FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

Knowledge Distillation

Final Presentation

Vrinda Gupta
Seminar Advanced Deep Learning, Friedrich-Alexander-Universität Erlangen-Nürnberg
February 1, 2023

2014

Distilling the Knowledge in a Neural Network

NEURAL INFORMATION PROCESSING SYSTEMS

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

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

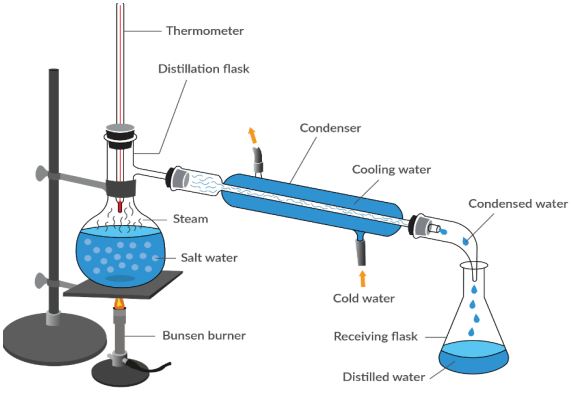
Jeff Dean
Google Inc.
Mountain View
jeff@google.com

Distilling the knowledge in a neural network
G Hinton, O Vinyals, J Dean - arXiv preprint arXiv:1503.02531, 2015 - arxiv.org
A very simple way to improve the performance of almost any machine learning algorithm is to train many different models on the same data and then to average their predictions. Unfortunately, making predictions using a whole ensemble of models is cumbersome and …
☆ Cited by 3740 Related articles »

compress the knowledge in an ensemble into a single model which is much easier to deploy and we develop this approach further using a different compression technique. We achieve some surprising results on MNIST and we show that we can significantly improve the acoustic model of a heavily used commercial system by distilling the knowledge in an ensemble of models into a single model. We also introduce a new type of ensemble composed of one or more full models and many specialist models which learn to distinguish fine-grained classes that the full models confuse. Unlike a mixture of experts, these specialist models can be trained

Introduction

Knowledge distillation: learning a small model from a larger model. [1]

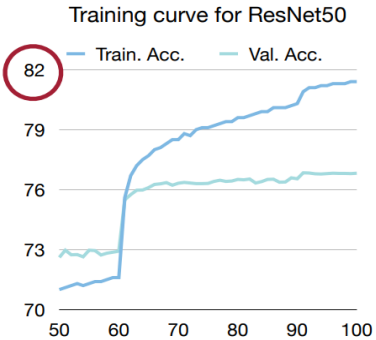


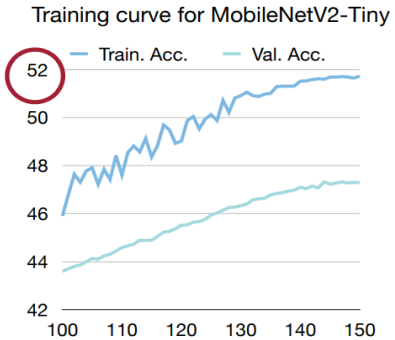
Source: https://www.google.com/imghp?hl=EN

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Challenges

1. **Tiny models are hard to train**

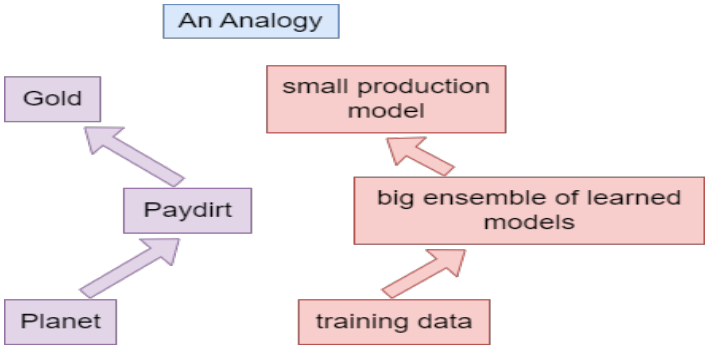




Source: https://www.google.com/imghp?hl=EN

Challenges

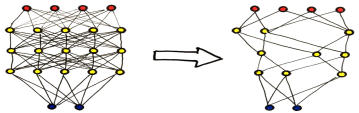
2. **Training and inference have different requirement**



- During training: Dropout, ensemble, increase depth, larger dataset
- During inference: Accuracy, speed

