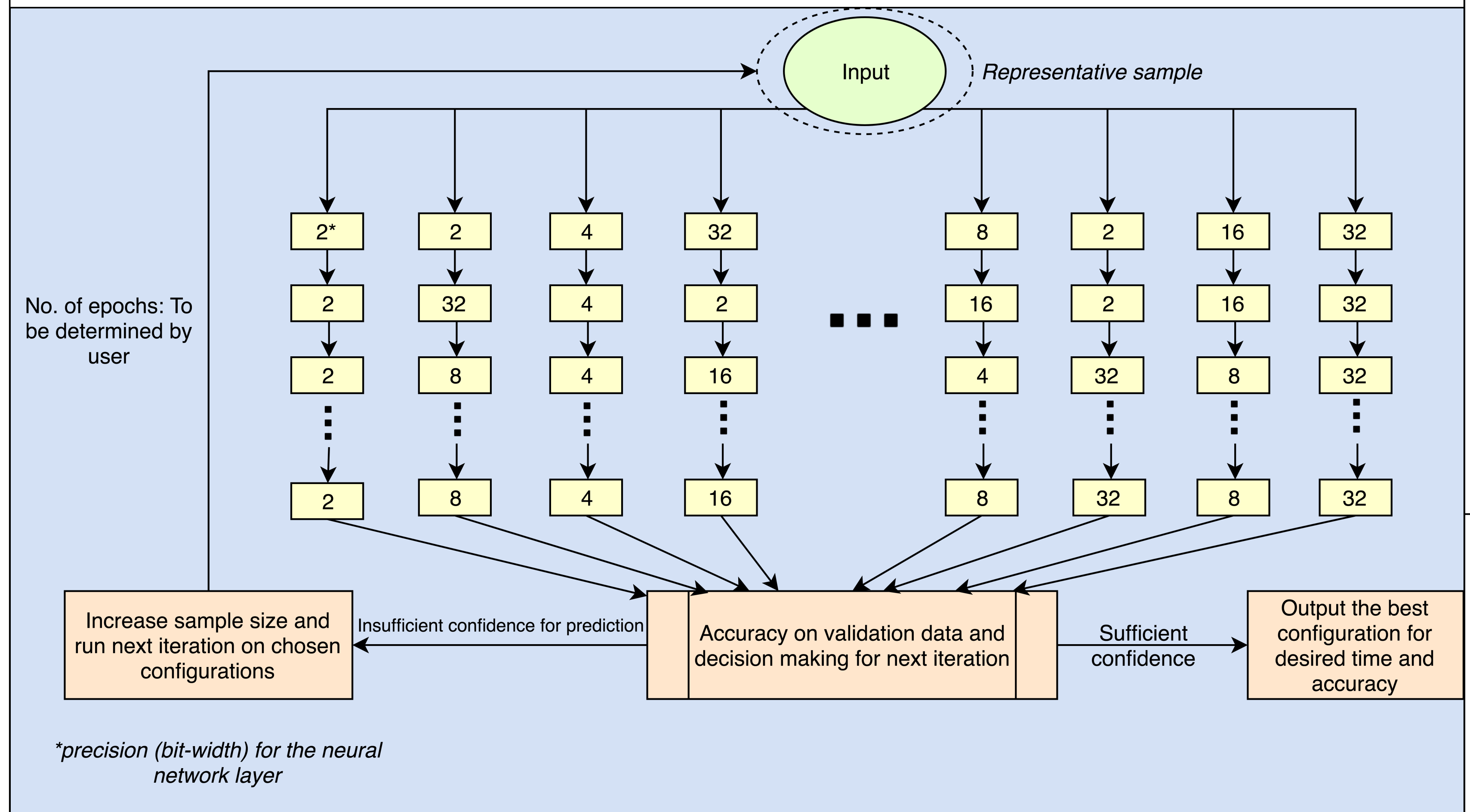
TOWARDS AN ADAPTIVE FINE-GRAINED FRAMEWORK TO OPTIMIZE NEURAL NETWORKS

Jonathan Campbell, Dipanjan Dutta, Clark Verbrugge

Problem statement

Neural networks are among the most important tools in solving problems using machine learning. However, network training time continues to rise for demanding applications. We are investigating the possibility of reducing precision of operations in order to decrease training time without a great loss in accuracy. We intend to propose an adaptive JIT-like framework to determine optimal per-layer weight/activation precision/bit-width. This framework would perform incremental learning on a representative dataset sample, training several network models each of different precision/bit-width configurations simultaneously to determine optimal values to be used for the full training period.


