Summary: The authors introduce a novel dataflow called Row Stationary (RS) for acceleration of convolutional neural networks (CNNs) on spatial architectures. The authors additionally develop an analysis framework to evaluate the energy efficiency of existing CNN dataflows (including weight stationary and several flavors of output stationary) under fixed hardware constraints such as area. Using this framework, they demonstrate that RS is 1.4x to 2.5x more energy-efficient than existing dataflows for convolutional layers and at least 1.3x more efficient for fully-connected layers in AlexNet. The RS dataflow achieves this by optimizing for all types of data movement across the storage hierarchy (from the register file level to the DRAM level), which involves jointly maximizing input data reuse and minimizing partial sum accumulation costs.

2 strengths:

1. The observation that optimizing for all types of data movement in the storage hierarchy simultaneously is theoretically sound and is discussed in sufficient detail. In addition, the developed RS dataflow is backed by practical results. In particular, their dataflow demonstrates significant improvements in energy efficiency over existing approaches for both the convolutional layers and fully-connected layers of AlexNet.
2. The authors describe a comprehensive analysis framework for the fair comparison of energy efficiency of different CNN dataflows by imposing the same hardware constraints. Although several CNN dataflows had been proposed prior to this paper, it was difficult to compare them owing to the lack of such a framework. Their analysis framework is holistic and considers the energy cost for both input data access and partial sum accumulation under hardware constraints. In addition, the paper points notes that although their analysis model does not always model real costs accurately, their results are still conservative compared to the other dataflows.

2 weaknesses:

1. The evaluation described in the paper is limited to AlexNet, which is a relatively old and simple architecture that is not representative of state-of-the-art CNNs. Although CNNs such as VGG16 and ResNet had been developed prior to the publishing of this paper (in 2014 and 2015 respectively), the RS dataflow is not evaluation on either of these networks. Testing on diverse CNN models would provide stronger evidence for the generalizability of the RS dataflow's energy-efficiency improvements. Beyond this, it is not completely clear why the authors selected the batch sizes or PE array sizes they did for evaluation.
2. Although the paper took note of the weaknesses of other dataflows, they fail to critically analyze the RS dataflow they propose. In particular, there is a lack of discussion of limitations or tradeoffs that the RS dataflow introduces. For instance, from the description in the paper, the dataflow seems to introduce additional complexity in terms of control logic. In addition, certain details of the dataflow are not clear. The discussion of the first and second phases of folding could have benefited from a visualization depicting how exactly these are performed. Further, there is no discussion of folding in the context of the analysis framework.

Suggested Improvement: The authors could explore the potential for dynamic reconfiguration of the RS dataflow depending on the characteristics of different CNN layers. In particular, the paper mentions that both mapping steps occur "statically prior to runtime". This limits the applicability of the dataflow to a single convolutional shape before new mappings are computed. A logical extension is then to allow adjustments of mapping strategy on-the-fly depending on the properties of each layer. Although this may lead to certain additional hardware costs, it could potentially lead to improved energy efficiency by reducing data movement requirements based on the layer.