Challenges

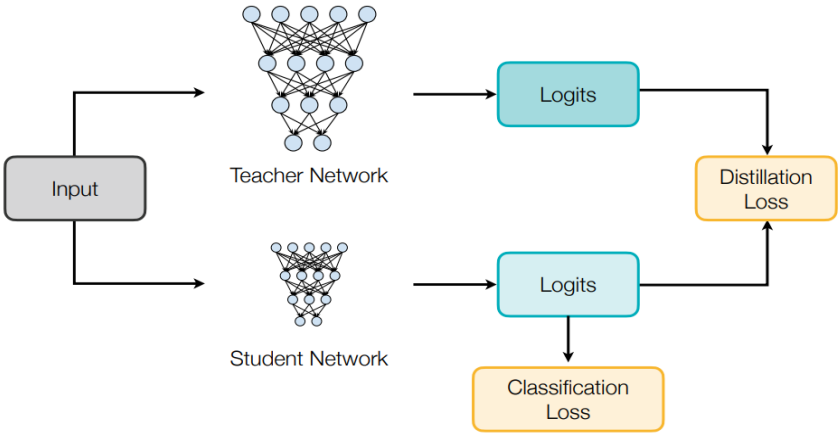
3. Limited hardware resources



	Cloud AI	Tiny AI
Computation (fp32)	19.5 TFLOPS	MFLOPs
Memory	80GB	256kB
Neural Network	ResNet ViT-Large …	MCUNet MobileNetV2-Tiny …

Source: https://www.google.com/imghp?hl=EN

Illustration of knowledge distillation



Loss = Classification loss + Distillation loss

Source: https://www.google.com/imghp?hl=EN

It all begins with Maths!

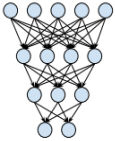
Classification loss

$$q_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{1}$$

where the z_i values are the elements of the input vector

- each probability in the result is in the range 0...1
- the probabilities sum to 1

Classification loss



Teacher Network

	Logits	Probabilities
Cat	5	0.982
Dog	1	0.017

$$\frac{\exp(5)}{\exp(5)+\exp(1)}$$

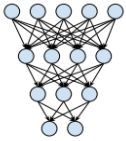
$$\frac{\exp(1)}{\exp(5)+\exp(1)}$$

Student Network

	Logits	Probabilities
Cat	3	0.731
Dog	2	0.269

The student model is less confident

Classification loss



Teacher Network

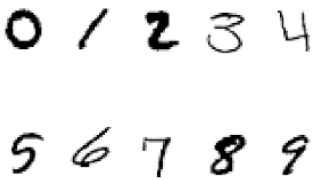
Student Network

	Logits	Probabilities
Cat	5	0.982
Dog	1	0.017

	Logits	Probabilities
Cat	3	0.731 ↑
Dog	2	0.269 ↓

Learning from DARK...! knowledge

Softmax hides relative similarities between classes. For ex: in MNIST, 2 is more similar to 3 than 7



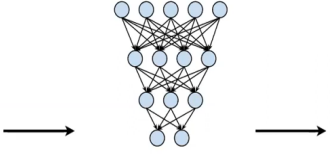
Modification in softmax function

$$q_i = \frac{\exp\left(z_i / T\right)}{\sum_j \exp\left(z_j \ / \ T\right)} \tag{2}$$

where value of T is hyperparameter

Concept of Temperature

- **low temperature model**: usually good at hard predictions
- **high temperature model**: considered to be having a dark knowledge



$$\frac{\exp(5/1)}{\exp(5/1)+\exp(1/1)}$$

	Logits	Probabilities (T=1)	Probabilities (T=10)
Cat	5	0.982	0.599
Dog	1	0.017	0.401

$$\frac{\exp(5/10)}{\exp(5/10)+\exp(1/10)}$$

Overall loss

Loss = Classification loss + Distillation loss

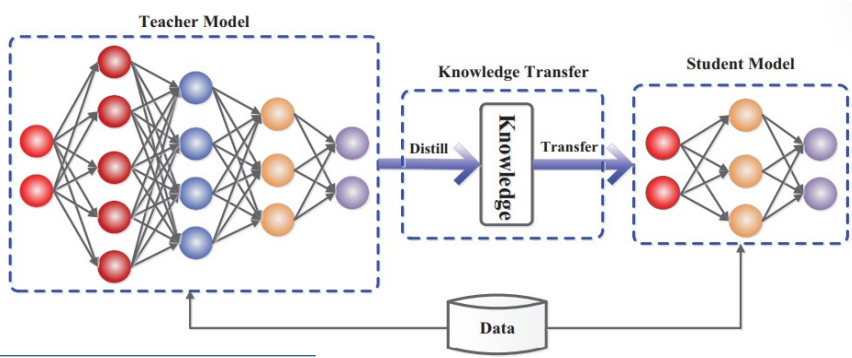
$$Loss = (1-\alpha)L_{CE}\left(\sigma\left(Z_s\right), \hat{y}\right) + 2\alpha T^2 L_{CE}\left(\sigma\left(\frac{Z_t}{T}\right), \sigma\left(\frac{Z_s}{T}\right)\right) \tag{3}$$

where
$L_{c\varepsilon}()$: Cross entropy loss
$\sigma 0$: Softmax
Z_s : Output logits of student network
Z_t : Output logits of teacher network
$\hat{y}$: Ground truth(one-hot)
α : Balancing parameter
T: Temperature

