Knowledge Distillation

Bruno Melício,
brunomelicio.ai@gmail.com

Budapest, 17/03/2021

Content

- Introduction
  - Motivation
- Knowledge Distillation
  - Background
  - Standard approach
  - Variations
  - Knowledge Distillation with Teacher Assistants
- Conclusion

Introduction

Motivation

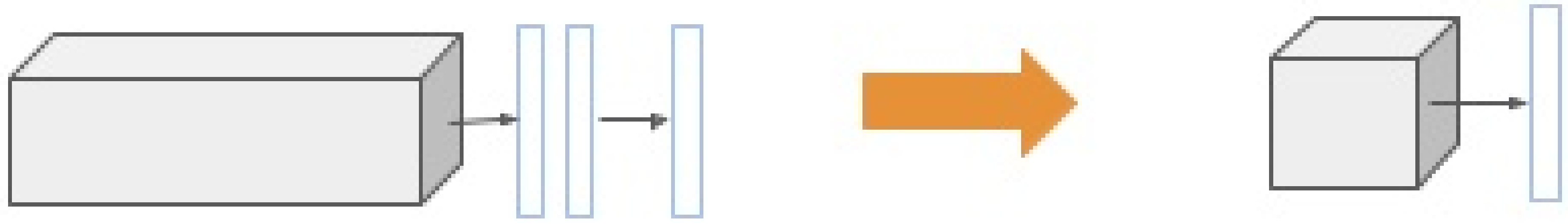
- Deep learning-based algorithms have achieved state of the art results on complex tasks that require Human Intelligence. However, these algorithms are trained on massive datasets resulting on huge models with a lot of parameters that restricts them to cloud computing for real time applications.
- **Thus, they cannot be deployed on edge devices.**

- A more suitable model for deployment would be a smaller model with less parameters but as accurate as a cumbersome[1] model.

1 Cumbersome - large or heavy and therefore difficult to carry or use; unwieldy.

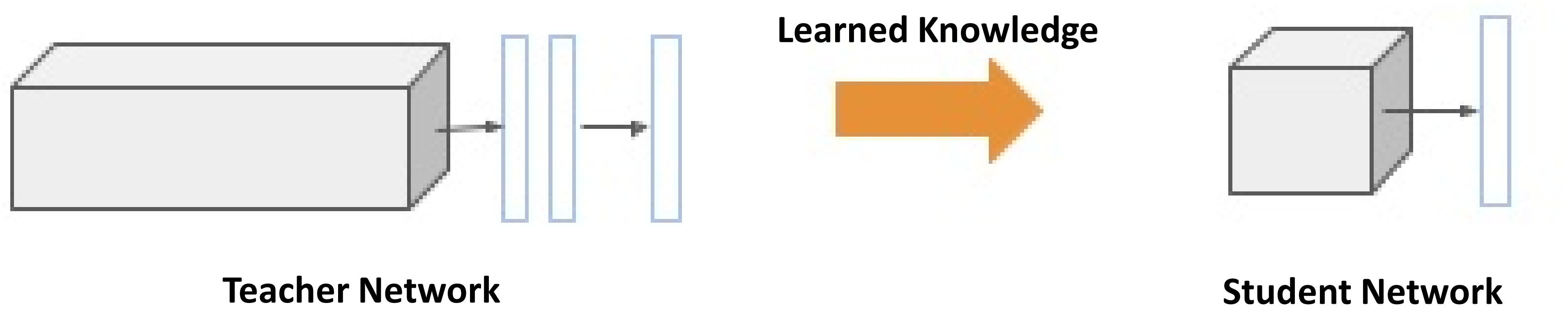
Knowledge Distillation

- Knowledge distillation is a compression technique that **transfers knowledge** from a **large model** to a **smaller model**.



Knowledge Distillation

- A big network with a lot of parameters, called **Teacher Network**, is trained on a huge dataset. Then, using a different kind of training, called **“distillation”**, the **learned knowledge is transferred** from the cumbersome model to a smaller network with fewer parameters, called **Student Network,** that is more suitable for deployment.

