Machine Learning 2

DS 4420 - Spring 2020

Green AI

Byron C. Wallace

Today

- *Green Artificial Intelligence*: The surprisingly large carbon footprint of modern ML models and what we might do about this

The problem

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
Transformer$_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer$_{big}$	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT$_{base}$	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT$_{base}$	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Cost of development

"The sum GPU time required for the project totaled 9998 days (27 years)

		Estimated cost (USD)	
Models	Hours	Cloud compute	Electricity
1	120	$52–$175	$5
24	2880	$1238–$4205	$118
4789	239,942	$103k–$350k	$9870

Table 4: Estimated cost in terms of cloud compute and electricity for training: (1) a single model (2) a single tune and (3) all models trained during R&D.

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Conclusions

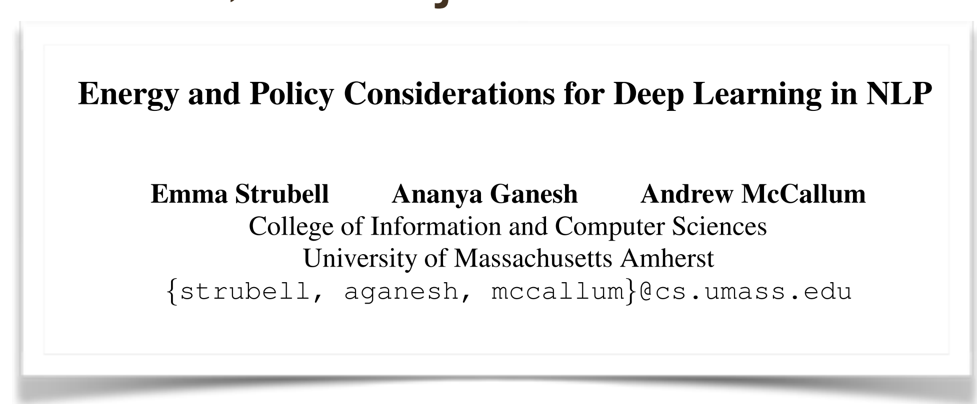
- Researchers should report training time and hyper parameter sensitivity
  - ★ And practitioners should take these into consideration

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Conclusions

- Researchers should report training time and hyper parameter sensitivity
  - ★ And practitioners should take these into consideration
- We need new, more efficient methods; not just ever larger architectures!



Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Towards Green AI

Green AI

Roy Schwartz[*◊] Jesse Dodge[*◊♣] Noah A. Smith[◊♡] Oren Etzioni[◊]

[◊]Allen Institute for AI, Seattle, Washington, USA
[♣] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[♡] University of Washington, Seattle, Washington, USA

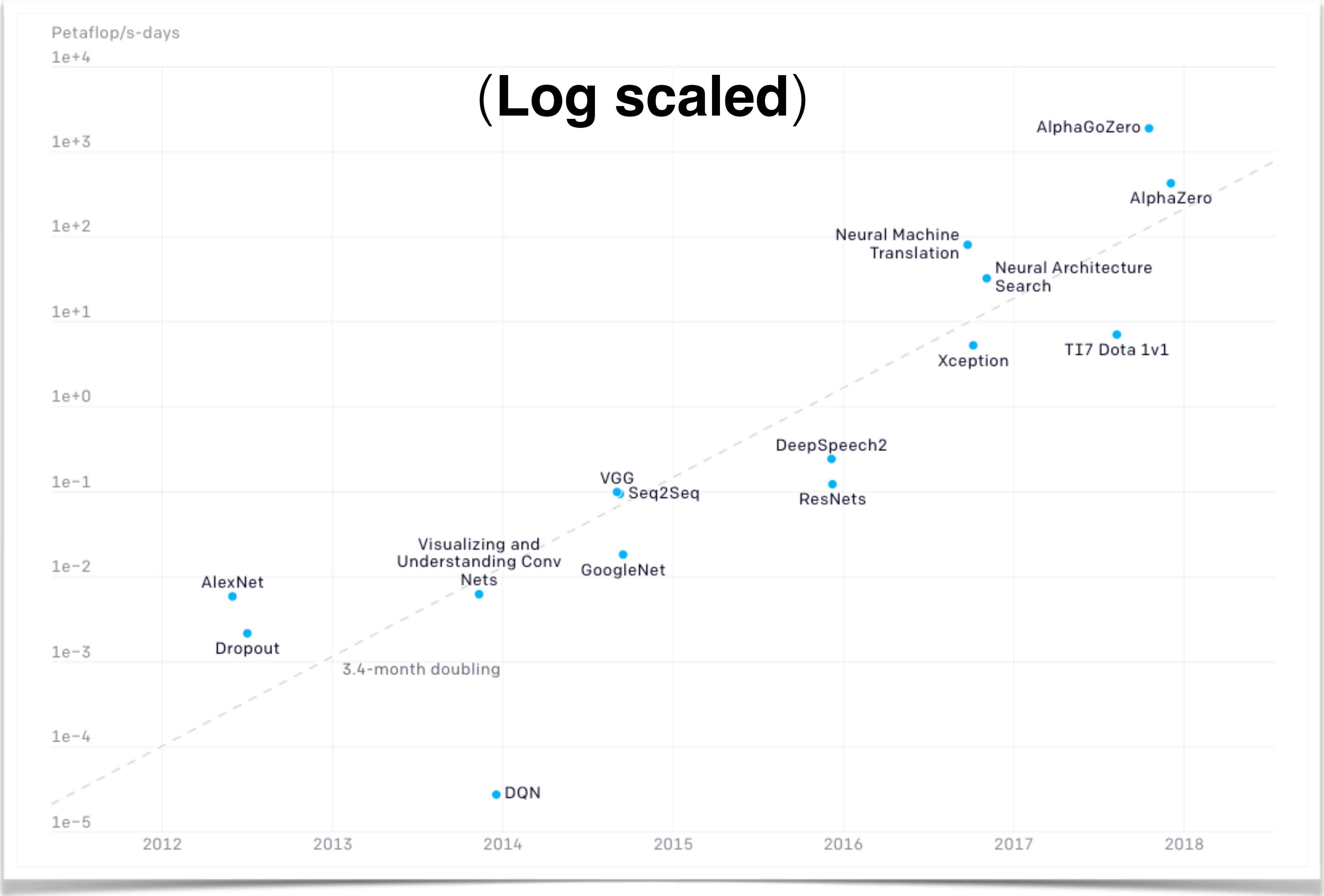
Towards Green AI

- Argues for a pivot toward research that is environmentally friendly and inclusive; not just dominated by huge corporations with unlimited compute

Green AI

Roy Schwartz[*◊] Jesse Dodge[*◊♣] Noah A. Smith[◊♡] Oren Etzioni[◊]

[◊]Allen Institute for AI, Seattle, Washington, USA
[♣] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[♡] University of Washington, Seattle, Washington, USA