Model Architecture

Three key components: [2]

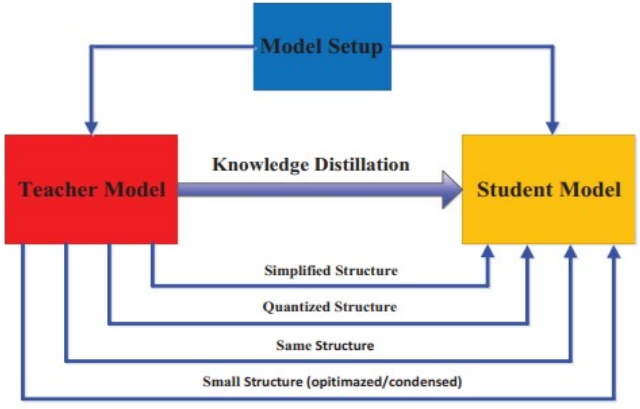
1. **teacher-student architecture**
2. distillation algorithm
3. knowledge



[2] Gou, Jianping et. al Knowledge Distillation: A Survey, 2021

Teacher-Student Architecture

- the quality of distillation depends on design the teacher and student networks
- "Student finds a right teacher"

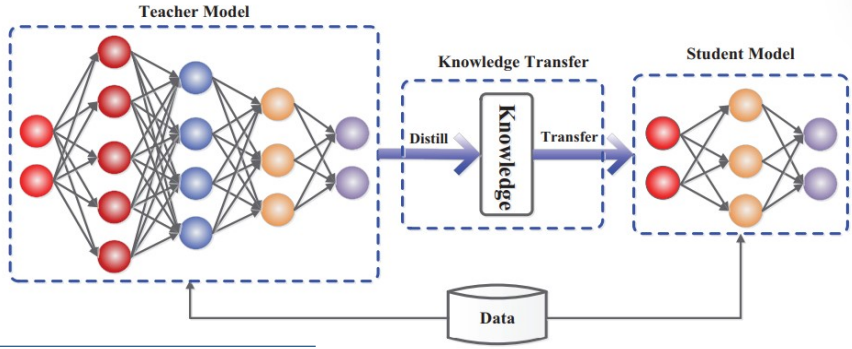


Source: Gou, Jianping, et al. "Knowledge Distillation: A Survey, 2021

Model Architecture

Three key components: [3]

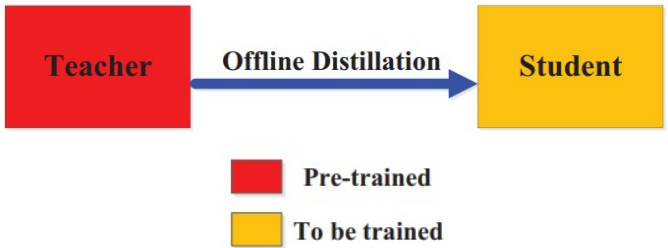
1. knowledge
2. **distillation schemes**
3. teacher-student architecture



[3] Gou, Jianping et. al Knowledge Distillation: A Survey, 2021

Distillation Schemes

Offline Distillation

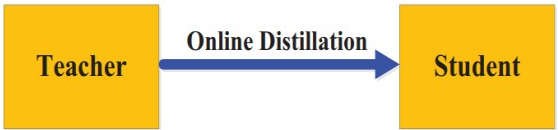


- What are disadvantage of fixed large teachers?
  1. training overhead
  2. large space to save logits

Distillation Schemes

Online Distillation [4]

- the teacher model and the student model are updated simultaneously
- both trainable

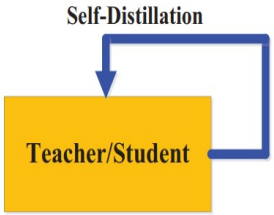


[4] Deep Mutual Learning [Zhang et al., CVPR 2018]

Distillation Schemes

Self Distillation

- the same networks are used for the teacher and the student models



Self-Distillation

Born-Again Networks(BANs) outperform their teacher's

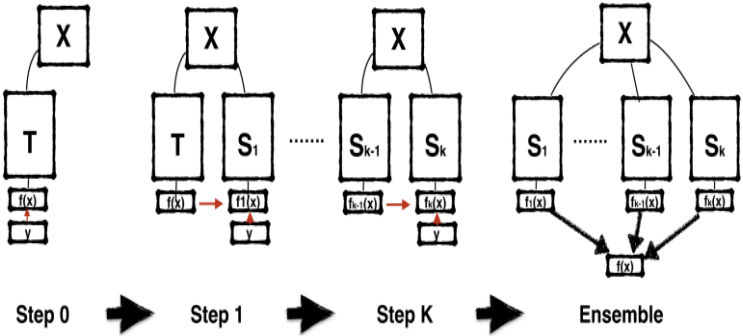
Born-Again Neural Networks

Tommaso Furlanello [1] **Zachary C. Lipton** [2 3] **Michael Tschannen** [4] **Laurent Itti** [1] **Anima Anandkumar** [5 3]

