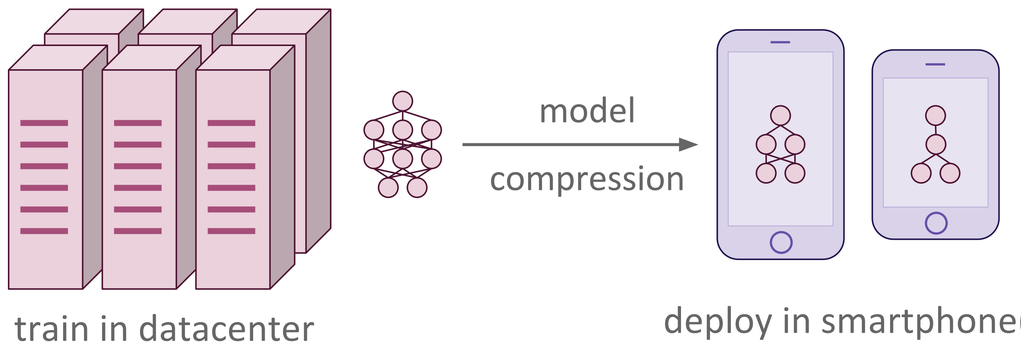
Data-Free Knowledge Distillation for Deep Neural Networks

Raphael Gontijo Lopes Stefano Fenu Thad Starner

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

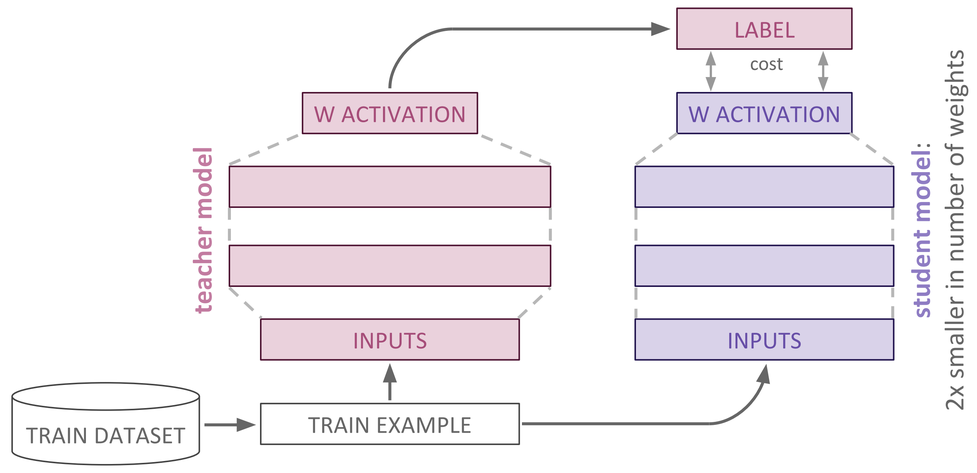
INTRODUCTION

Model compression: important for deployment in embedded systems, where memory and computation is limited. Usually involves compressing into specific hardware-optimized models.



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.

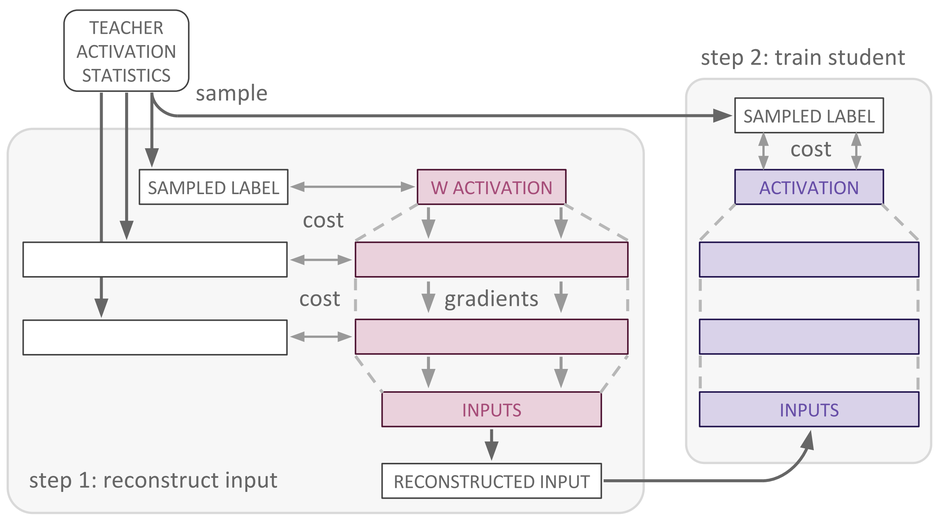
METHOD

Many datasets encounter hurdles to release in the form of privacy or security concerns, as in the case of biometric and sensitive data.