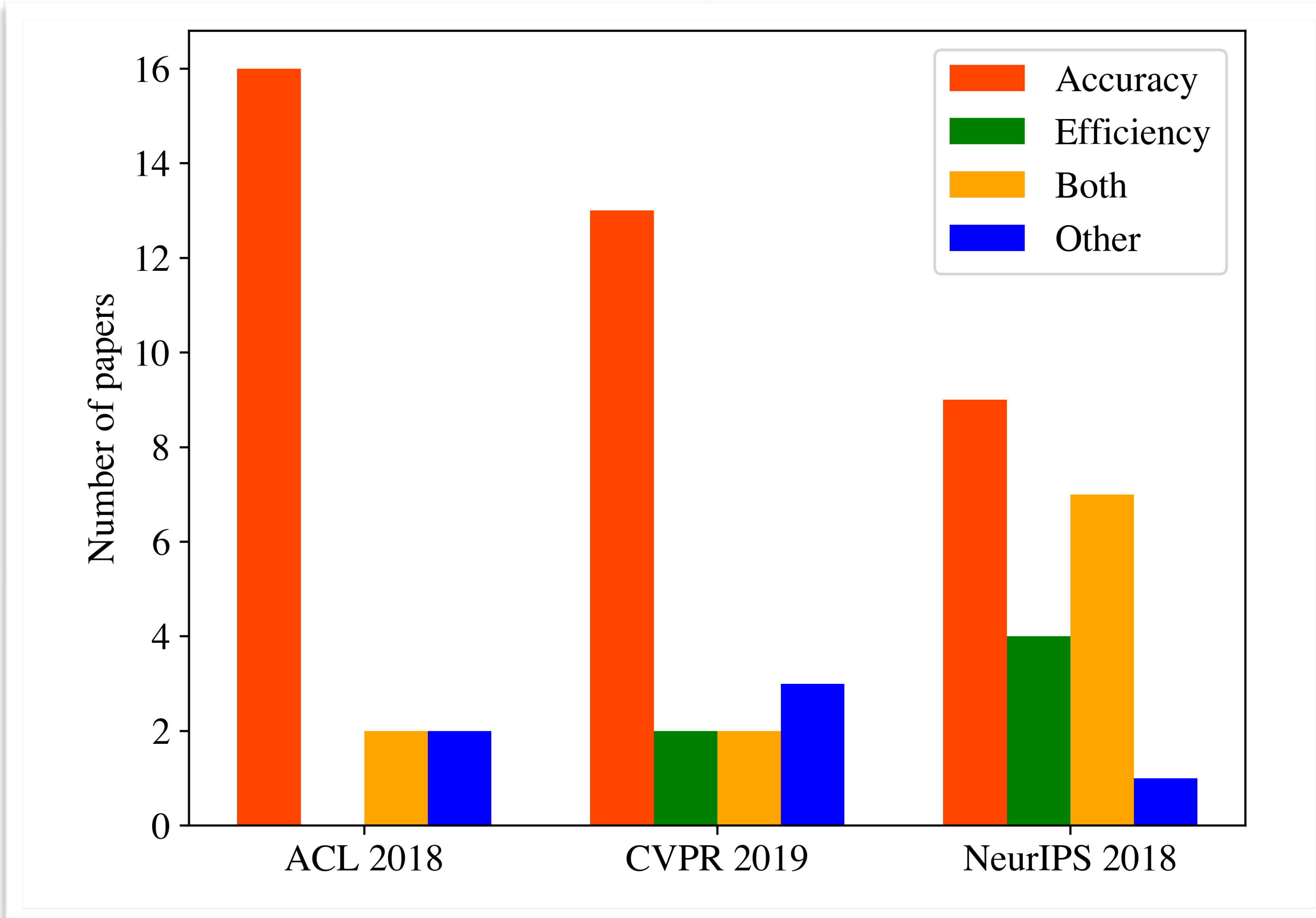
https://openai.com/blog/ai-and-compute/

Dario Amodei & Danny Hernandez

(ORIGINAL POST)

Girish Sastry, Jack Clark, Greg Brockman & Ilya Sutskever

(ADDENDUM)



Does the community care about efficiency?

Green AI

Roy Schwartz* ♢ Jesse Dodge*♢♣ Noah A. Smith♢♡ Oren Etzioni♢

♢Allen Institute for AI, Seattle, Washington, USA
♣ Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
♡ University of Washington, Seattle, Washington, USA

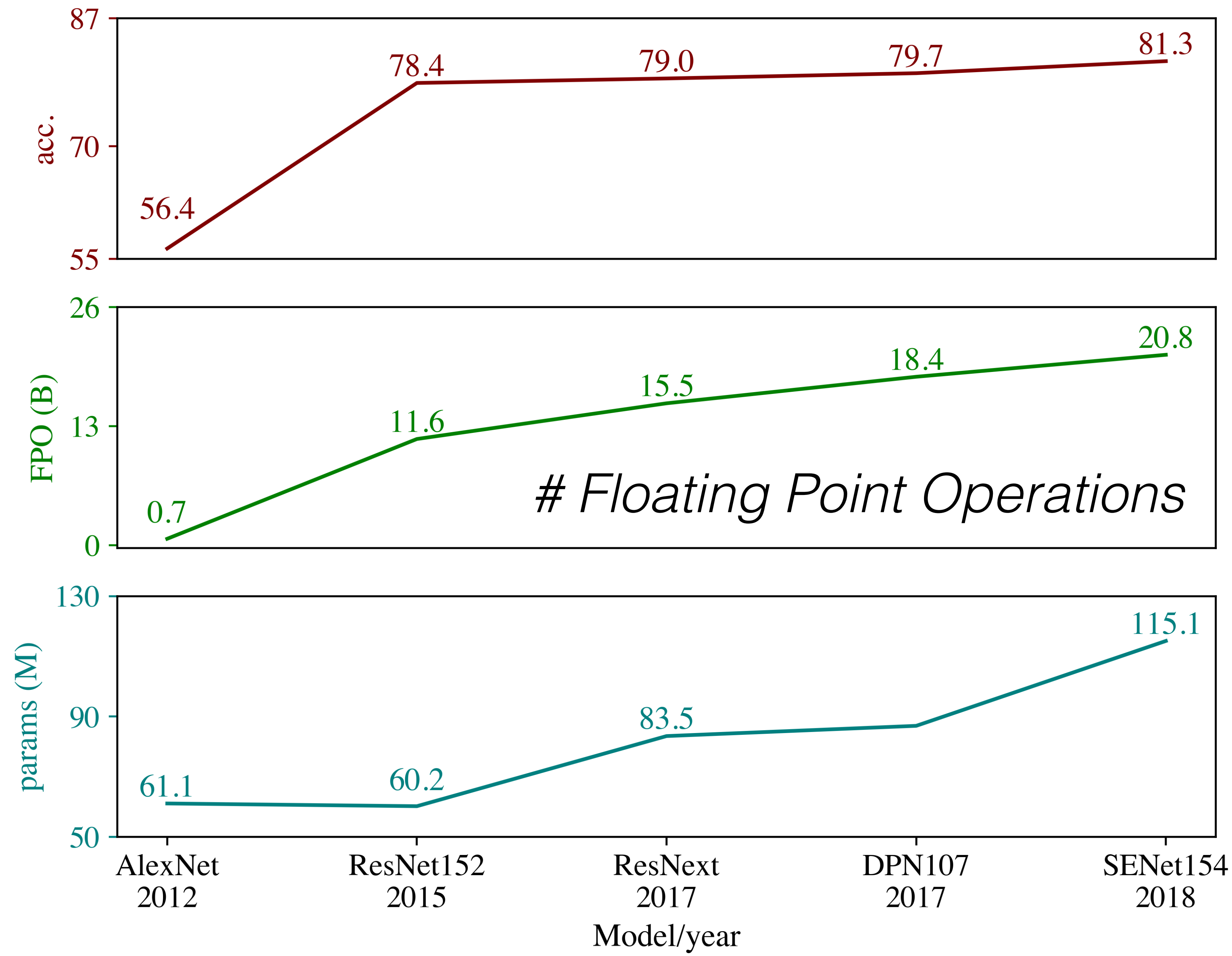
$$\boxed{Cost(R) \propto E \cdot D \cdot H}$$

Equation 1: The equation of Red AI: The cost of an AI (R)esult grows linearly with the cost of processing a single (E)xample, the size of the training (D)ataset and the number of (H)yperparameter experiments.

