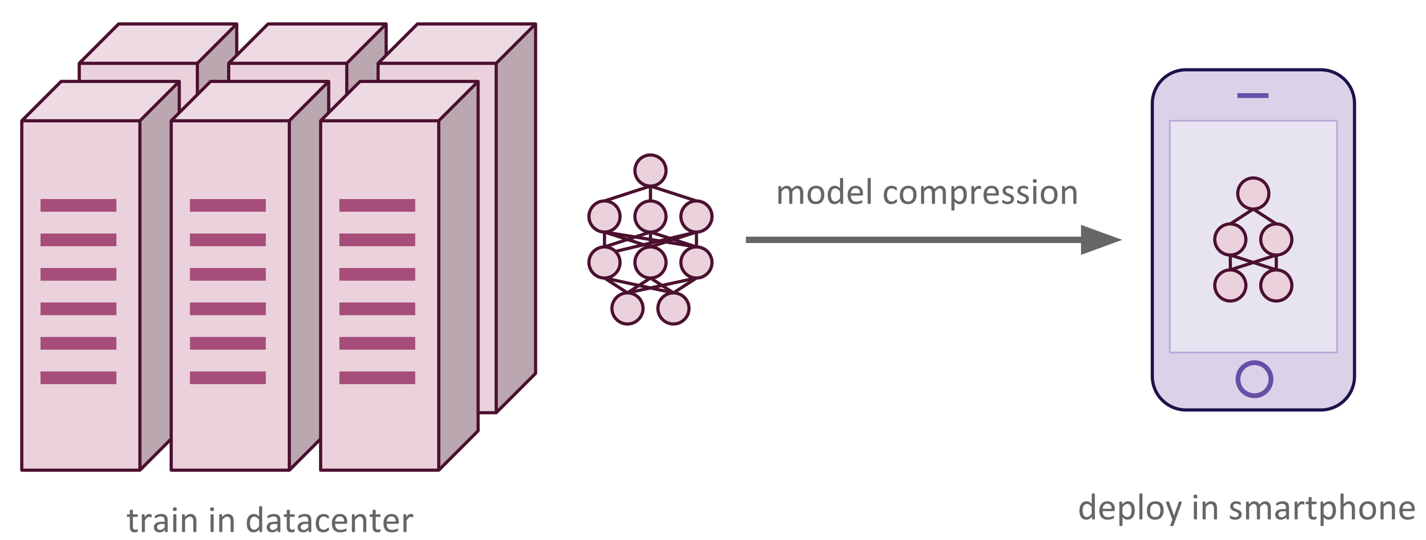
Neural Network Knowledge Distillation with No Training Data