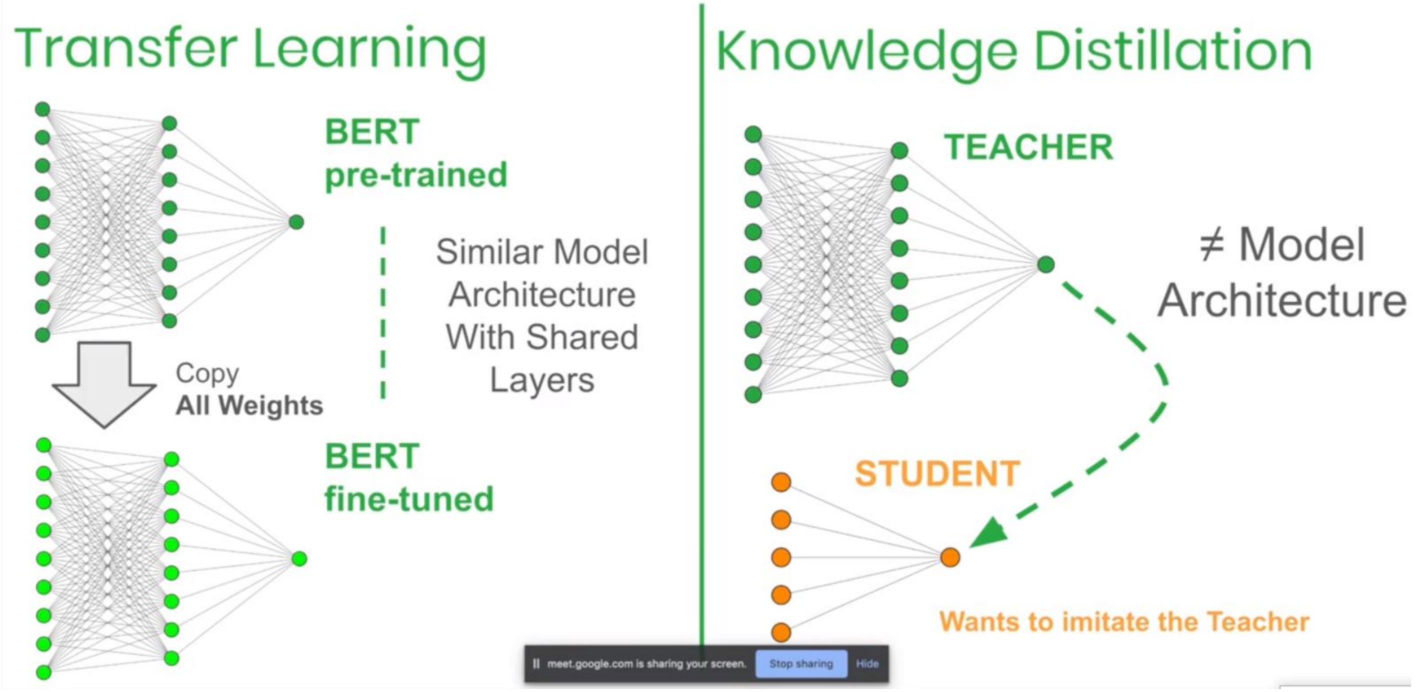


1 Cumbersome - large or heavy and therefore difficult to carry or use; unwieldy.

Knowledge Distillation

Background

- Knowledge Distillation is different than Transfer Learning
- Knowledge Distillation is a compression technique



Knowledge Distillation

Background

- Given a **dataset D = (X,Y)** we want to train a Neural Network to **learn a function $f_\theta(x)$** and **find the optimal parameters θ** such that the **loss $L(f_\theta(x), Y)$ is minimal**.
- $L = \sum_{(x,y)\in D} (y - f_\theta(x))^2$