Green AI

Roy Schwartz[*◊] Jesse Dodge[*◊♣] Noah A. Smith[◊♡] Oren Etzioni[◊]

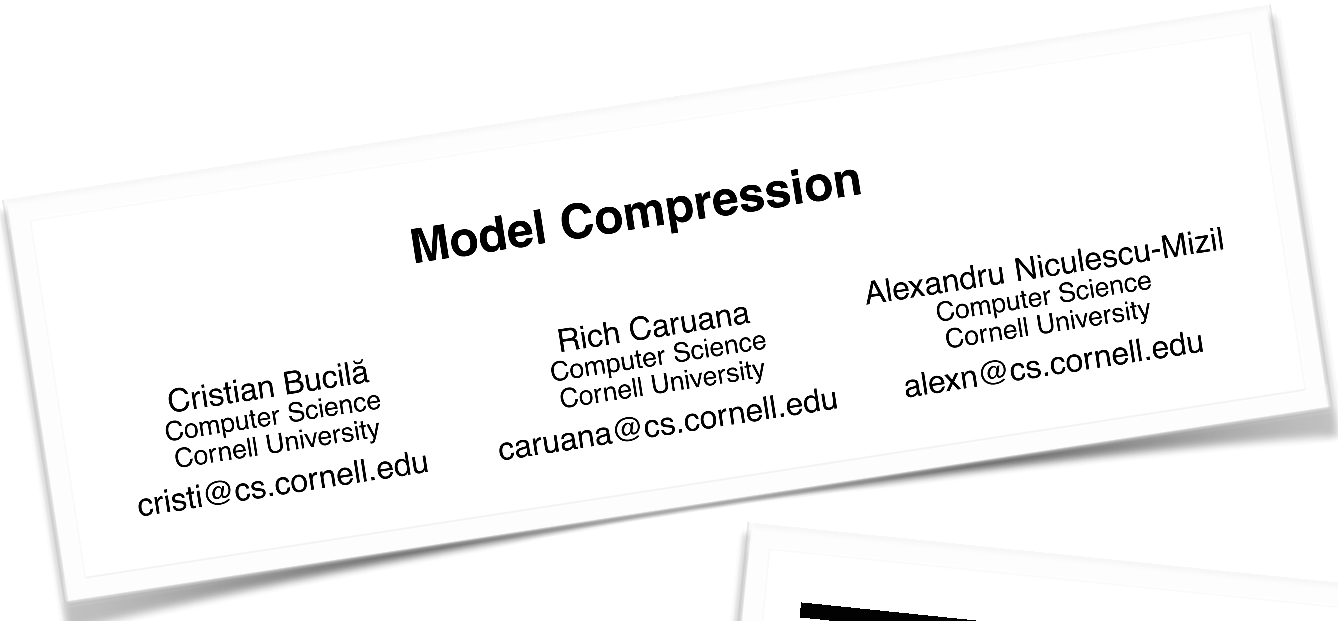
[◊]Allen Institute for AI, Seattle, Washington, USA
[♣] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[♡] University of Washington, Seattle, Washington, USA



(a) Different models.

Large increase in FPO —> Small gains in acc

Model distillation/compression

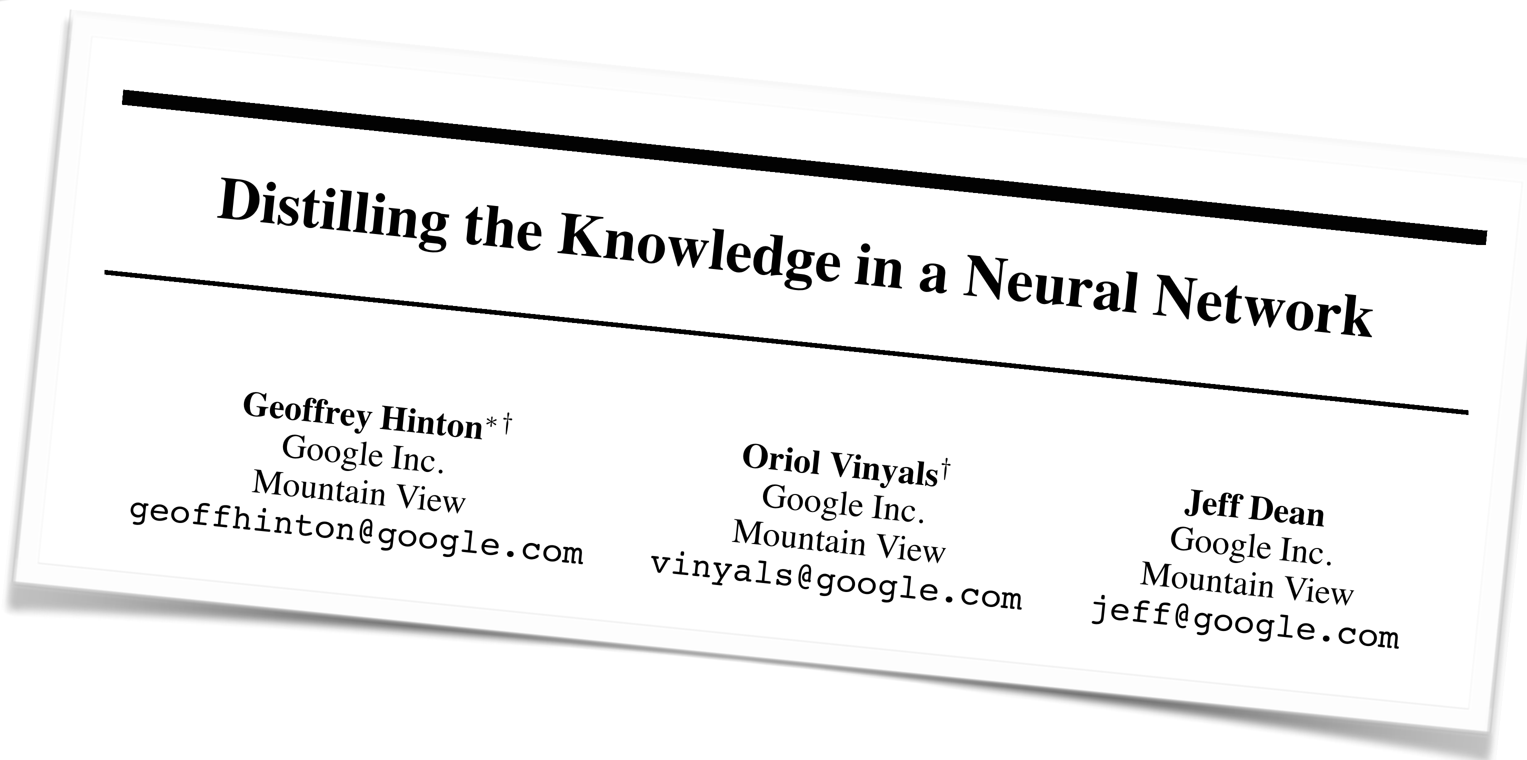


Model Compression

Cristian Bucilă
Computer Science
Cornell University
cristi@cs.cornell.edu

Rich Caruana
Computer Science
Cornell University
caruana@cs.cornell.edu

Alexandru Niculescu-Mizil
Computer Science
Cornell University
alexn@cs.cornell.edu