Abstract

Knowledge Distillation (KD) consists of transferring "knowledge" from one machine learning model (the *teacher*) to another (the *student*). Commonly, the teacher is a high-capacity model with formidable performance, while the student is more compact. By transferring knowledge, one hopes to benefit from the student's compactness, without sacrificing too much performance. We study various forms of knowledge distillation on the ImageNet dataset and show that, surprisingly, it is possible to significantly improve the performance of a student over the teacher when the student has the same architecture as the teacher. We call these models born-again networks (BANs), and we show that they can compete with and even outperform the original teacher models. Interestingly, given such a powerful ensemble, one can often find a simpler model — no more complex than one of the ensemble's constituents — that mimics the ensemble and achieves its performance. Previously, in *Born-Again Trees* Breiman & Shang (1996) pioneered this idea, learning single trees that match the performance of multiple-tree predictors. These born-again trees approximate the ensemble decision but offer some desired properties of individual

Born-Again Neural Networks[5]

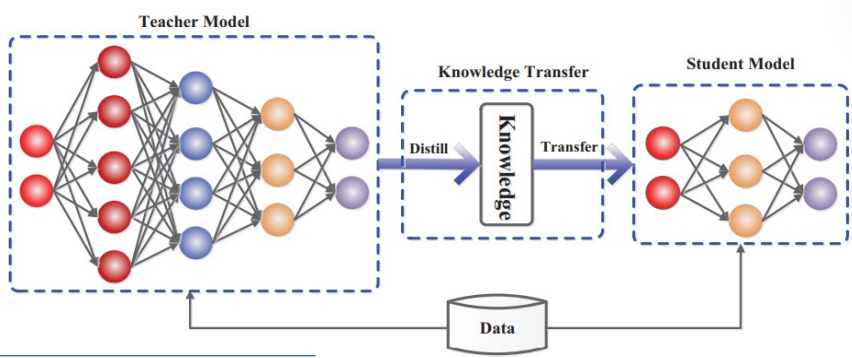


[5] Born-Again Neural Networks [Furlanello et al., ICML 2018]

Model Architecture

Three key components: [6]

1. teacher-student architecture
2. distillation schemes
3. **knowledge**

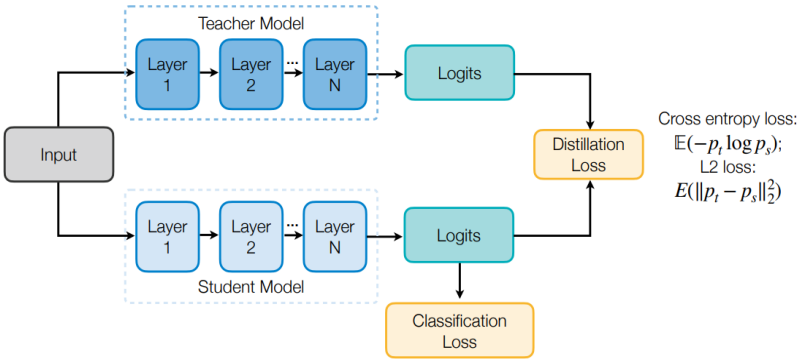


[6] Gou, Jianping et. al Knowledge Distillation: A Survey, 2021

Knowledge - What to match ?

Output Logits

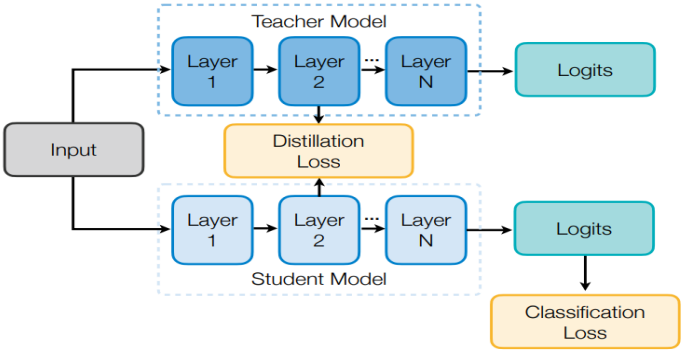
- Response based
- E.g: soft targets (by Hinton)



Knowledge - What to match ?

