DENSO IT LAB

ICLR

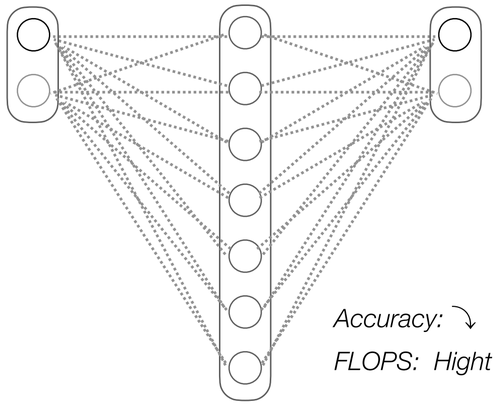
Structured Activation Sparsification

GPU Compatible DNN Acceleration By Utilizing Sparsity In Activation

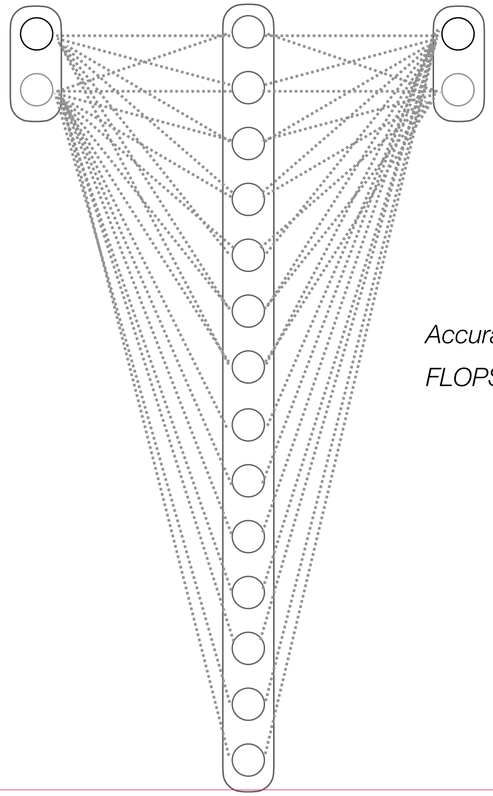
Yusuke Sekikawa, DENSO IT Lab. Inc.,

Background

- Wide network (having more channel) yield good accuracy
- However, It consumes more FLOPS



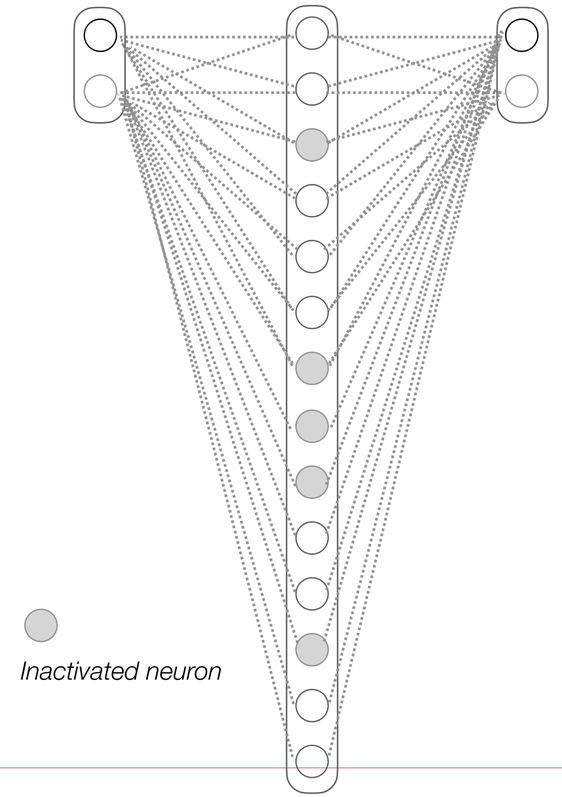
Accuracy: ↘
FLOPS: Hight



Accuracy: ↗
FLOPS: Hight

Research question

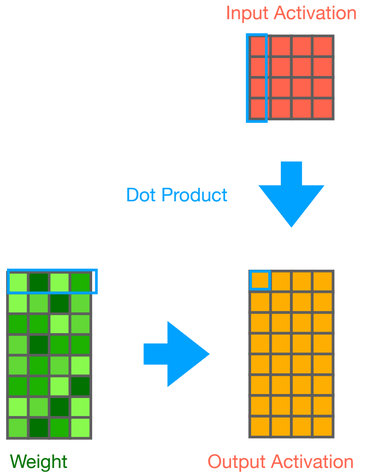
- Can we utilize the sparsity in activation?
  - Sparsity induced by activation function (e.g., ReLU) is input dependent and unstructured -> Hard to utilize on vector process such as GPU