x = 

Label / Hard Label

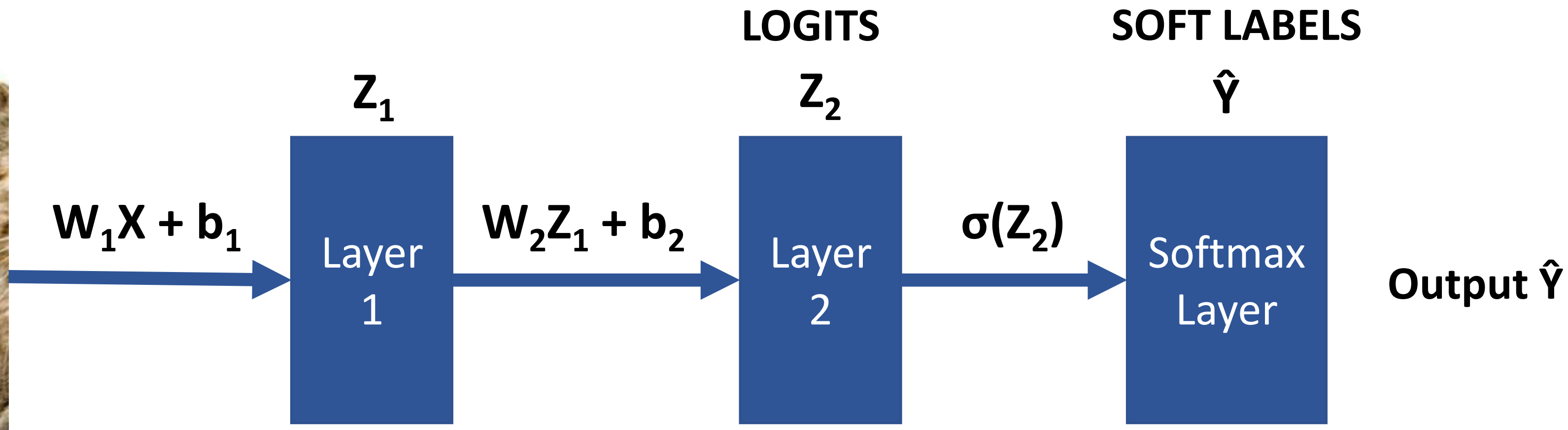
y = [dog, cat, fungus, plant]^T

y = [1, 0, 0, 0]^T

Knowledge Distillation

Background

Input X



$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

The network learned: $\hat{Y} = f(x) = \sigma(W_2\,(W_1X + b_1) + b_2)$

Knowledge Distillation

Standard approach (Hinton et al. 2015)[2]

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Input X

$W_1X + b_1$ → Z_1: Layer 1 → … → LOGITS Z_{10} → $\sigma(Z_{10})$ → SOFT LABELS $\hat{Y}$ | HARD LABELS Y

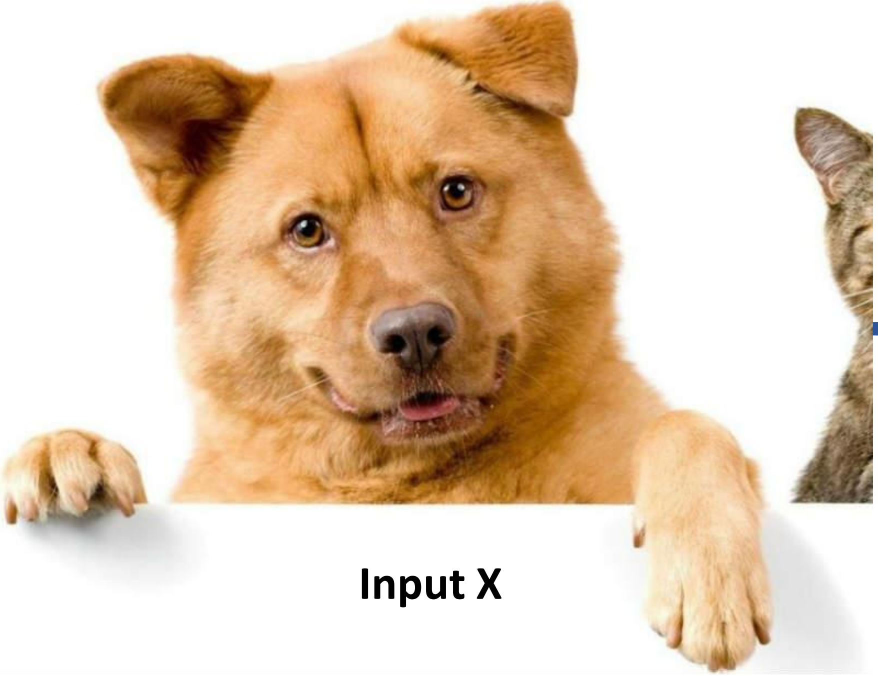
LOGITS Z_{10}	SOFT LABELS $\hat{Y}$	HARD LABELS Y
2.3	0.98	1
-2.3	0.01	0
-12	1e-06	0
-14	1e-07	0

Output $\hat{Y}$

2 Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the knowledge in a neural network. https://arxiv.org/abs/1503.02531

Knowledge Distillation

Standard approach (Hinton et al. 2015)[2]



Input X

Z_1

Layer 1

...