Raphael Gontijo Lopes   Thad Starner

Georgia Institute of Technology | School of Interactive Computing | Ubicomp Lab

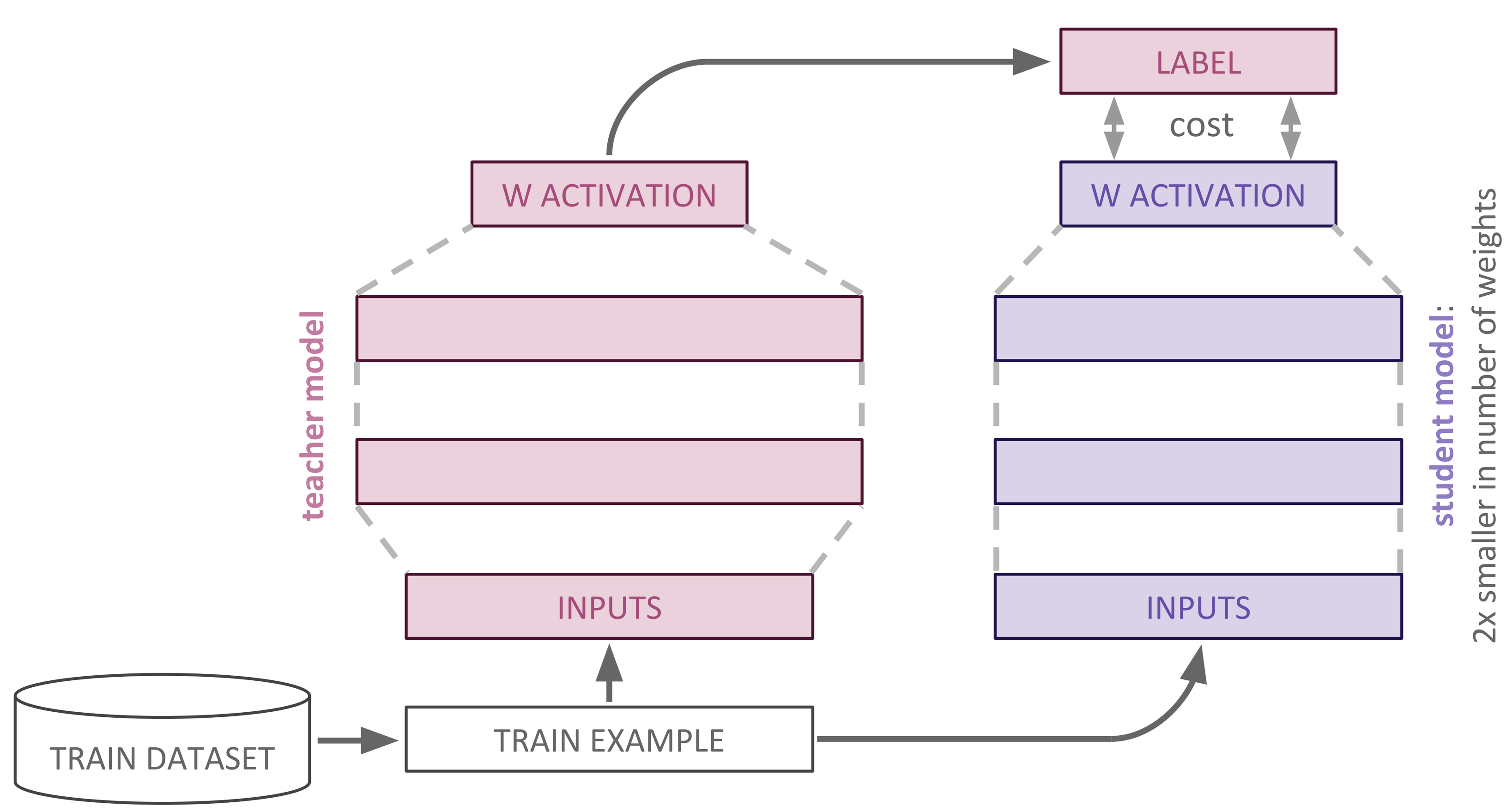
INTRODUCTION

Model compression: important for deployment in embedded systems, where memory and computation is limited.



Why not train the smaller architecture directly?
Training bigger deep learning models leads to better accuracies, due to techniques like dropout, which enables generalization. Smaller models can theoretically learn these functions [1], but training is hard.

Knowledge Distillation [2]: training a student model on the weighted activations of a teacher model on the train set.