Input
Representative sample
No. of epochs: To be determined by user
2* 2 2 2
2 32 8 8
4 4 4 4
32 2 16 16
...
8 16 4 8
2 2 32 32
16 16 8 8
32 32 32 32
Increase sample size and run next iteration on chosen configurations
Insufficient confidence for prediction
Accuracy on validation data and decision making for next iteration
Sufficient confidence
Output the best configuration for desired time and accuracy
*precision (bit-width) for the neural network layer

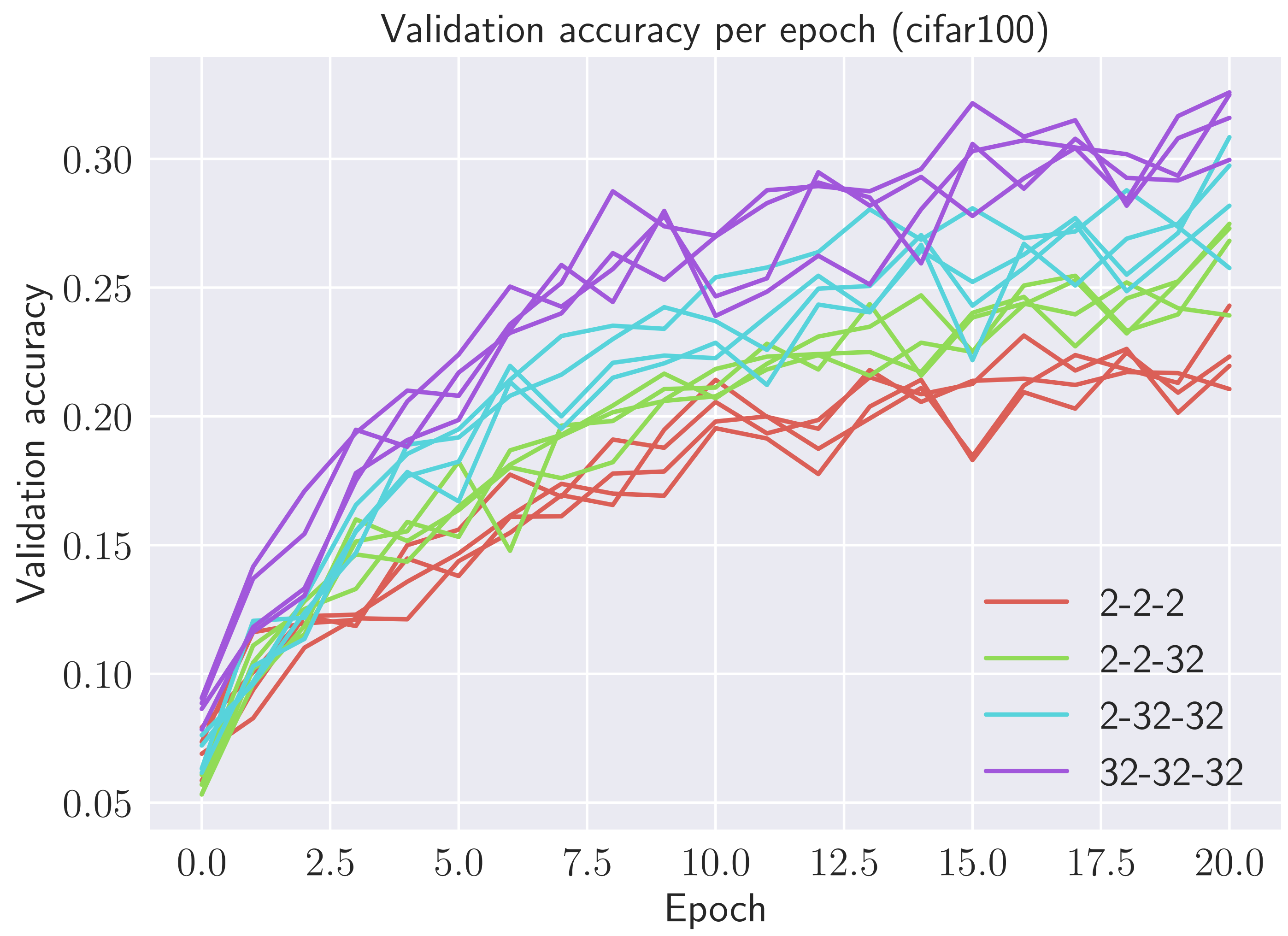
Our approach for adaptive neural network configuration framework

Research Questions

- Can we determine approximate accuracy of final model quickly enough?
- What metrics do we use to determine accuracy of final model during training?