Feature based

- intermediate weights, features
- gradients



What to match? intermediate weights

2015

FitNets: Hints for Thin Deep Nets

ICLR

Adriana Romero[1], **Nicolas Ballas**[2], **Samira Ebrahimi Kahou**[3], **Antoine Chassang**[2], **Carlo Gatta**[4] & **Yoshua Bengio**[2]†
[1]Universitat de Barcelona, Barcelona, Spain.
[2]Université de Montréal, Montréal, Québec, Canada. †CIFAR Senior Fellow.

Fitnets: Hints for thin deep nets

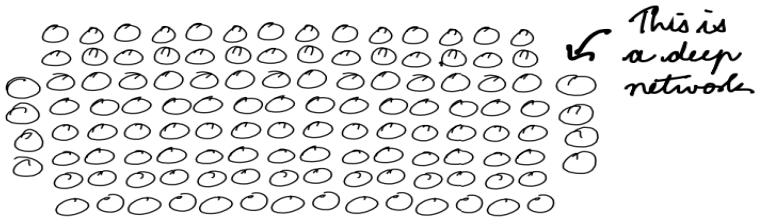
A Romero, N Ballas, SE Kahou, A Chassang… - arXiv preprint arXiv …, 2014 - arxiv.org

While depth tends to improve network performances, it also makes gradient-based training more difficult since deeper networks tend to be more non-linear. The recently proposed knowledge distillation approach is aimed at obtaining small and fast-to-execute models, and …

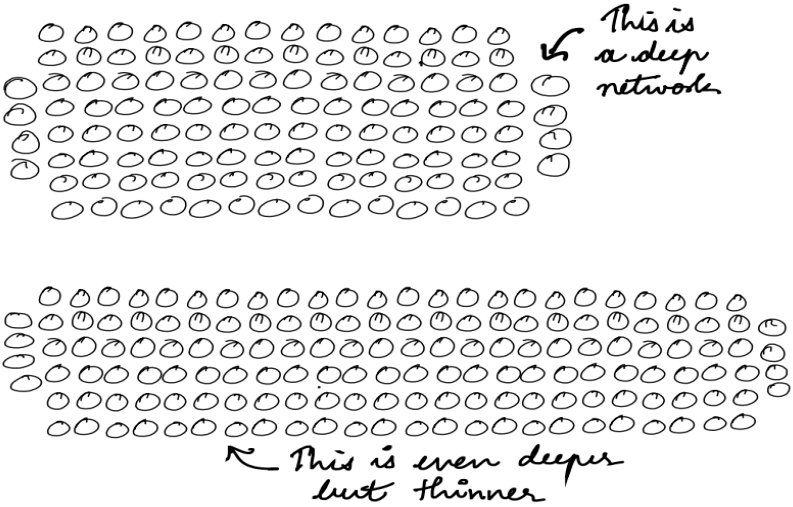
☆ 99 Cited by 1035 Related articles All 13 versions »

cently proposed knowledge distillation approach is aimed at obtaining small and fast-to-execute models, and it has shown that a student network could imitate the soft output of a larger teacher network or ensemble of networks. In this paper, we extend this idea to allow the training of a student that is deeper and thinner than the teacher, using not only the outputs but also the intermediate representations learned by the teacher as hints to improve the training process and final performance of the student. Because the student intermediate hidden layer will

