Summary

The paper looks at data partitioning and schedule strategies employed by DNN accelerators, which directly impacts the performance and the energy efficiency. The impact of the choice of dataflows on the utilization and efficiency, along with the tools and methodologies used to help architects explore the design space is explored.

The paper proposes a data-centric approach towards acceleration, as opposed to compute-centric approach followed before. It also shows how these directives can be used to optimize for data reuse better. This is used to define MAESTRO that takes in as input

- a DNN model with a set of layers
- a dataflow description for each layer
- a hardware configuration
  Maestro estimates the end-to-end execution time, energy, Network-on-Chip costs and so on. MAESTRO cost model can also be used to determine optimal parameters with given area, energy, or throughput budget for a given model. The MAESTRO model uses concepts of Spatial, Temporal and Spatio-Temporal Reuse to better model the efficiency, as opposed to a simple compute-oriented loop structure.

Strengths

- Data centric flow helps in visualisation of the biggest bottleneck in accelerating LLMs - memory bandwidth, and helps in working around through smart usage of better L1 memory for data reuse through PE to PE interconnects
- Allows for effective design space exploration, highlighting design preferences according to layers which varies dramatically

Weaknesses

- Cannot model content of datacells apart from simpler methods; i.e. in case of transformers, wherein sparse layers are common, there might be a lot of compute involving zeros which are wasted compute that can be skipped
- performs comparison of dataflows, does not take into actual designs which may have custom blocks
-

Improvements

The knowledge of the wasted compute i.e. the ones which involve zeros, can help speeding up the calculations much better, as these are wasted compute. A better use of the chunks i.e instead of a simple offset, the offset should be non-continous and accoording to the distribution of the data, can help in modelling those and finding optimisations faster.