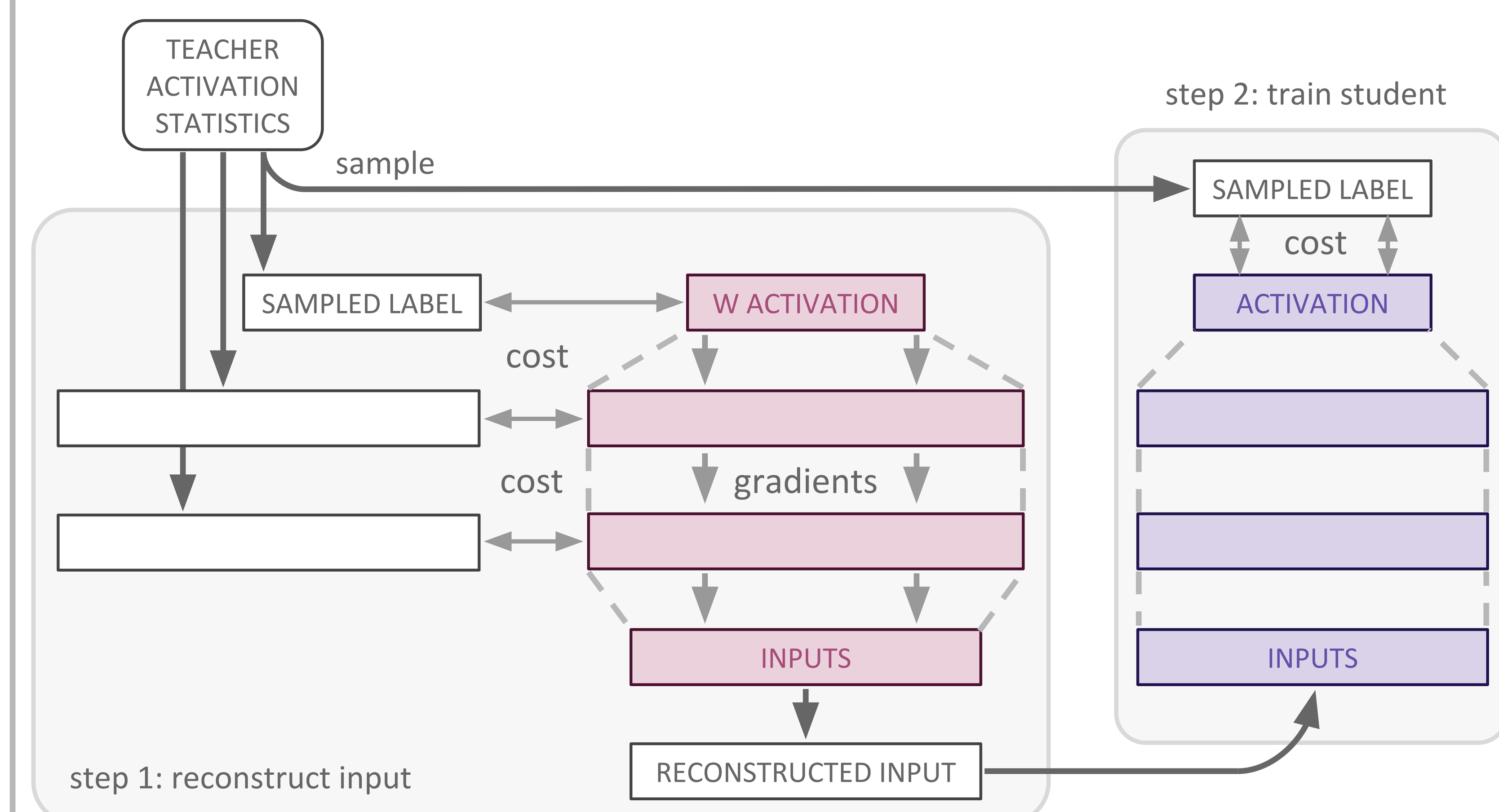
The weighted activations carry more information about how the teacher model generalizes.

METHOD

As datasets get larger, their release becomes prohibitively expensive. Even when a big dataset is released [3], it usually represents a small subset of a much larger internal dataset, used to train many of state of the art models.

