Compression with Attention: Learning in Lower Dimensions

Gaurav Singh & Kia Bazargan *{singh431, kia}@umn.edu*

University of Minnesota – Twin Cities

1. Introduction

Deep learning (DL) models are ballooning in size

- They are built on the principle of stacking linear functions with fixed non-linear activations.
- Multi-layer Perceptrons (MLPs) require many hidden neurons. For example, 75% of the parameters in GPT-3 175B are MLPs!
- Significant efforts have already been made to compress Deep Neural Networks (DNNs) through quantization, distillation, and pruning.

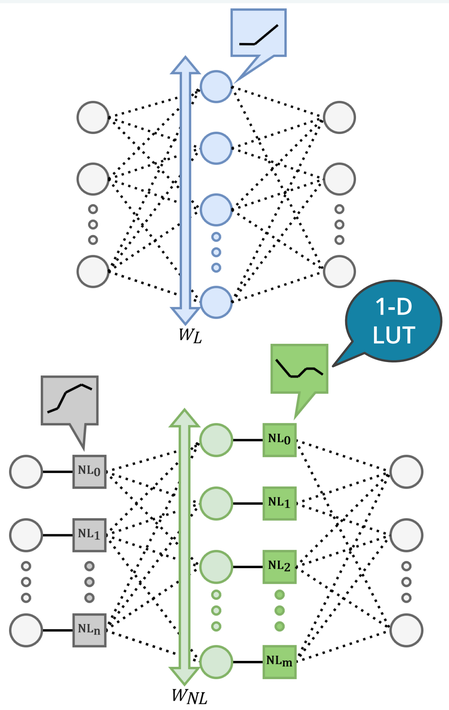
Can we reduce the MLP's hidden size further without sacrificing accuracy?

2. Methodology

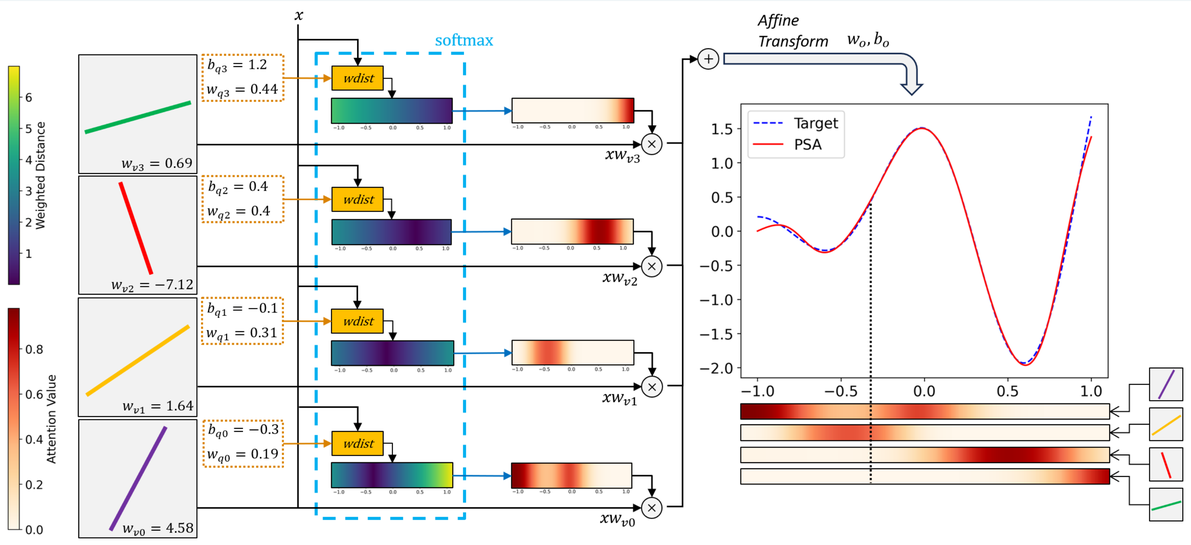
- Separate learning between linear multi-dimensional functions and non-linear 1-D functions.
- Utilize **distillation** to transfer knowledge into these non-linear functions.
- **Quantize** the model.
- Implement non-linear functions as a simple look-up table.

A Look-up Table has O(1) complexity; hence, any reduction in hidden size will save total model OPS!

How can we efficiently learn arbitrary non-linear 1-D functions in DL?