FitNets

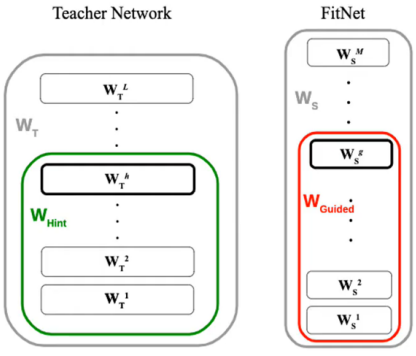


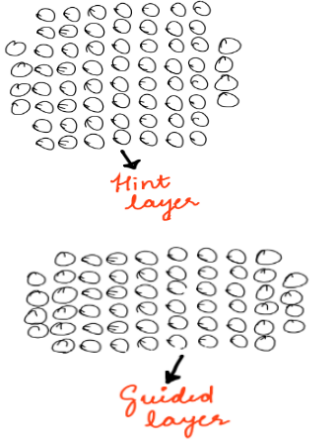
FitNets



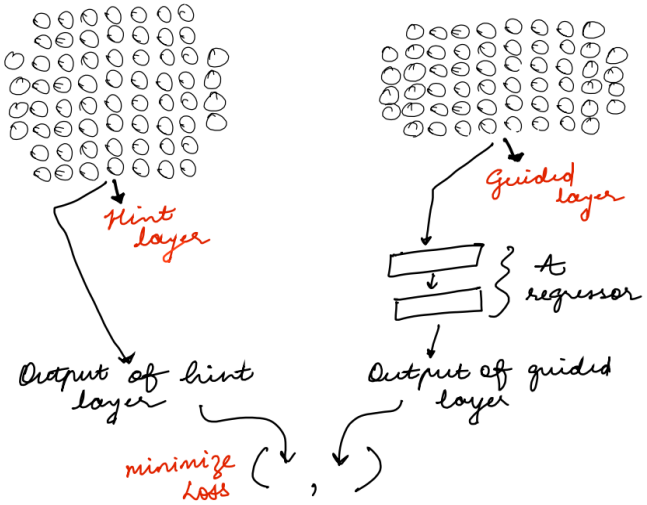
Fitnets

Teacher-Student Architecture



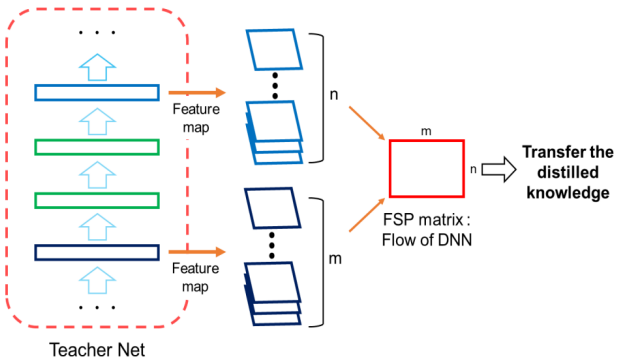


FitNets


Hint layer
Guided layer
A regressor
Output of hint layer
Output of guided layer
minimize Loss

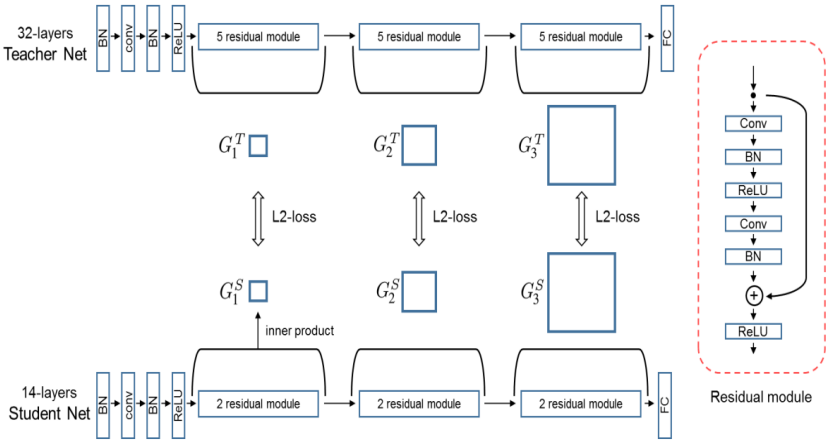
What to match? Relational Information

- Explores the relationships between different layers or data samples
  FSP Matrix Relation between different layers : [7]



[7] Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning [Yim et al., CVPR 2017]

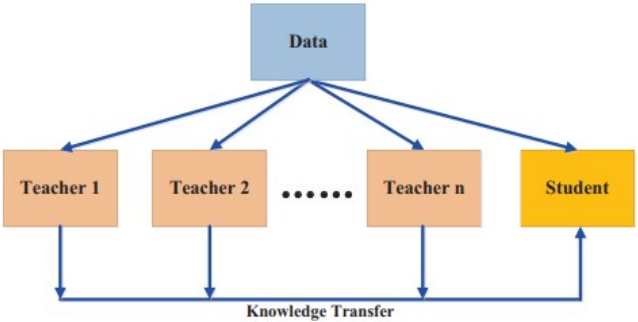
What to match? Relational Information



Distillation Algorithm

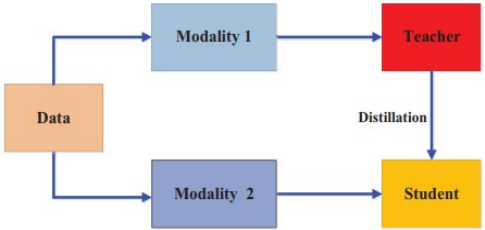
Various distillation methods for knowledge transfer

1. **Multi-Teacher Distillation**
   transfer knowledge from multiple teachers, e.g., use average response of all teachers



Source: Gou, Jianping et. al Knowledge Distillation: A Survey, 2021

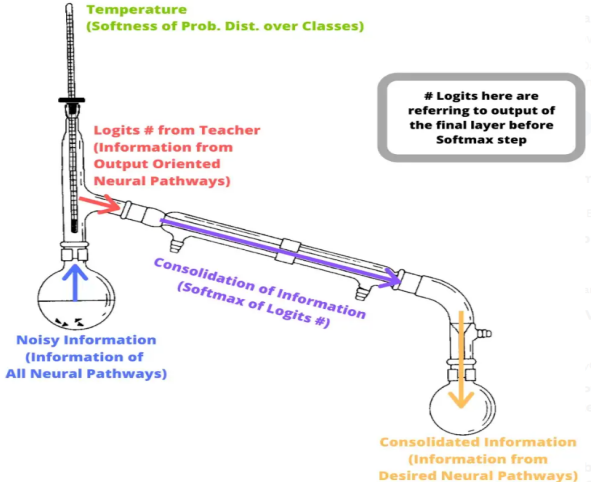
2. Cross-Modal Distillation



Modality for Teacher	Modality for Student
RGB images	Depth images
Vision	Sound
Textual Modality	Visual Modality
Temporal data	spatial data