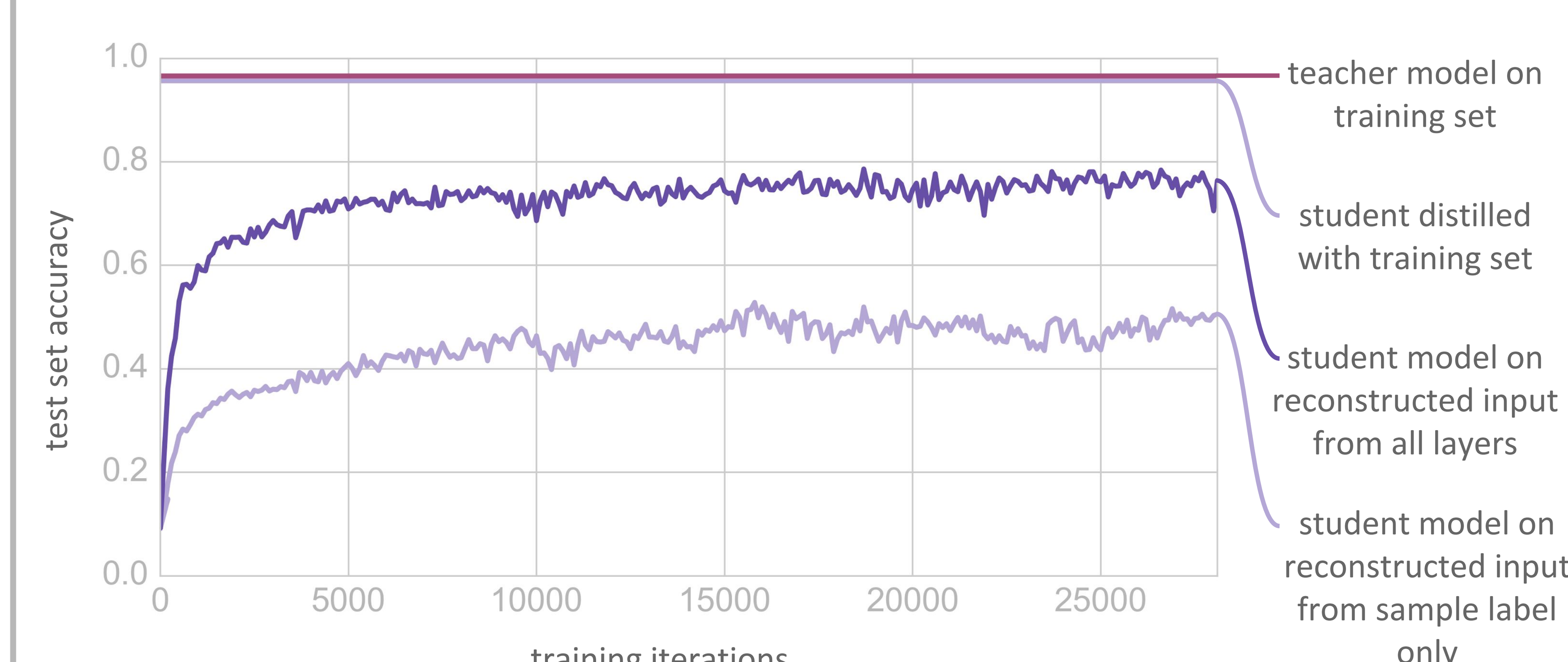
Problem: is there metadata can be provided with a pre-trained model to enable more efficient compression, even when no training data is available?

Idea: keep per-layer activation statistics for the teacher model. Reconstruct an input that matches those statistics using Gradient Descent. Inspired by Draelos et al. [4].