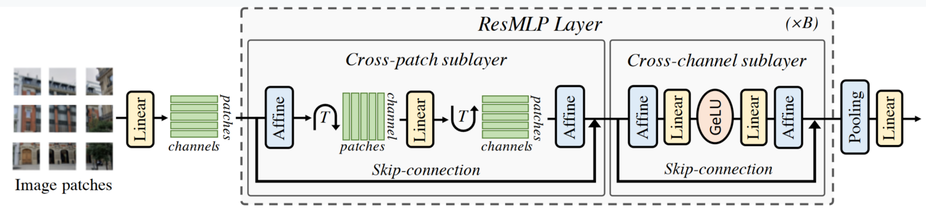
3. Personal Self-Attention



Break a 1D function down into h different linear segments. Each segment has its own dedicated center. Pay ***attention*** to each segment based on a ***distance score***.

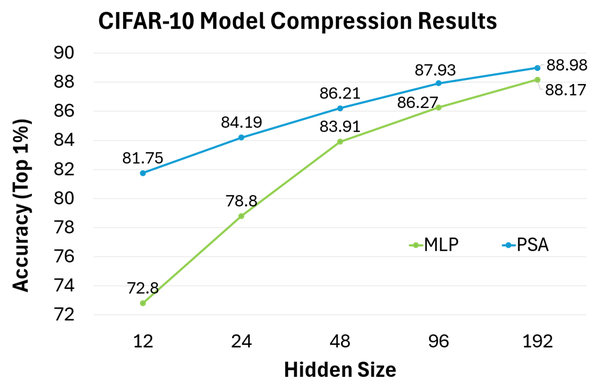
- Fully compatible with backpropogation.
- Continuous with polynomial like smoothing.
- Adaptive to any non-symmetric non-linearity.

4. Training & Hardware Results



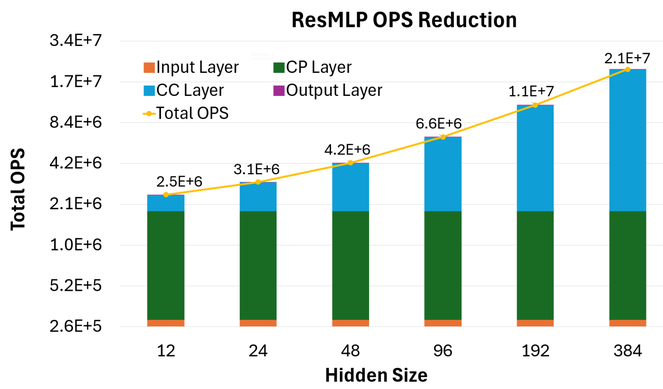
Testing on ResMLP [3]

- Isolates features in MLP.
- Simple unrolled hardware implementation.
- Image classification on CIFAR10 and SVHN.



Full flow:

1. Train the teacher model with no compression.
2. Add PSA at the input and hidden neurons.
3. Distill the teacher into a student with smaller hidden sizes.
4. Quantize the full model to INT8.
5. Implement PSA as a 1-D 8-bit Look-up Table.



	FINN [1]	FracBNN [2]	Ours - MLP	Ours - PSA	Ours - PSA
Accuracy (Top 1%)	80.1	89.1	86	85.9	85.9
Device	Zynq ZC706	Zynq ZU3EG	Alveo U250	Alveo U250	Kintex XC7K160T
Throughput (FPS)	21900	2806.9	29450.7	29450.7	20454.9
LUTs	46253	51444	62025	66462	75655
DSPs	-	126	1083	697	600
BRAM	186	212	84	140	212
Frequency (MHz)	200	250	200	200	143
Power (W)	3.6	4.1	5.7	4.7	2.0
FPS/LUT	0.47	0.05	0.47	0.44	0.27
FPS/DSP	-	22.28	27.19	42.25	34.09
FPS/W	6083	685	5167	6266	10227

5. Conclusion & Future Work

A new operation called Personal Self-Attention is proposed.

- It shows how such an operation can compress an MLP, reducing hidden size by 2x and reducing total model OPS by 1.57x.
- A hardware accelerator using PSA is demonstrated, reducing DSP count by 1.55x.

Our future work will include:

- Testing on larger models and datasets (Imagenet, BERT, etc).
- Other Machine Learning tasks like regression.
- Optimizing PSA hardware implementation.

This work was supported in part by the Cisco Systems, Inc. under Grant 1085913, and in part by the National Science Foundation under Grant PFI-TT 2016390.

[1] Umuroglu, Yaman, et al. 'FINN: A Framework for Fast, Scalable Binarized Neural Network Inference'. Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Association for Computing Machinery, 2017, pp. 65–74
[2] Zhang, Yichi, et al. 'FracBNN: Accurate and FPGA-Efficient Binary Neural Networks with Fractional Activations'. The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Association for Computing Machinery, 2021, pp. 171–182
[3] Touvron, Hugo, et al. 'ResMLP: Feedforward Networks for Image Classification with Data-Efficient Training'. arXiv [Cs.CV], 2021, http://arxiv.org/abs/2105.03404