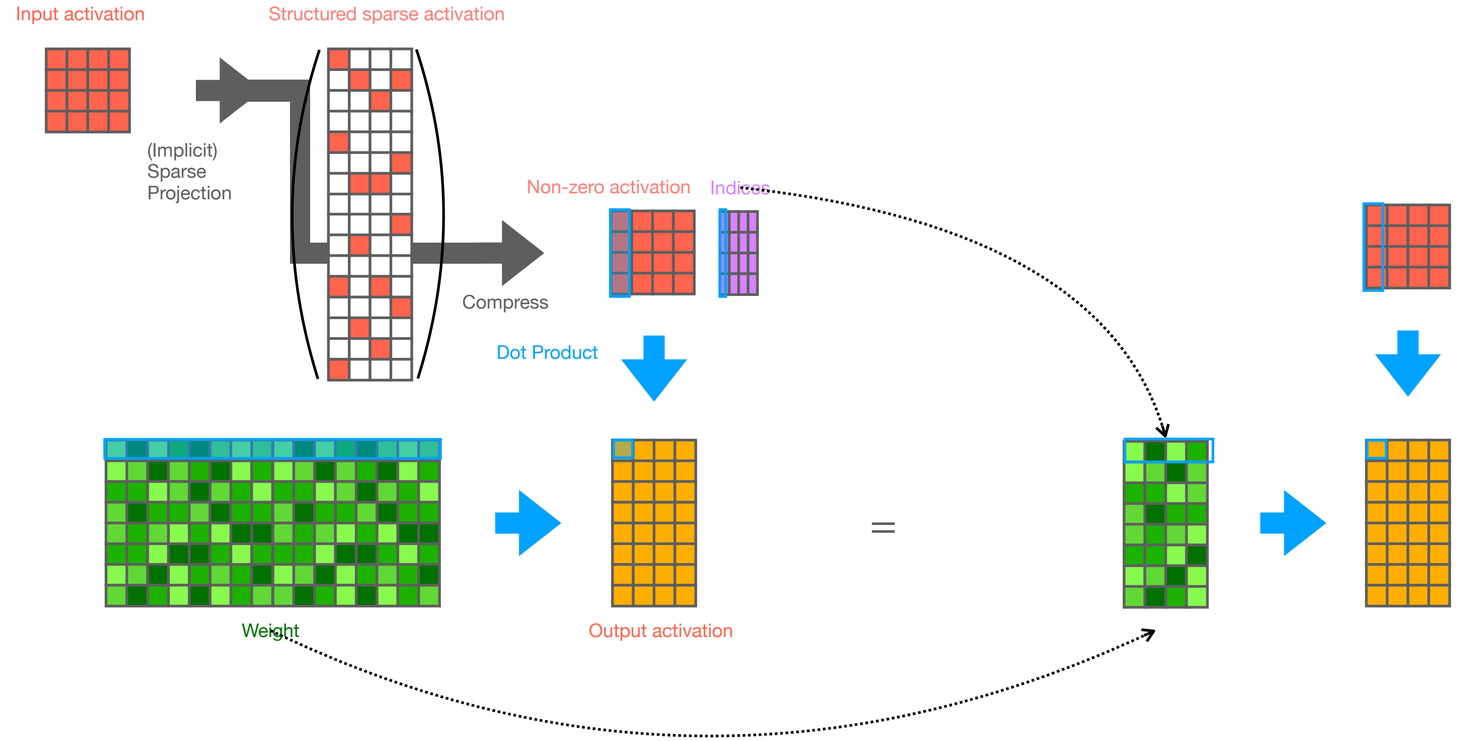
Structured Activation Sparsification

Core Idea: Structured Weight by Projection

- Realize wide network with structurally sparse activation by implicit projection