Distilling the Knowledge in a Neural Network

Geoffrey Hinton[*†]
Google Inc.
Mountain View
geoffhinton@google.com

Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

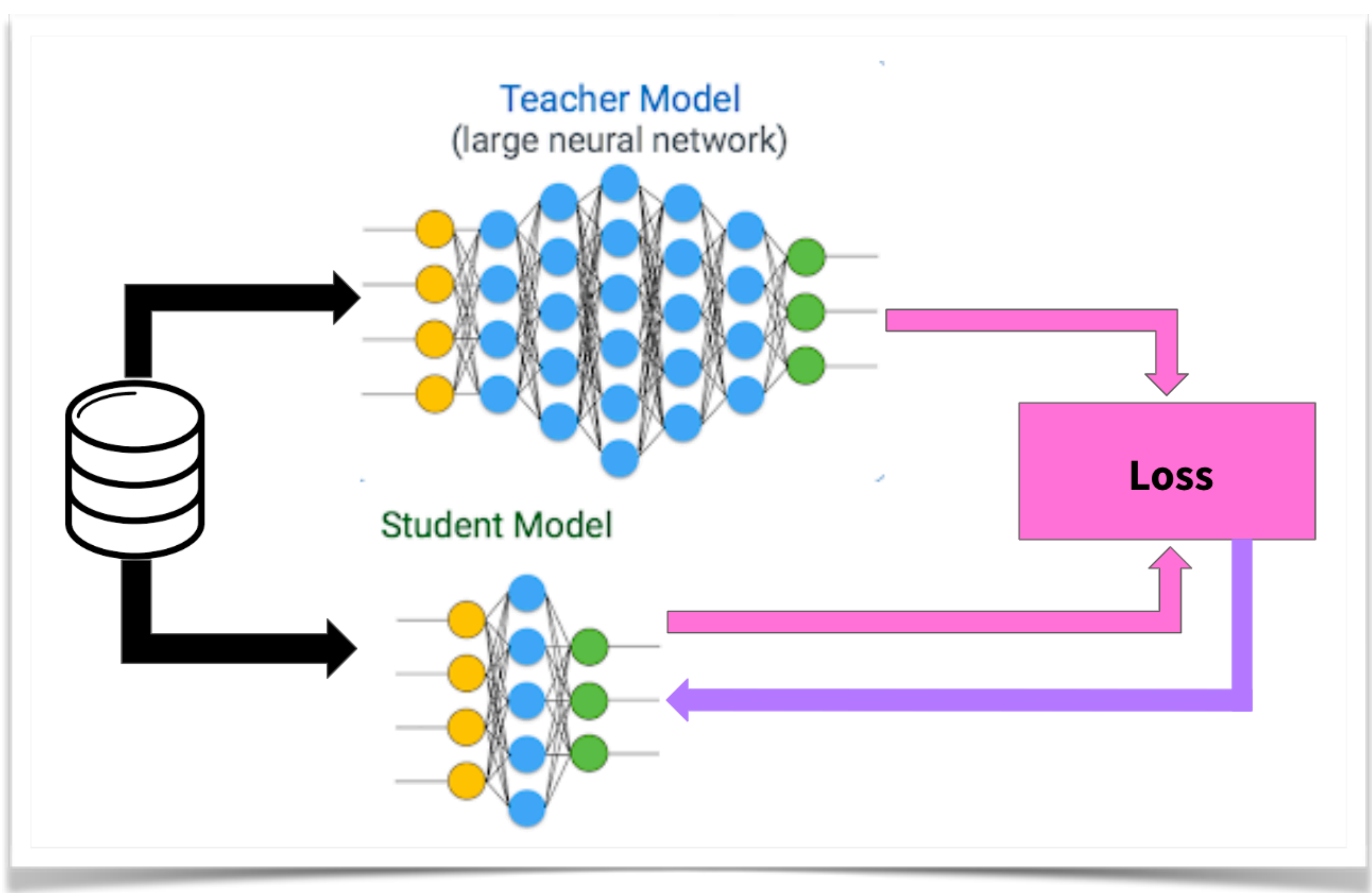
Jeff Dean
Google Inc.
Mountain View
jeff@google.com

Model distillation

Idea: Train a smaller model (**the student**) *on the predictions/outputs of a larger model* (**the teacher**)

Model distillation

Idea: Train a smaller model (**the student**) *on the predictions/outputs of a larger model* (**the teacher**)



https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764


Prakhar Ganesh

Model Compression

Cristian Bucilă
Computer Science
Cornell University
cristi@cs.cornell.edu

Rich Caruana
Computer Science
Cornell University
caruana@cs.cornell.edu

Alexandru Niculescu-Mizil
Computer Science
Cornell University
alexn@cs.cornell.edu

KDD, **2006**

The idea

- Learn a "fast, compact” model (**learner**) that approximates the predictions of a big, inefficient model (**teacher**)

The idea

- Learn a "fast, compact” model (**learner**) that approximates the predictions of a big, inefficient model (**teacher**)

- Note that we have access to the **teacher** so can train the **learner** even on “unlabeled” data — we are trying to get the **learner** to mimic the **teacher**

The idea

- Learn a "fast, compact” model (**learner**) that approximates the predictions of a big, inefficient model (**teacher**)

- Note that we have access to the **teacher** so can train the **learner** even on “unlabeled” data — we are trying to get the **learner** to mimic the **teacher**

- This paper considers a bunch of ways we might generate synthetic “points” to pass through the **teacher** and use as training data for the **learner**. In many domains (e.g., language, vision) real unlabeled data is easy to find (so we do not need to generate synthetic samples)