Discussion

Overall picture



Source: https://towardsdatascience.com/distillation-of-knowledge-in-neural-networks-cc02f79698b6

My Experiment

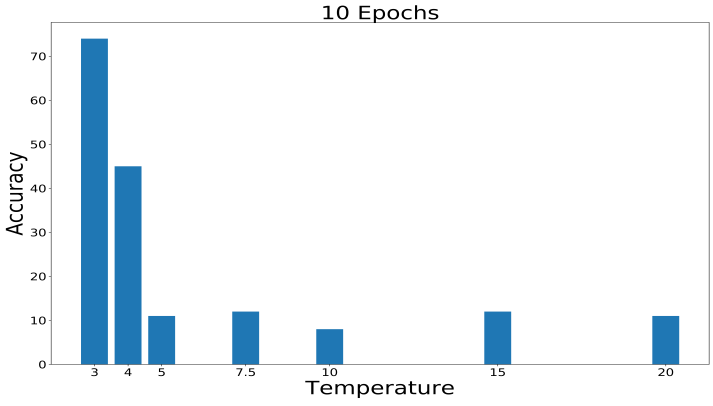
- Offline distillation
- Response based

My Experiment

Teacher Model	97%	31.8secs	BigNet (conv1): Conv2d $(1, 32,$ kernel_size =(5,5), stride =(1,1)) (conv2): Conv2d $(32, 32,$ kernel_size =(5,5), stride=(1,1) (conv3): Conv2d $(32, 64,$ kernel_size =(5,5), stride =(1,1)) (fc1): Linear (in_features =576, out_features =256, bias = True) (fc2): Linear(in_features=256, out_features=10, bias=True)
Smaller Model	11%	18secs	SmallNet (conv1): Conv2d $(1, 16,$ kernel_size =(5,5), stride =(1,1)) (conv2): Conv2d $(16, 32,$ kernel_size =(5,5), stride =(1,1) (fc1): Linear (in_features=3200, out_features =10, bias = True)
Student Model	45%	21secs	DistilledNet SmallNet Regularized by soft target [T=4]

Ablation Study

Temperature vs Accuracy



Advantages

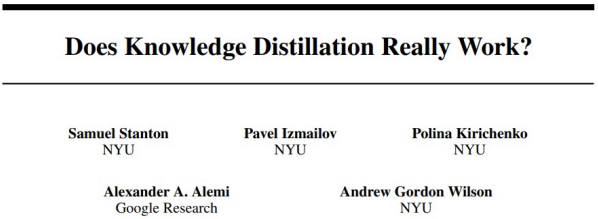
Using distillation, one could reduce the size of models like **BERT** by **87%** and still retain **96%** of its performance.

Basically, distillation enables one to:

- get state-of-the-art accuracy
- with lighter model
- within less time
- could run models on CPUs
- can be used even when there are fewer training data available for the student model.

Further comments

Knowledge distillation seems a practical and effective solution, but just how well does it really work?



"In short: **Yes**, in the sense that it often improves student generalization, though there is room for further improvement. **No**, in that knowledge distillation often fails to live up to its name, transferring very limited knowledge from teacher to student." [8]

[8] Does Knowledge Distillation Really Work? [Wilson et al., Advances in Neural Information Processing Systems 2021]

Efficacy of Knowledge Distillation

1. Imagenet Dataset : Paying more attention to attention

4.2.2 IMAGENET

To showcase activation-based attention transfer on ImageNet we took ResNet-18 as a student, and ResNet-34 as a teacher, and tried to improve ResNet-18 accuracy. We added only two losses in the 2 last groups of residual blocks and used squared sum attention F^2_{sum}. We also did not have time to tune any hyperparameters and kept them from finetuning experiments. Nevertheless, ResNet-18 with attention transfer achieved 1.1% top-1 and 0.8% top-5 better validation accuracy (Table. 5 and Fig. 7(a), Appendix), we plan to update the paper with losses on all 4 groups of residual blocks.