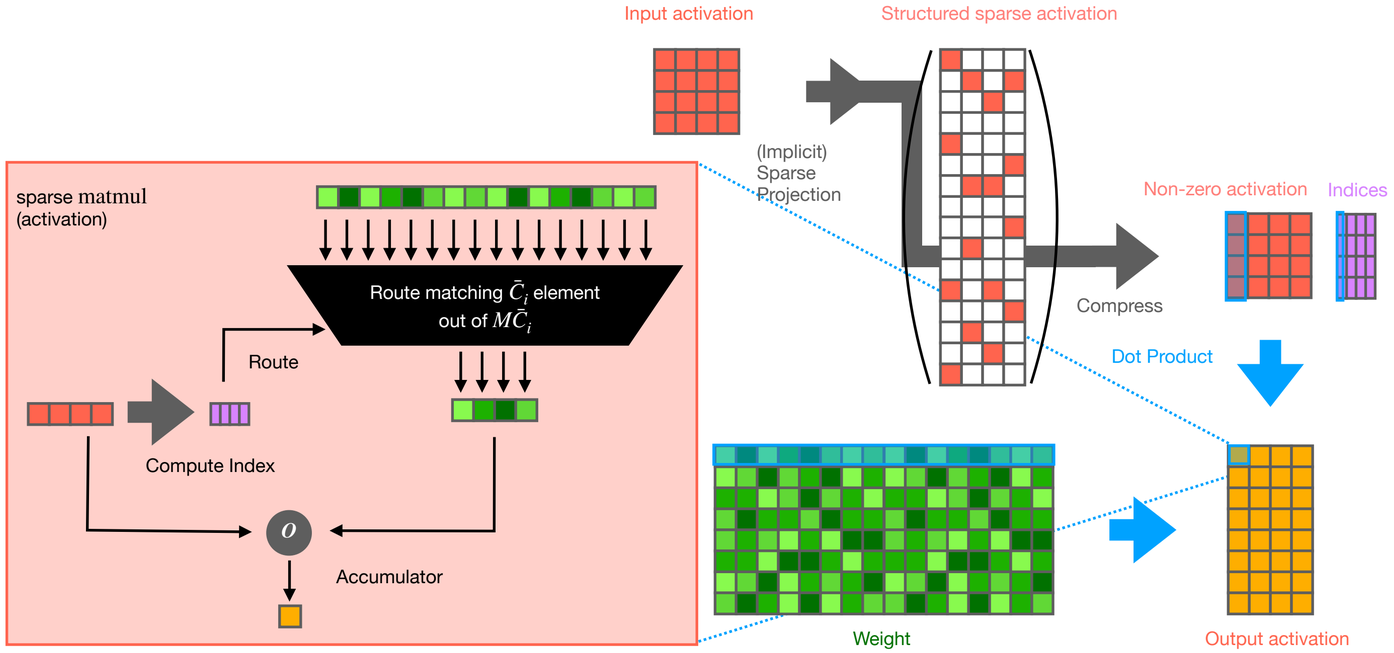
Dense Matrix Multiplication



SAS Matrix Multiplication (1:4)