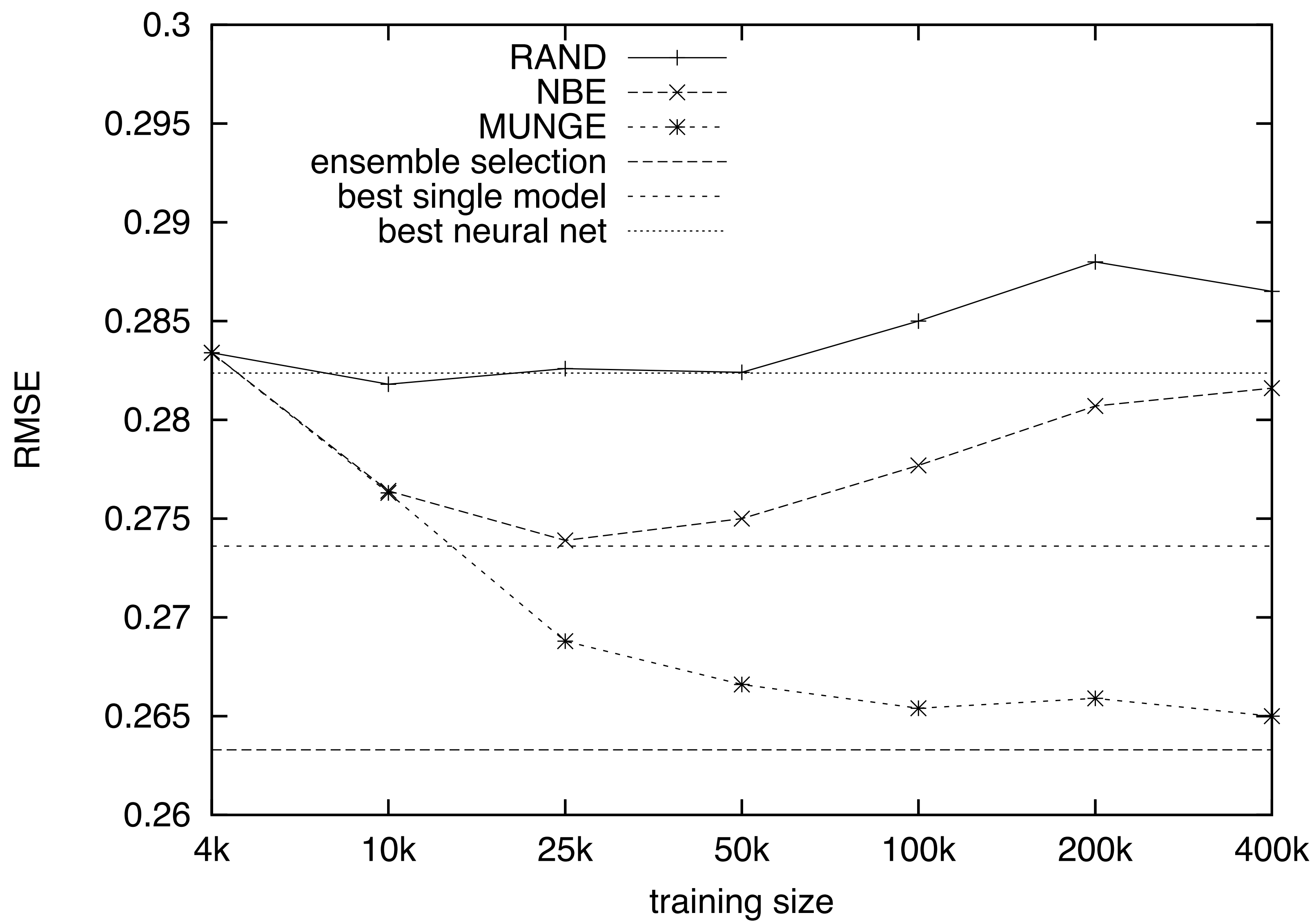


Figure 2: Average perf. over the eight problems.

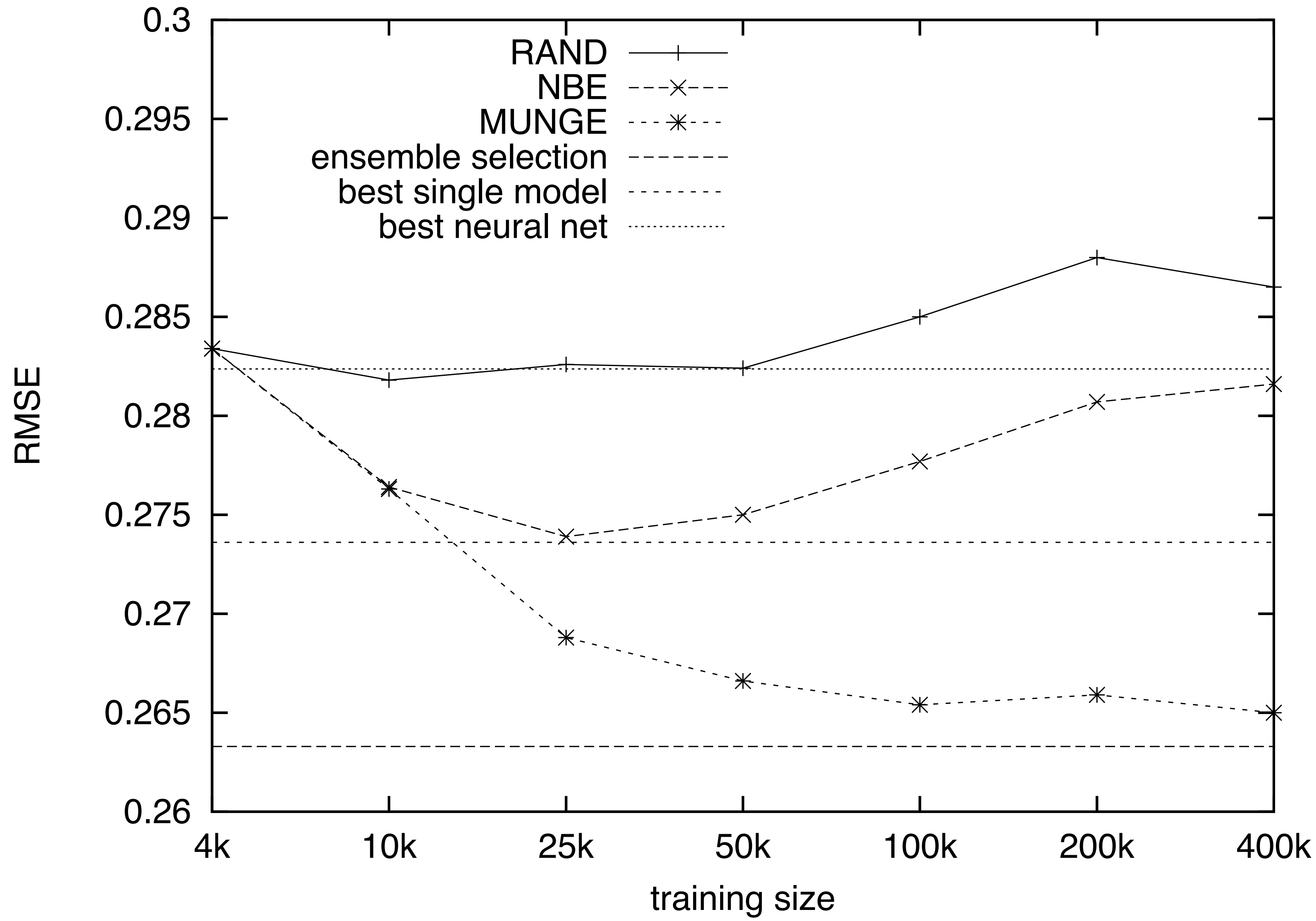
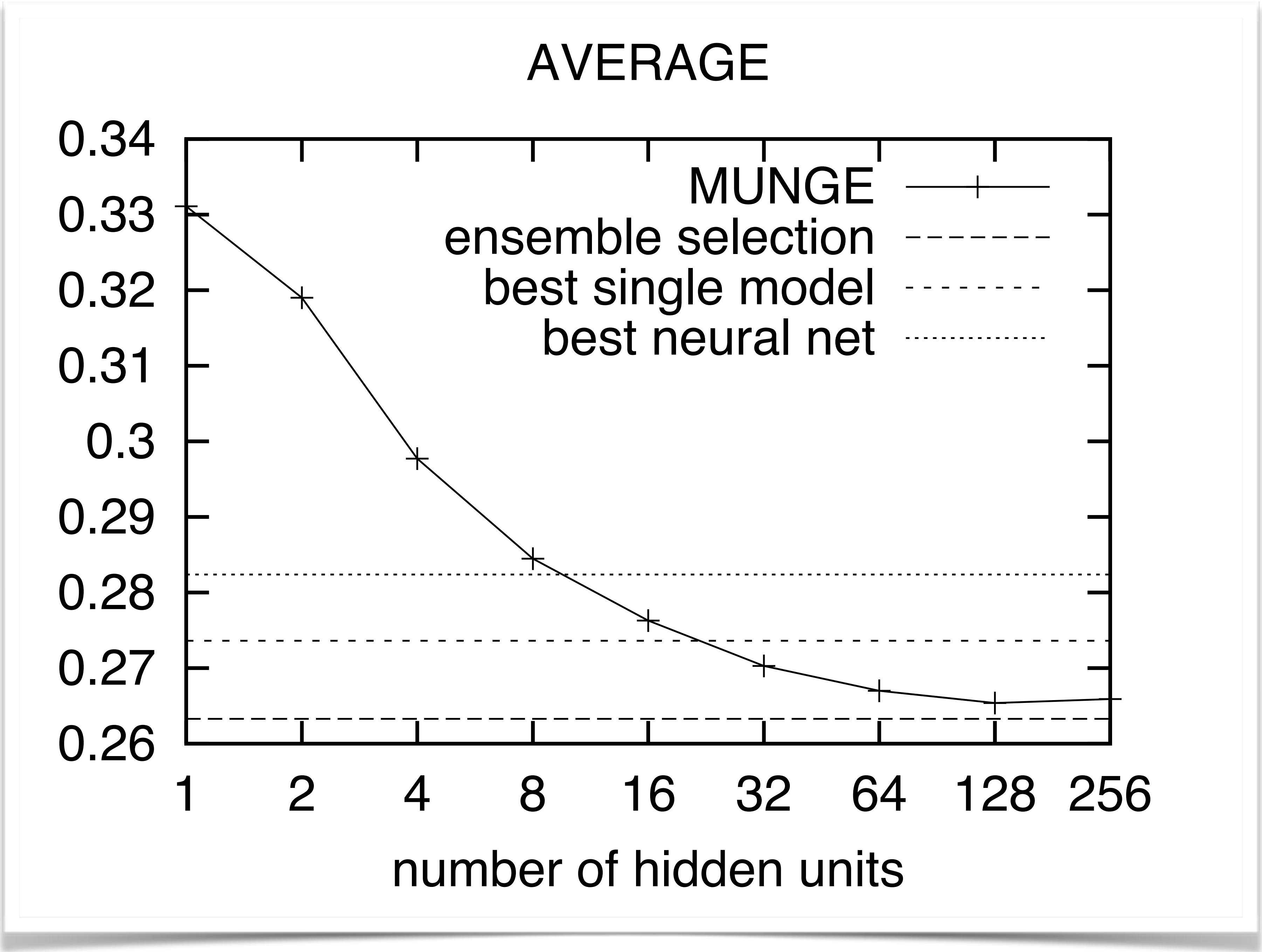


