) explorando a esparsidade dinâmica das conexões entre os nós do grafo.

As Graph Neural Networks (GNNs) têm se tornado essenciais para modelagem de dados estruturados em grafos, sendo aplicadas em diversas áreas, como recomendação de conteúdo, análise de redes sociais e bioquímica. No entanto, a inferência desses modelos ainda representa um grande desafio devido ao alto custo computacional, ao grande número de operações necessárias para processar a vizinhança de cada nó e caráter irregular de grafos. O artigo propõe uma abordagem para acelerar a inferência de GNNs por meio da exploração da esparsidade dinâmica, permitindo uma execução mais eficiente sem comprometer significativamente a precisão dos resultados.

A principal contribuição do artigo é a introdução de um mecanismo adaptativo que identifica padrões de esparsidade durante a inferência, em vez de depender de estruturas estáticas de esparsidade definidas previamente. Esse mecanismo permite que o modelo ajuste dinamicamente quais conexões devem ser mantidas ou descartadas com base na importância das informações propagadas entre os nós. Dessa forma, é possível reduzir a carga computacional eliminando operações desnecessárias, com pouca perda de qualidade nas previsões. Dynasparce é implementado em FPGAs pela possibilidade de definir uma organização de memória específica para diferentes diretivas computacionais de acordo com esparsidade do grafo, definir um mecanismo eficiente de hardware para transformação do formato dos dados e identificação da esparsidade, e desenvolver um *soft-processor* para o mapeamento dinâmico de kernels.

2 Justificativa

Conforme mencionado, o avanço de GNNs tem impulsionado aplicações em diversas áreas. Entretanto, o uso eficiente dessas redes ainda apresenta desafios para CPUs e GPUs. O artigo propõe uma abordagem para acelerar a inferência de GNNs, reduzindo o custo computacional sem comprometer significativamente a precisão. O uso de FPGAs como alternativa para execução de inferência dos modelos representa uma alternativa a métodos tradicionais por permitir a configuração de hardware especializado e reconfigurável. Além disso, o artigo propõe uma abordagem interessante para contornar a esparsidade variável da representação dos grafos e se destaca por propor co-design de software de pré-processamento dos modelos e hardware em FPGA.

References

[1] Bingyi Zhang and Viktor Prasanna. Dynasparse: Accelerating gnn inference through dynamic sparsity exploitation. In *2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 233–244. IEEE, 2023.