Summary

Sparse CNN accelerator architecture that employs a compressed data format for the sparse weights and activations, eleminating unnecessary data transfers and reduces storage requirements. It also provides an efficient NoC that facilitates efficient delivery of the weights and activations to the PE array and focuses on its reuse. The accumulation is then performed through a novel accumulation array. It works on the principle that in a CNN, every input weight will be multiplied with every filter weight. The accumulator network then distributes after the Cartesian product between input and filter weights.

Strengths

- Provided the arrays are sparse enough (which they are through factors such as dropout), computations are cut down as opposed to a dense multiplier.
- Sparse encoding frees up bandwidth to PE

Weaknesses

- The tiling methodology introduces halos, which cuts down data reuse as the same data has to be held by multiple PEs, decreasing compute effectiveness
- Works only on CNNs as every value of the filter gets multiplied with the input at some point of time; for normal matrix multiplications, the partial sums generated may not be useful all the time since it first computes it and then decides how to distribute the product to the outputs.

Suggested Improvements

- Dataflow can be replaced from a coordinate system to more compressed formats such as CPSR
- Improvements in the NoC between PEs and accumulator arrays according to the format used