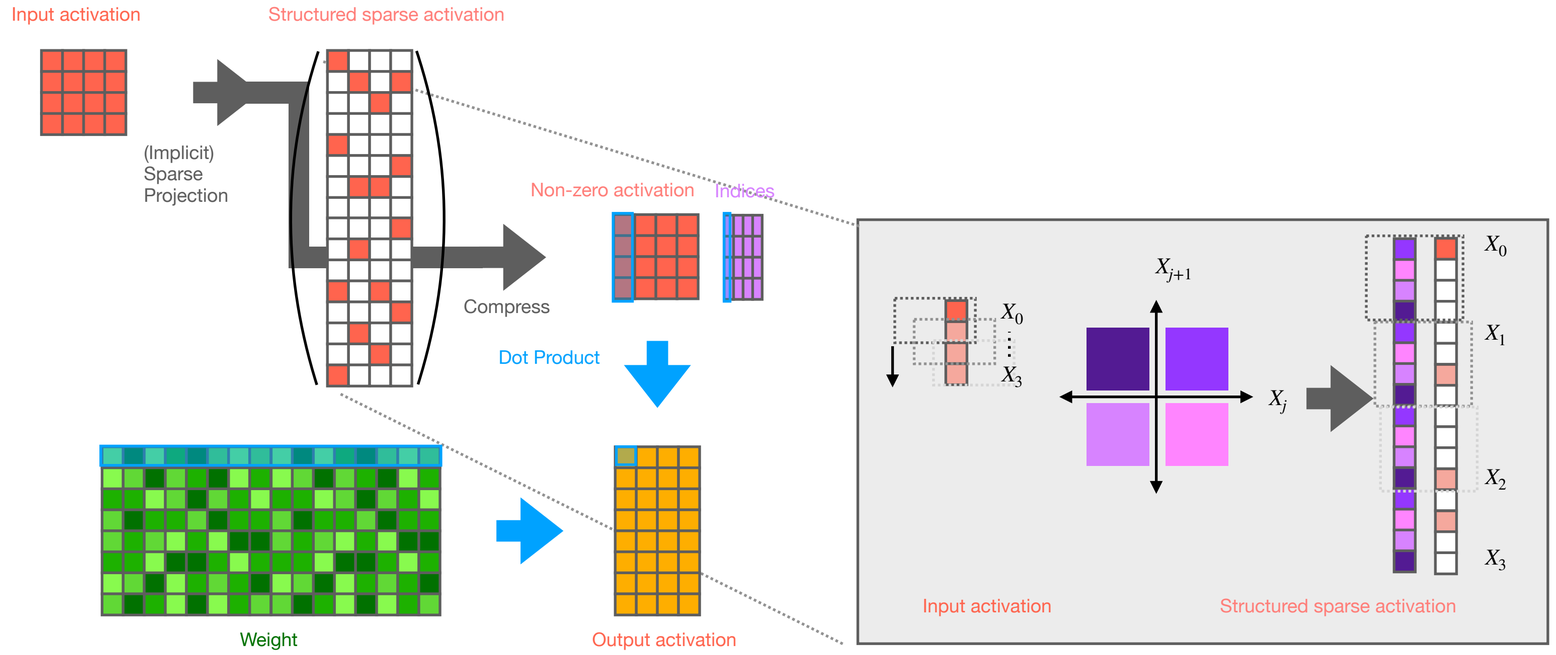
GPU compatible

- Increase wide for $\boldsymbol{M}\times$ without increasing FLOP on commercial GPU
  - Utilize *SparseTensorCore* developed for sparse-weight



Sparse projection mechanism

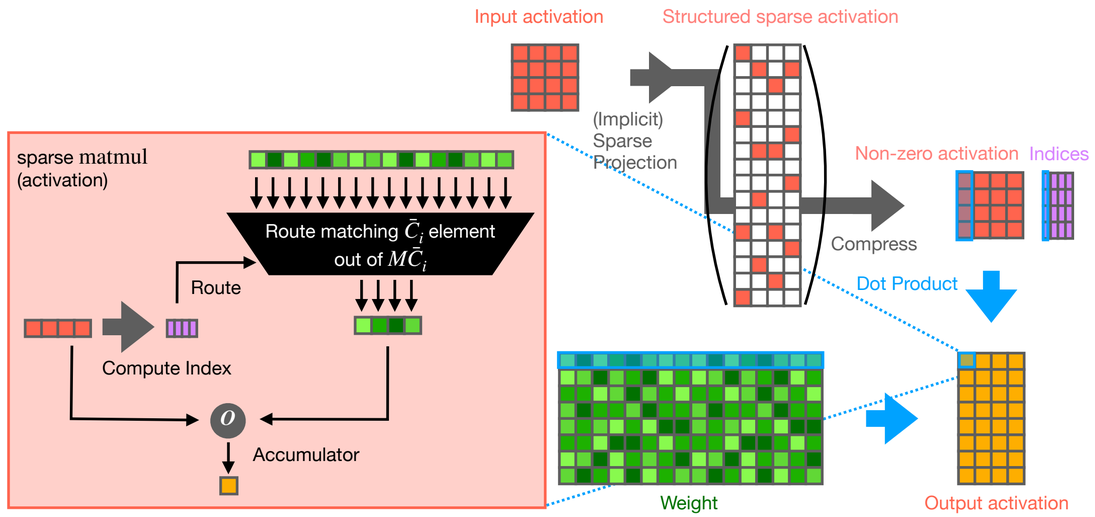
- Input-dependent implicit sparse projection
  - We do not actually make sparse activation by directly computing the index for nonzero