Figure 2: Average perf. over the eight problems.

We can train a neural network student to mimic a big ensemble — this does much better than net trained on labeled data only

Performance vs complexity



Time (a proxy for energy)

Table 3: Time in seconds to classify 10k cases.

	MUNGE	ENSEMBLE	ANN	SINGLE
ADULT	7.88	8560.61	3.94	48.31
COVTYPE	4.46	3440.99	1.05	37.31
HS	12.09	1817.17	3.85	3.85
LETTER.P1	2.59	1630.21	0.25	0.25
LETTER.P2	2.59	2651.95	0.74	526.34
MEDIS	4.78	190.18	2.85	2.85
MG	6.98	1220.04	1.80	53.58
SLAC	3.60	23659.03	2.85	74.48
AVERAGE	5.62	5396.27	2.17	93.37

teacher

Distilling the Knowledge in a Neural Network

Geoffrey Hinton[*†]
Google Inc.
Mountain View
geoffhinton@google.com

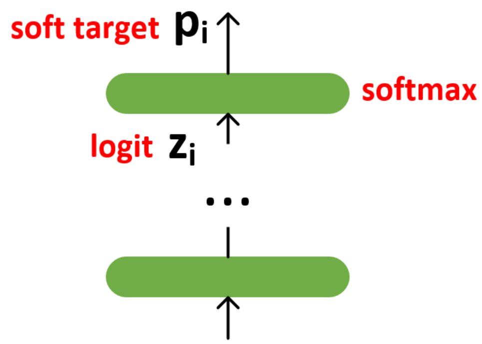
Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

Jeff Dean
Google Inc.
Mountain View
jeff@google.com

NeurIPs (workshop), **2014**

Soft targets

- The key idea is to fit the **learner** on **soft targets** (i.e., raw outputs or *logits*) from the **teacher** model



$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

z_i : the logit, i.e. the input to the softmax layer

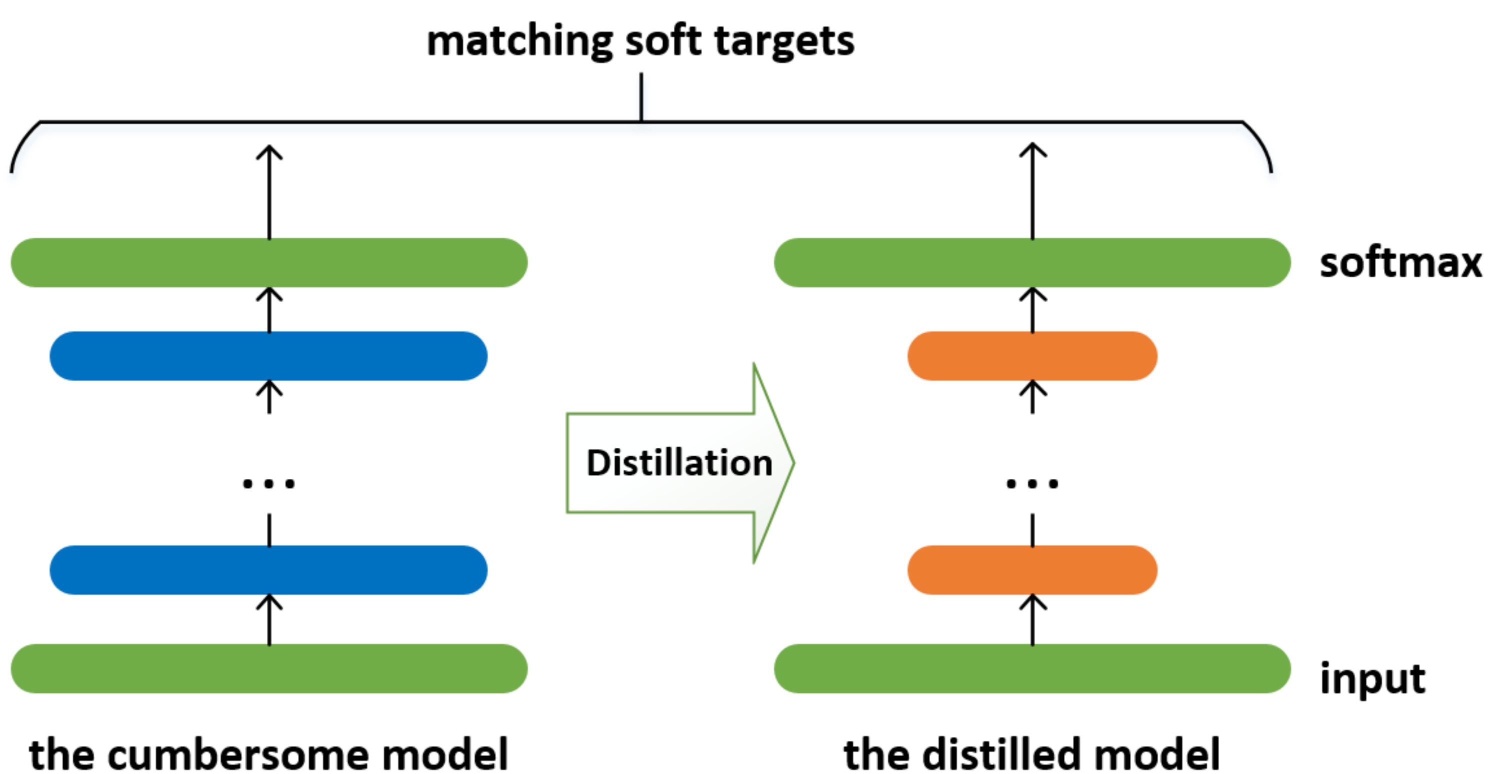
q_i : the class probability computed by the softmax layer