Yet, other entities might still want to compress models that have been pre-trained on it for hardware-optimized models.

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

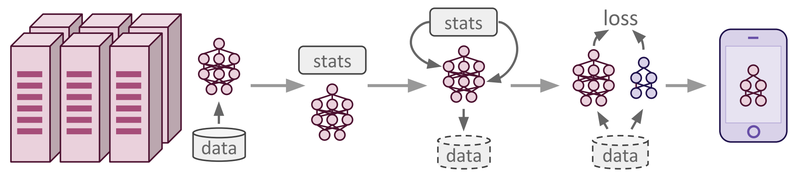
Idea: keep per-layer activation records for the teacher model. Reconstruct an input that matches those statistics using Gradient Descent. Inspired by Draelos et al. [3].



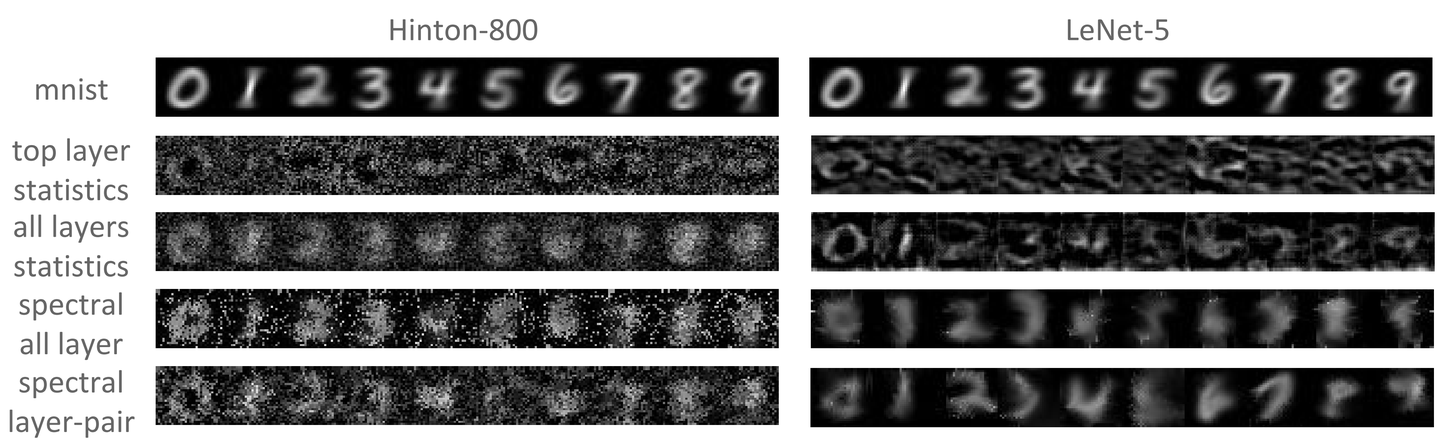
Since the input to the student was reconstructed from the teacher, it should ideally provide even more information about how the teacher generalizes.

We also use spectral methods to preserve inter-layer dynamics, which proved to be much more high-performing, with the tradeoff for size of records produced.

The final pipeline becomes:



RESULTS



(per-pixel means)

Model	HINTN-1200	HINTN-800	LENET-5	LENET-5	ALEXNET	ALEXNET
Dataset	MNIST	MNIST	MNIST	MNIST	CelebA	CelebA
Train	96.95%	95.70%	98.91%	98.65%	80.82%	
Distill		95.74%		98.91%		
Top Stats		68.75%		77.30%		54.12%
All Stats		76.38%		85.61%		
Spec All		89.41%		90.28%		77.56%
Spec Pair		91.24%		90.28%		76.94%

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] 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*, .