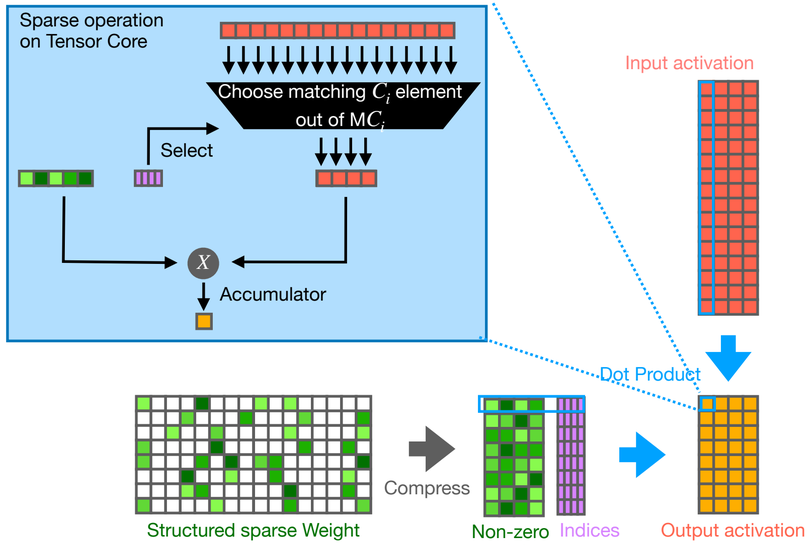
Is SAS better than SWS

- Both SAS and SWS increase network width while keeping the same FLOPS
  - Which is better, given the same FLOPS?

Note: SAS consume $M\times$ memory for weight than SWS



SAS: Structured Activation Sparsification (Ours)