T : a temperature that is normally set to 1

Image from Yangyang

Soft targets

- The key idea is to fit the **learner** on **soft targets** (i.e., raw outputs or *logits*) from the **teacher** model



teacher **learner**

Image from Yangyang

System	Test Frame Accuracy
Baseline	58.9%
10xEnsemble	61.1%
Distilled Single model	60.8%

Let’s implement this…
(“in class” exercise on distillation:

"In class" exercise 3/19

Availability: Item is hidden from students. It will be available after Mar 19, 2020 12:30 PM.
Starter notebook: https://colab.research.google.com/drive/1XymnfdS4wMY3Q6aEuFedoh5lSIR-Uzrh

Pruning models

Pruning models

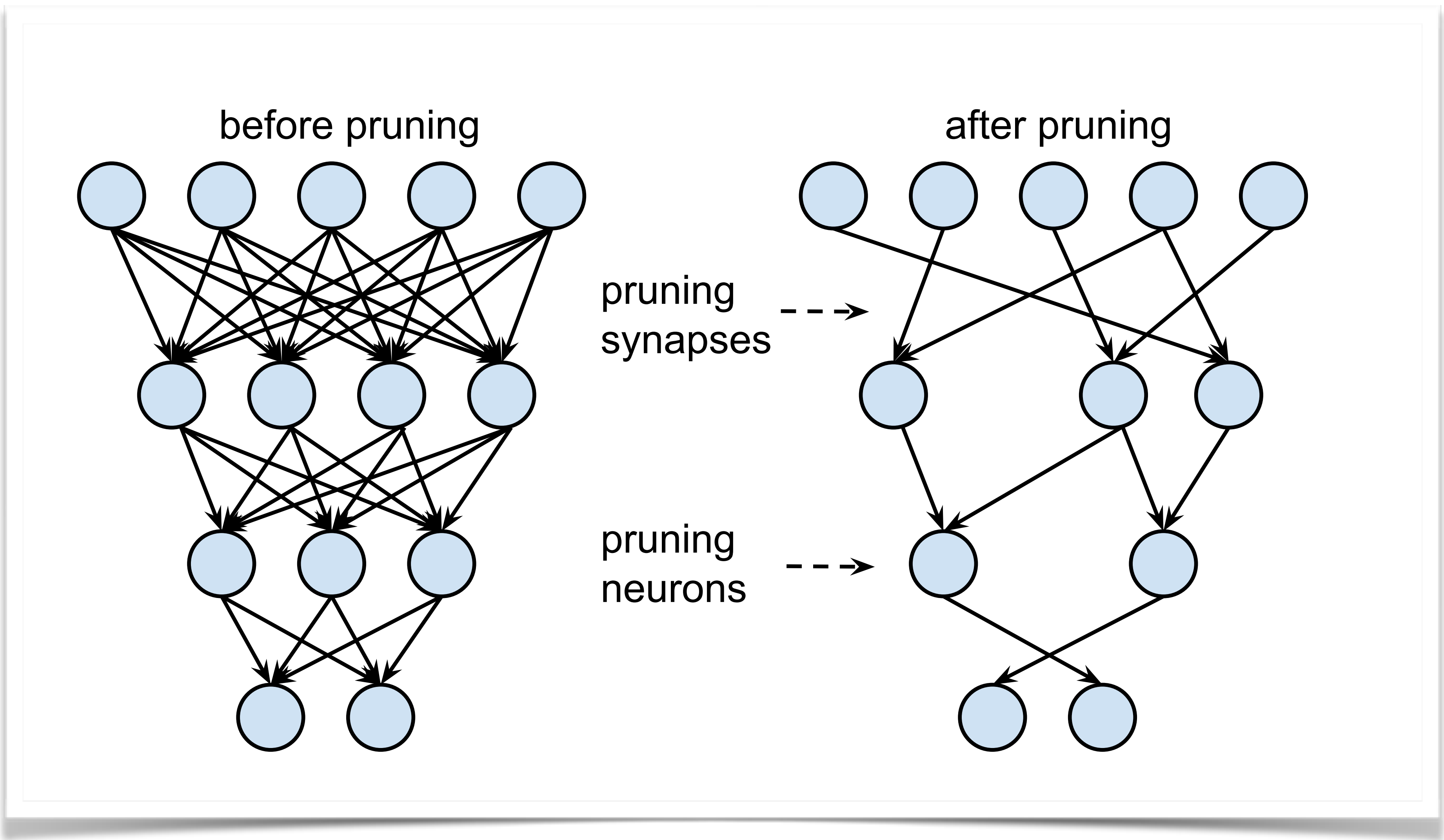


Image from Han *et al*. NeurIPs 2015

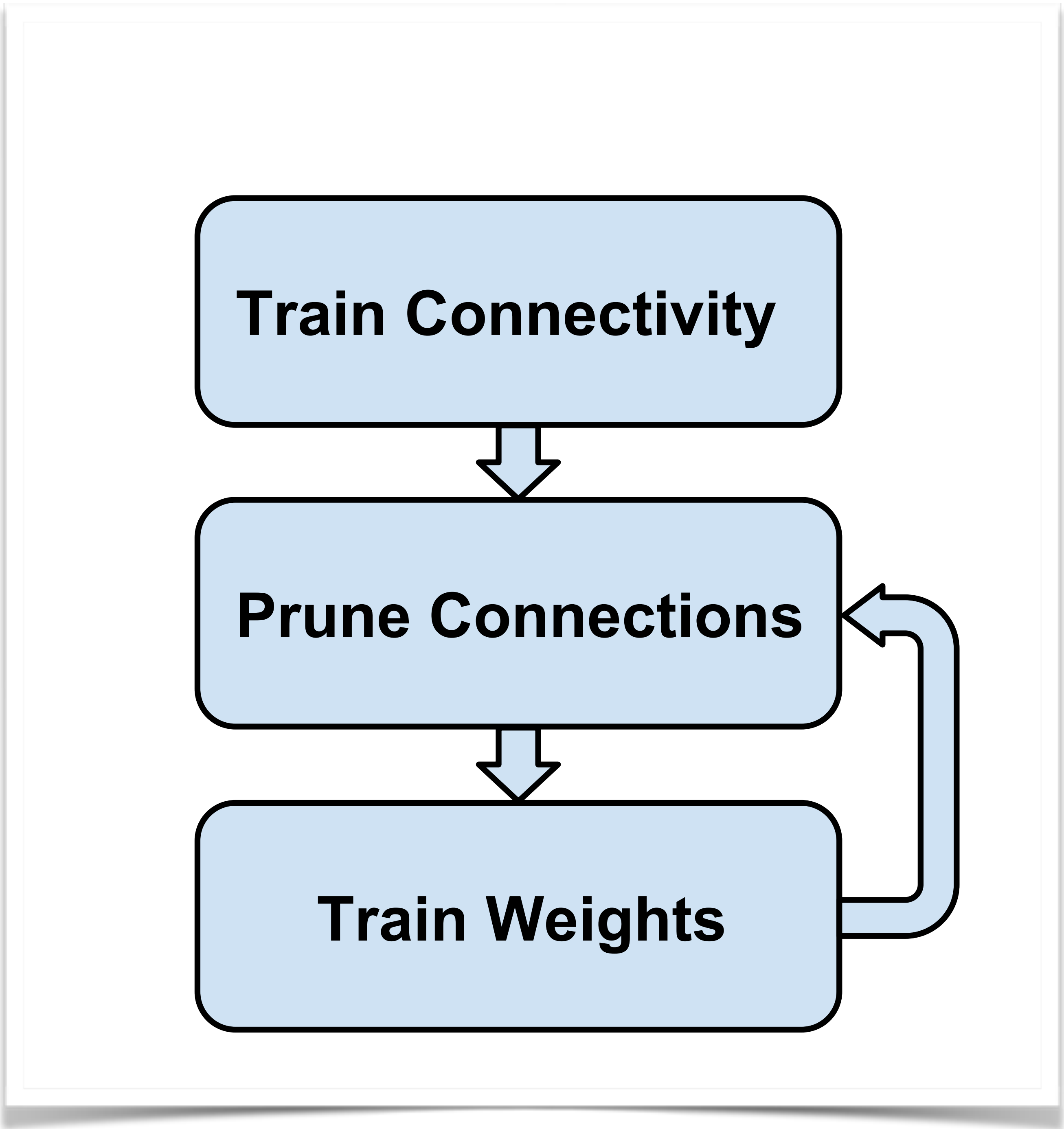
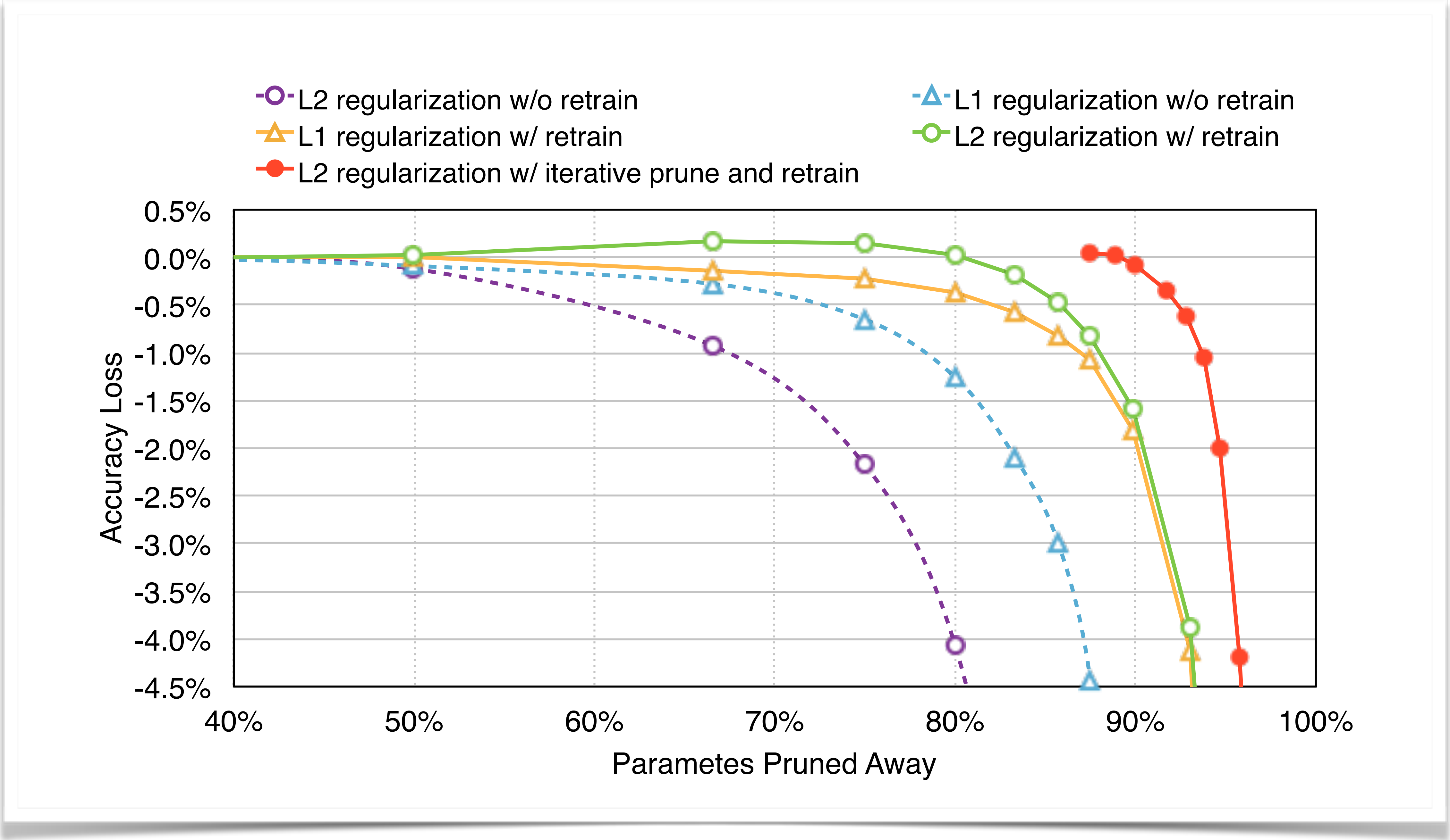


Image from Han *et al*. NeurIPs 2015

Network	Top-1 Error	Top-5 Error	Parameters	Compression Rate
LeNet-300-100 Ref	1.64%	-	267K	
LeNet-300-100 Pruned	1.59%	-	**22K**	**12**$\times$
LeNet-5 Ref	0.80%	-	431K	
LeNet-5 Pruned	0.77%	-	**36K**	**12**$\times$
AlexNet Ref	42.78%	19.73%	61M	
AlexNet Pruned	42.77%	19.67%	**6.7M**	**9**$\times$
VGG-16 Ref	31.50%	11.32%	138M	
VGG-16 Pruned	31.34%	10.88%	**10.3M**	**13**$\times$



L2 regularization w/o retrain
L1 regularization w/o retrain
L1 regularization w/ retrain
L2 regularization w/ retrain