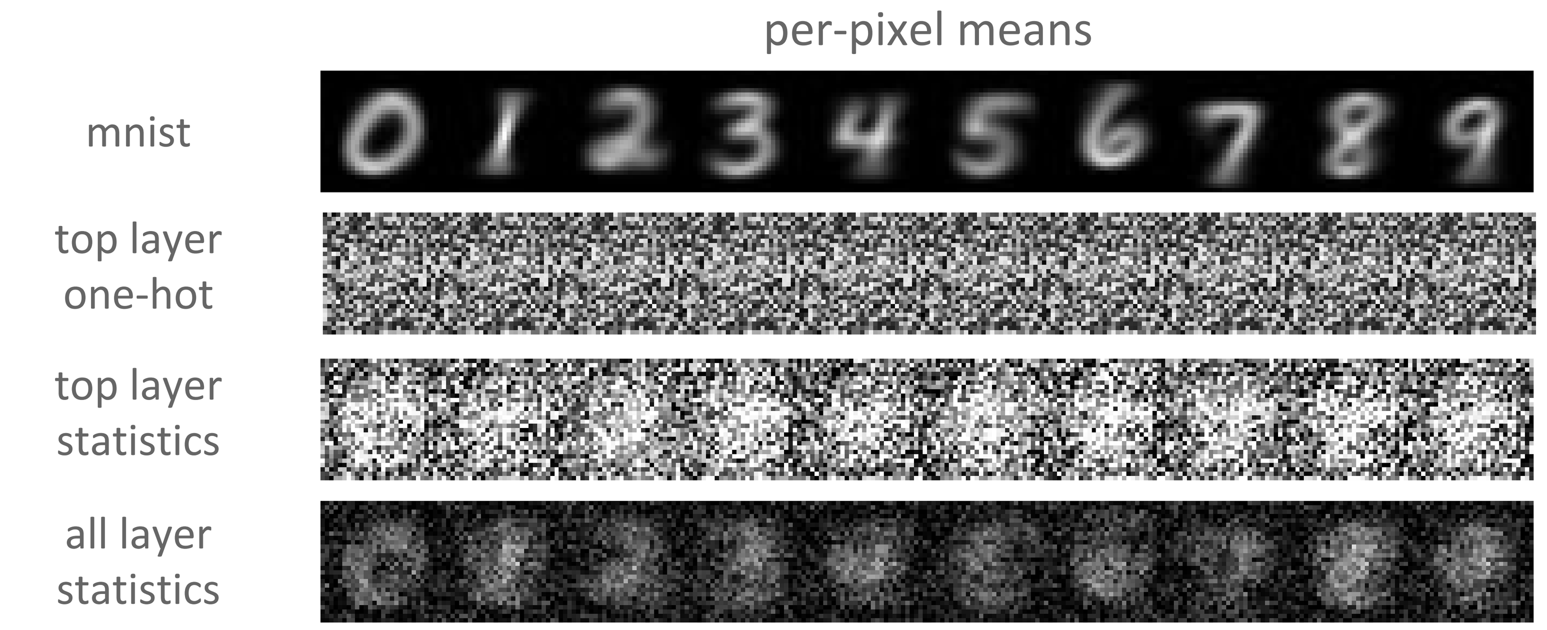


Since the input to the student was reconstructed from the teacher, it should ideally provide even more information about how the teacher generalizes.