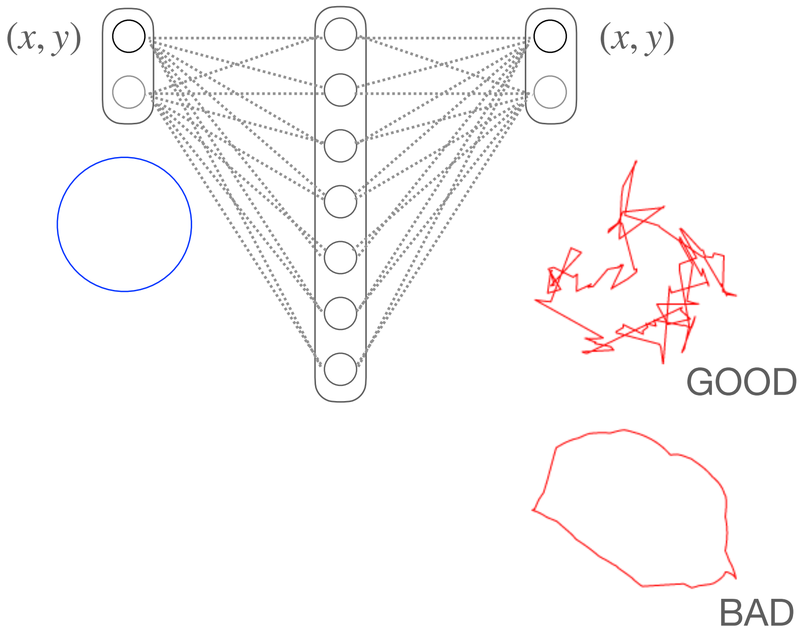


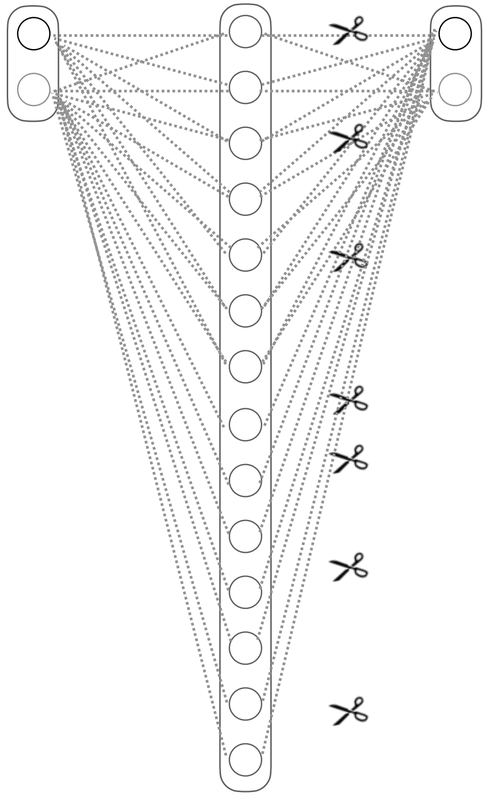
SWS: Structured Weight Sparsification (Nvidia)

Preliminary Experiment: Expressiveness by Trajectory Length

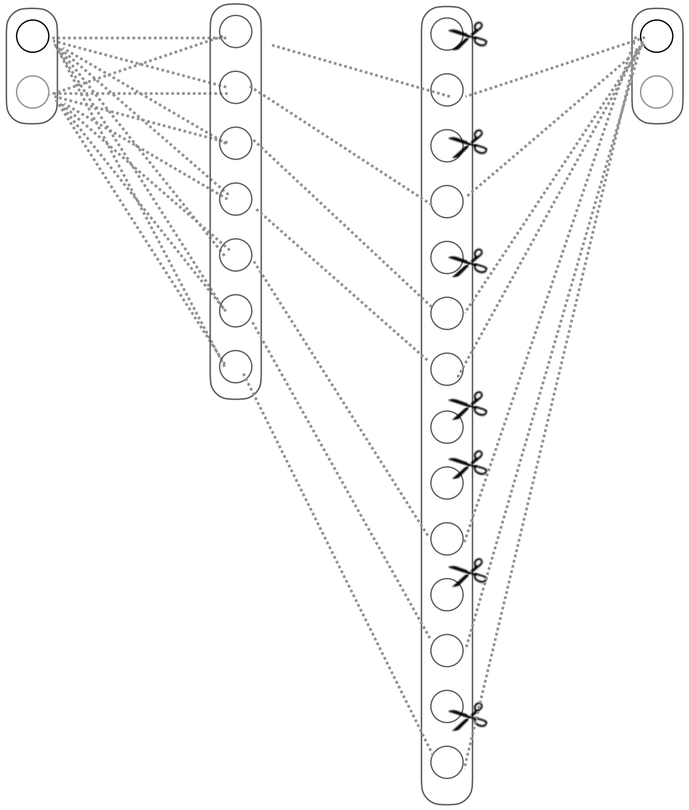
Trajectory Length: Longer length (complicated shape) indicates more expressiveness

"On the Expressive Power of Deep Neural Networks" https://proceedings.mlr.press/v70/raghu17a/raghu17a.pdf ICML2017