L2 regularization w/ iterative prune and retrain
0.5%
0.0%
-0.5%
-1.0%
-1.5%
-2.0%
-2.5%
-3.0%
-3.5%
-4.0%
-4.5%
Accuracy Loss
40%
50%
60%
70%
80%
90%
100%
Parametes Pruned Away

The lottery-ticket hypothesis

The Lottery Ticket Hypothesis. *A randomly-initialized, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.*

THE LOTTERY TICKET HYPOTHESIS:
FINDING SPARSE, TRAINABLE NEURAL NETWORKS

Jonathan Frankle
MIT CSAIL
jfrankle@csail.mit.edu

Michael Carbin
MIT CSAIL
mcarbin@csail.mit.edu

Finding winning tickets

1. Randomly initialize a neural network $f(x; \theta_0)$ (where $\theta_0 \sim \mathcal{D}_\theta$).
2. Train the network for j iterations, arriving at parameters θ_j.
3. Prune $p\%$ of the parameters in θ_j, creating a mask m.
4. Reset the remaining parameters to their values in θ_0, creating the winning ticket $f(x; m \odot \theta_0)$.

THE LOTTERY TICKET HYPOTHESIS:
FINDING SPARSE, TRAINABLE NEURAL NETWORKS

Jonathan Frankle
MIT CSAIL
jfrankle@csail.mit.edu

Michael Carbin
MIT CSAIL
mcarbin@csail.mit.edu

Results

- Consistently find winning tickets (less than 10-20% size of original models)
- These actually often yield **higher** test accuracy!
- Very much an ongoing research topic…

THE LOTTERY TICKET HYPOTHESIS: FINDING SPARSE, TRAINABLE NEURAL NETWORKS

Jonathan Frankle
MIT CSAIL
jfrankle@csail.mit.edu

Michael Carbin
MIT CSAIL
mcarbin@csail.mit.edu