We were not able to achieve positive results with KD on ImageNet. With ResNet-18-ResNet-34 student-teacher pair it actually hurts convergence with the same hyperparameters as on CIFAR. As it was reported that KD struggles to work if teacher and student have different architecture/depth (we observe the same on CIFAR), so we tried using the same architecture and depth for attention transfer. On CIFAR both AT and KD work well in this case and improve convergence and final accuracy, on ImageNet though KD converges significantly slower (we did not train until the end due to lack of computational resources). We also could not find applications of FitNets, KD or similar methods on ImageNet in the literature. Given that, we can assume that proposed activation-based AT is the first knowledge transfer method to be successfully applied on ImageNet.

Efficacy of Knowledge Distillation

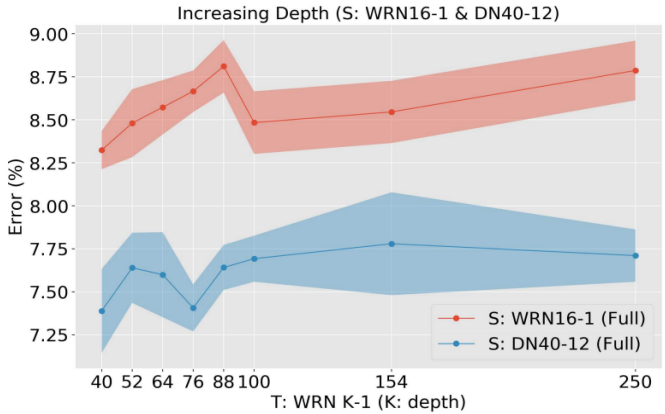
1. Imagenet Dataset : Paying more attention to attention [9]

Teacher	Teacher Error (%)	Student Error (%)
-	-	30.24
ResNet18	30.24	30.57
ResNet34	26.70	30.79
ResNet50	23.85	30.95

[9] Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer [Komodakis et al., CVPR 2016]

Efficacy of Knowledge Distillation

2. Bigger Model are better teachers?-No!



Reason: a mismatch between student and teacher capacities

Efficacy of Knowledge Distillation

3. Repeated knowledge distillation

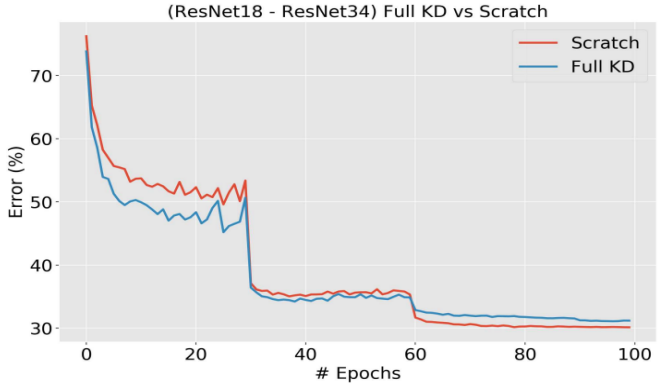
Training Procedure	Large Error (%)	Medium Error (%)	Small Error (%)
Large→Med.→ Small	4.41	4.80	8.04 (7.99 ± 0.24)
Med. → Small	-	5.34	**7.614** (**7.68 ± 0.26**)
Large → Small	4.41	-	7.98 (8.03 ± 0.14)

Table 6. Using sequential knowledge distillation to distill from a large model (WRN16-8) to a medium model (WRN16-3), and from the latter to a small model (WRN16-1) does not help. The optimal approach still is to distill directly from the medium model to the small model, even though the teacher in this case has lower accuracy.

Reason: the network architecture heavily determines the success of sequential knowledge distillation

Efficacy of Knowledge Distillation

4. Distillation adversely affects training



Reason: student may not have enough capacity to minimize both the training loss and KD loss
Solution: “Early-stopped” knowledge distillation (“ESKD”)

Efficacy of Knowledge Distillation

5. Deciding other factors
    - Configuration
    - Value of alpha
    - Temperature

Further Readings

- Applied to several machine learning and deep learning use cases
  1. Computer Vision : image classification, face recognition, object detection
  2. NLP: text generation, document retrieval
  3. Speech: language identification, speech recognition
- Distillation Metric: A tradeoff [10]

$$DS = \alpha * \left(\frac{\text{student}_s}{\text{teacher}_s} \right) + (1 - \alpha) * \left(1 - \frac{\text{student}_a}{\text{teacher}_a} \right) \tag{4}$$

where
DS - distillation score,
student_s and student_a - student size and accuracy,
teacher_s and teacher_a - teacher size and accuracy,
α - hyperparameter

[10] Knowledge Distillation in Deep Learning and its Applications[Alkhulaifi et al.,PeerJ Comput.Sci. 7:e474 (2021)]

My Remarks!

1. Imagenet- a tricky problem
2. KD in semi-supervised, self-supervised, reinforcement learning, federation learning
3. Deciding on student-teacher architecture and distillation techniques
4. Distill large datasets into a small-scale synthetic dataset
5. Can we distill from neural networks to other machine learning models like decision tree?

Thank you for your attention.

Questions?