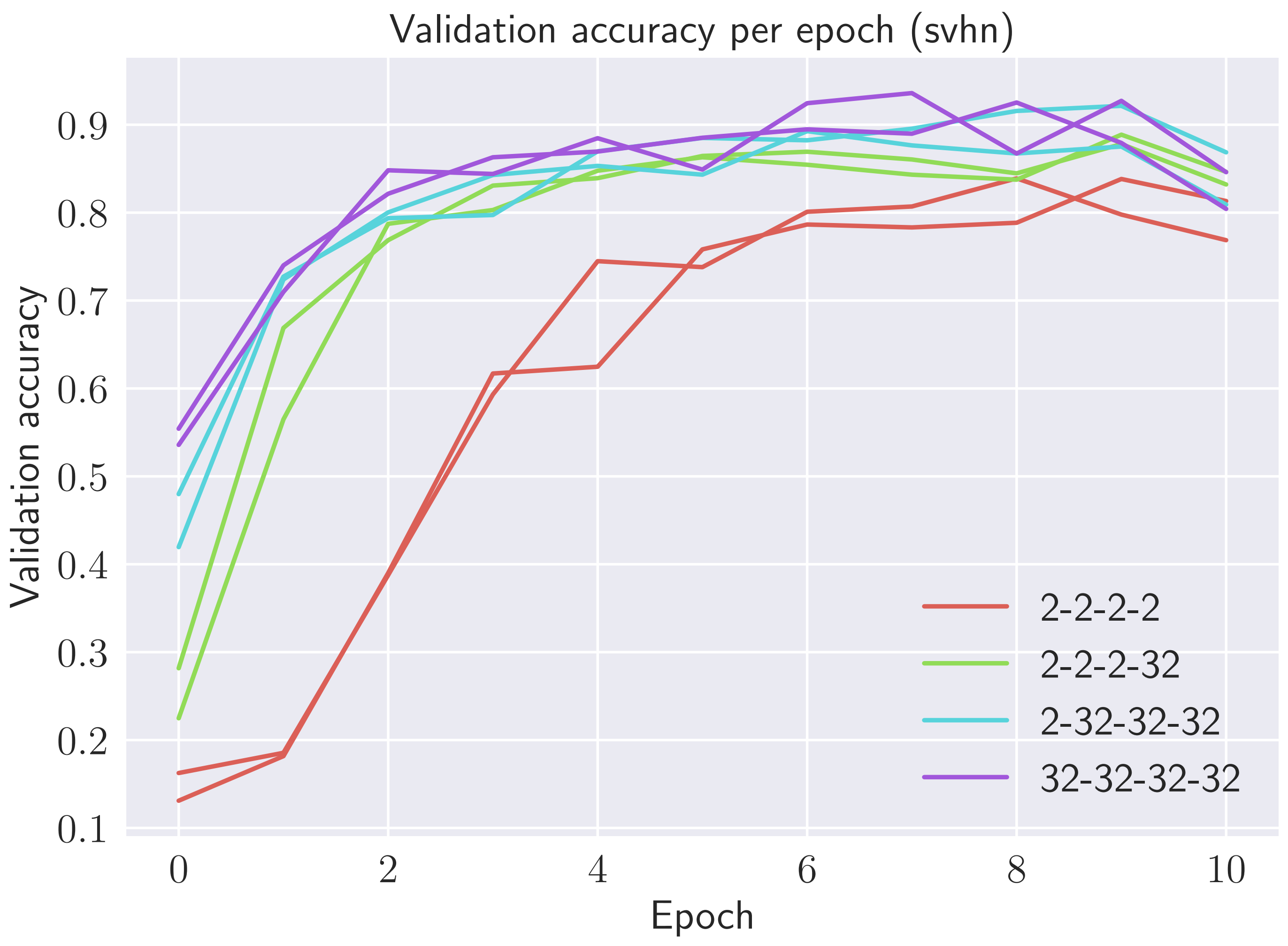
Initial Results

Current results show the possibility of using early training results through train/test accuracy correlation. Below we show validation accuracy during training for two popular datasets. Each colour represents a different combination of layer precisions (precisions were set using the QuantizedNeuralNetwork framework [1]). These preliminary results show correlation of early-epoch validation accuracy with final epoch (test) accuracy, although there can be large differences depending on dataset.



CIFAR-100

3 convolutional layers + 1 32-bit dense layer



SVHN

1 convolutional + 2 x 3 convolutional + 1 32-bit dense layer

Future Work

In future, we will continue to pursue our research questions with a focus on the generalizability of the shown results on other types of datasets and network architectures. We also plan to look into automatic adaptation of neural network architecture itself by adding or removing layers or nodes on the fly, as well as to provide an efficient way to decide on the optimal configuration that provides the best tradeoff between time and accuracy.

Acknowledgements

This research has been funded in part or completely by the Computing Hardware for Emerging Intelligent Sensory Applications (COHESA) project. COHESA is financed under the National Sciences and Engineering Research Council of Canada (NSERC) Strategic Networks grant number NETGP485577-15.

References

[1] B. Moons et al. "Minimum Energy Quantized Neural Networks." Asilomar Conference on Signals, Systems and Computers, 2017

NSERC CRSNG

McGill

COHESA