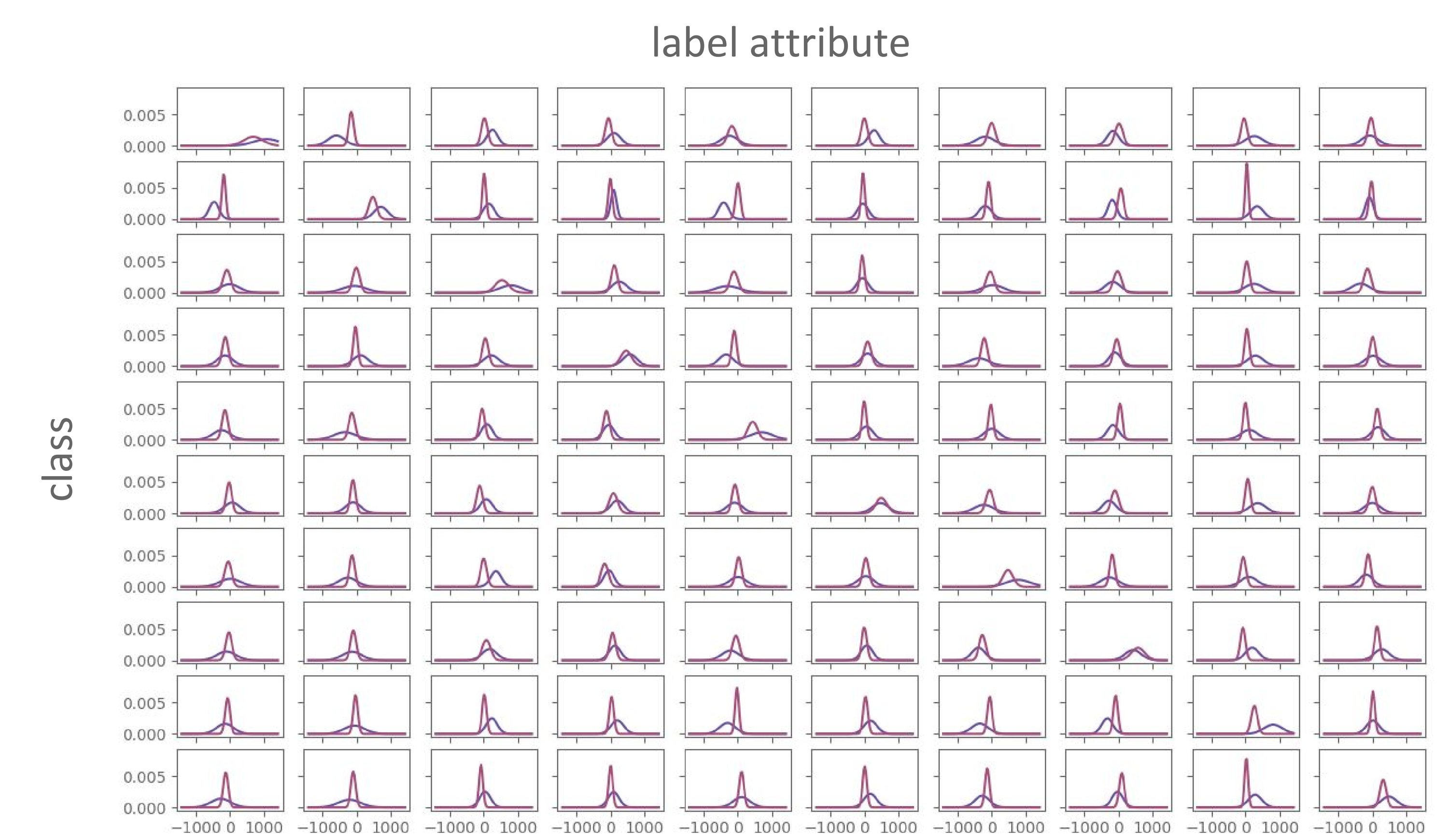
RESULTS



Input reconstruction works best when statistics for all layers (rather than just top layer) are stored and optimized for:



Teacher / Student activations on train set:



CONCLUSION

We have shown there is a relatively small amount of information that, if added to a release of a pre-trained model, can facilitate network compression.

We have also shown that it is possible to compress a network with no access to the original training set and have motivated further exploration into distribution formats for deep models.

REFERENCES

[1] Hornik, Kurt (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks, 4*, 251-257.

[2] Hinton, G. Vinyals, O. and Dean, J. Distilling knowledge in a neural network. In *Deep Learning and Representation Learning Workshop, NIPS*, 2014.

[3] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan & Sudheendra Vijayanarasimhan (2016). YouTube-8M: A Large-Scale Video Classification Benchmark. *CoRR, abs/1609.08675*, .

[4] Timothy J. Draelos, Nadine E. Miner, Christopher C. Lamb, Craig M. Vineyard, Kristofor D. Carlson, Conrad D. James & James B. Aimone (2016). Neurogenesis Deep Learning. *CoRR, abs/1612.03770*, .