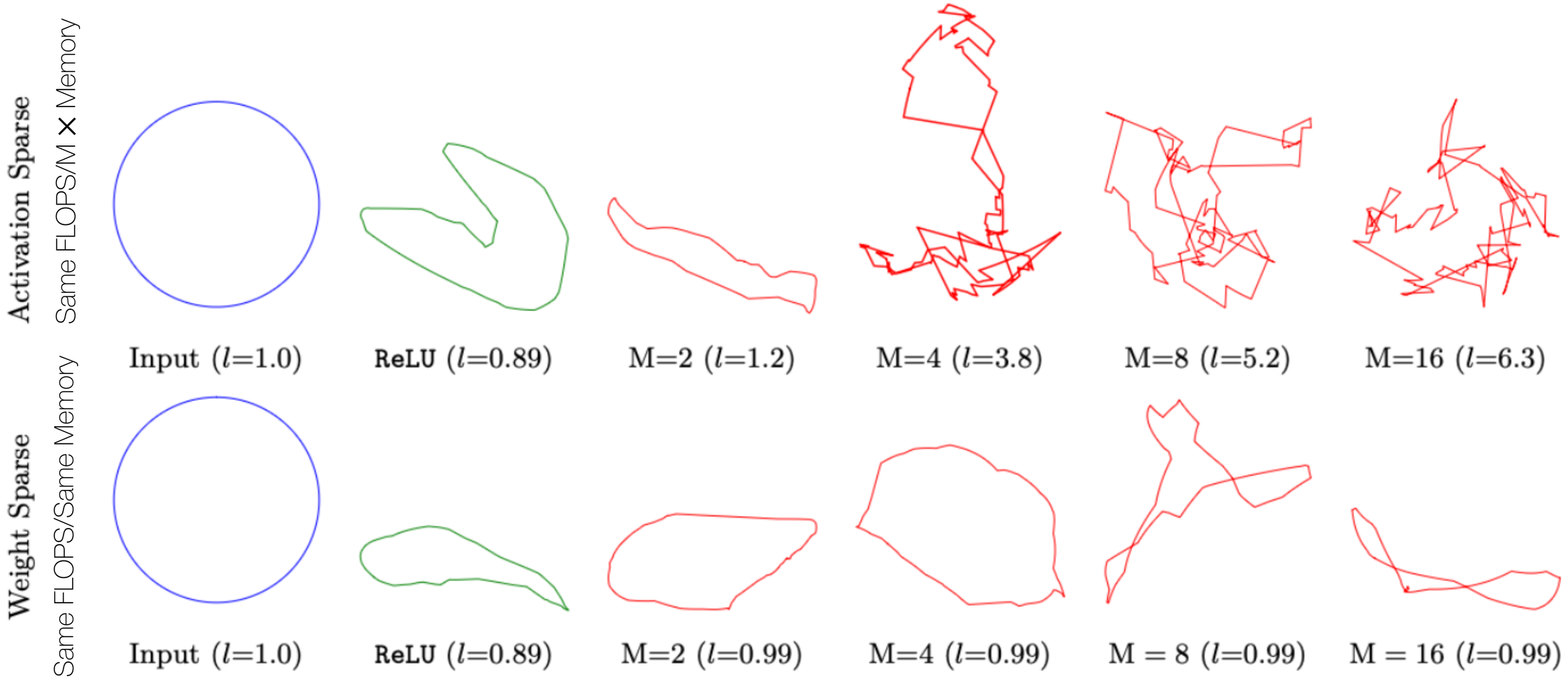


Structured Weight Sparsification (1:2 SWS, NVIDIA)



Structured Activation Sparsification (1-2 SAS, Ours)

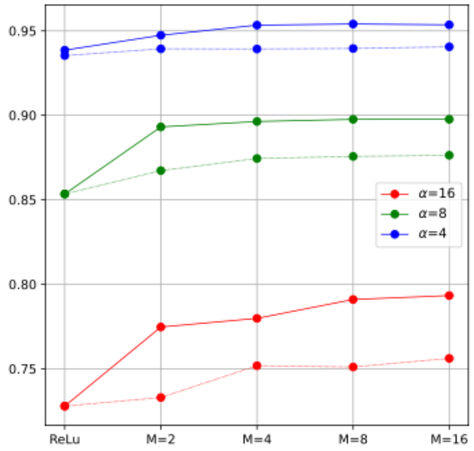
Trajectory Length: Evaluation result



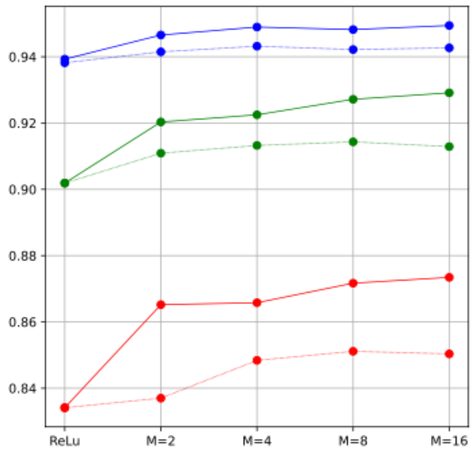
SAS yields longer trajectory indicating more expressiveness

Result on CiFAR10/100

CIFAR100

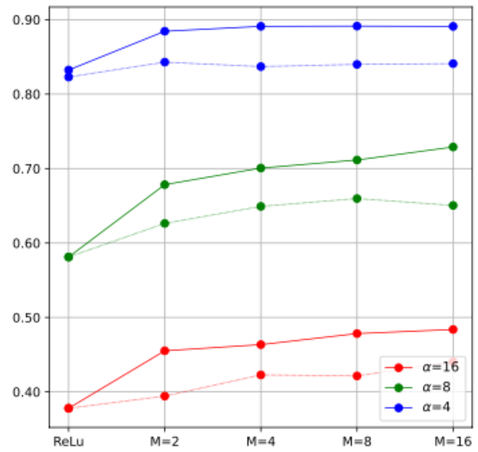


(a) Train summary

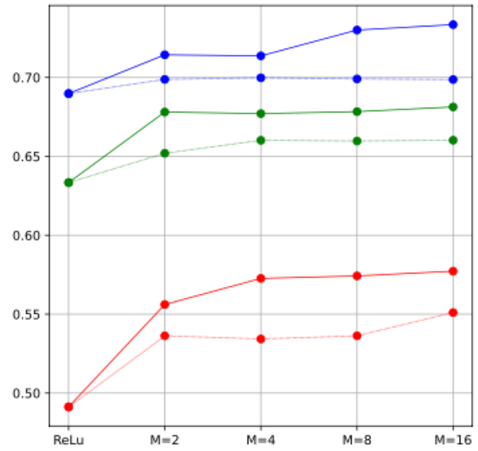


(e) Test summary

CIFAR100



(a) Train summary



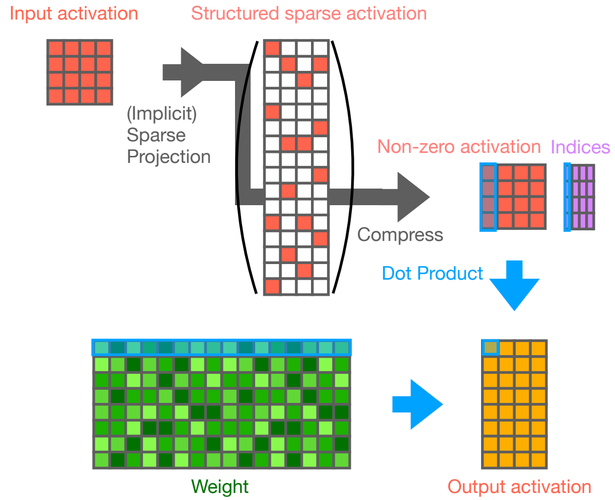
(e) Test summary

- Increasing sparsity (M)
  → Better accuracy without increasing FLOPS
- Better than SWS for the FLOPS (and sparsity)

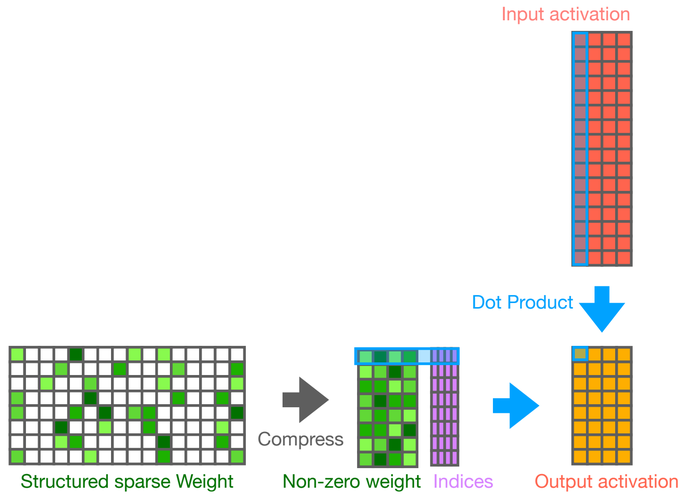
  Note: SAS consume $\boldsymbol{M}\times$ memory for weight than SWS

SAS Summary

- We explore the utilization of projected structured sparsity in activation
  - Increasing the width of NN by sparse projection increases capacity while keeping the same FLOPS
  - Better than SWS in terms of FLOPS/accuracy tradeoff (SAS consumes more memory for weight)
- Future work
  - Develop library to build SAS neural network
  - Combination wit quantization



SAS: Structured Activation Sparsification (Ours)



SWS: Structured Weight Sparsification