LOGITS

Z_{10}

2.3
-2.3
-12
-14

$\sigma(Z_{10} / T)$

T = 5

NEW SOFT LABELS

$\hat{Y}$

0.90
0.09
0.0001
1e-06

Output $\hat{Y}$

HARD LABELS

Y

1
0
0
0

Raising the temperature T $\sigma(\vec{z})_i = \dfrac{exp(z_i/T)}{\sum_j exp(z_j/T)}$

Knowledge Distillation

Standard approach (Hinton et al. 2015)[2] - Training the Student Network

- Given a **dataset D = (X,S)** where **S is the soft labels learned from the Teacher Network**, we want to train the Student Network **to learn a function $f_\theta(x)$** and **find the optimal parameters θ** which represent the learned knowledge from the Teacher such that the **loss $L(f_\theta(x), S)$ is minimal**.

Kullback Leibler divergence loss

$$L = KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

x =

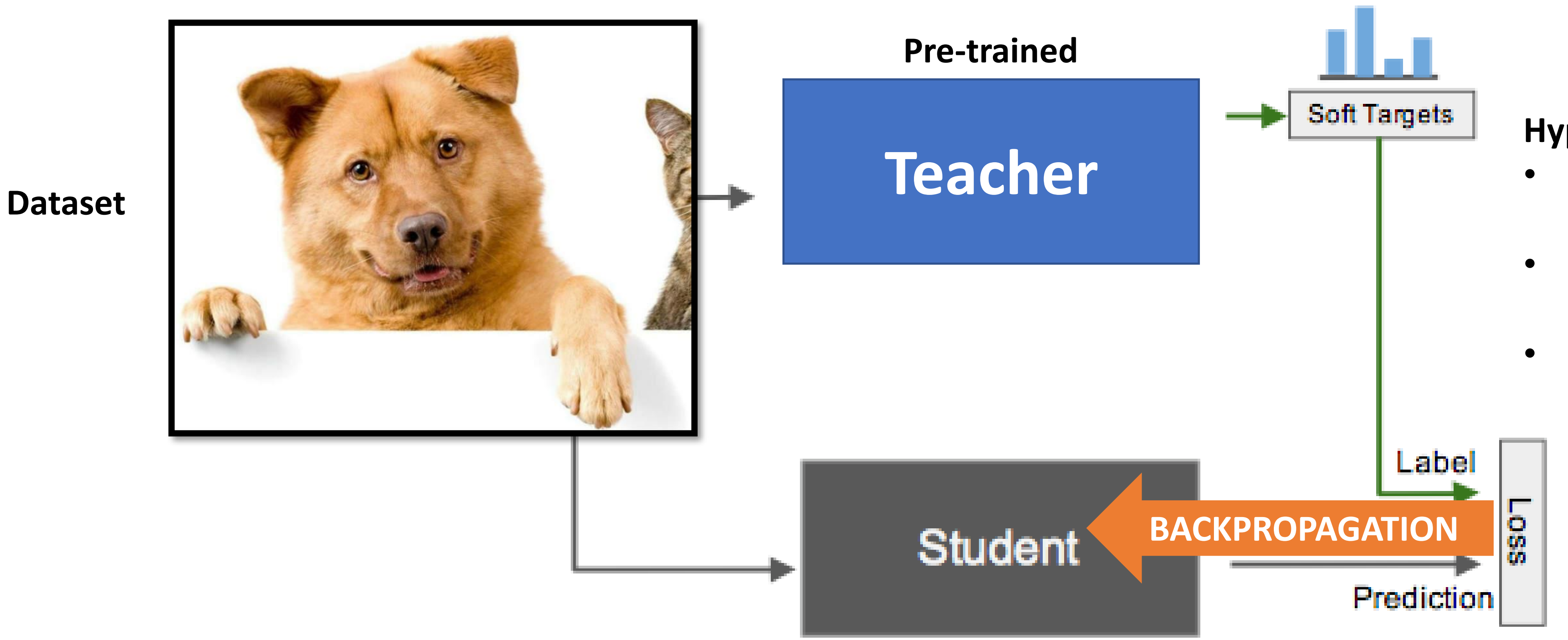
Soft Label

y = [dog, cat, fungus, plant]^T

y = [0.9, 0.09, 0.0001, 1e-06]^T

Knowledge Distillation

Standard approach (Hinton et al. 2015)[2] - Training the Student Network

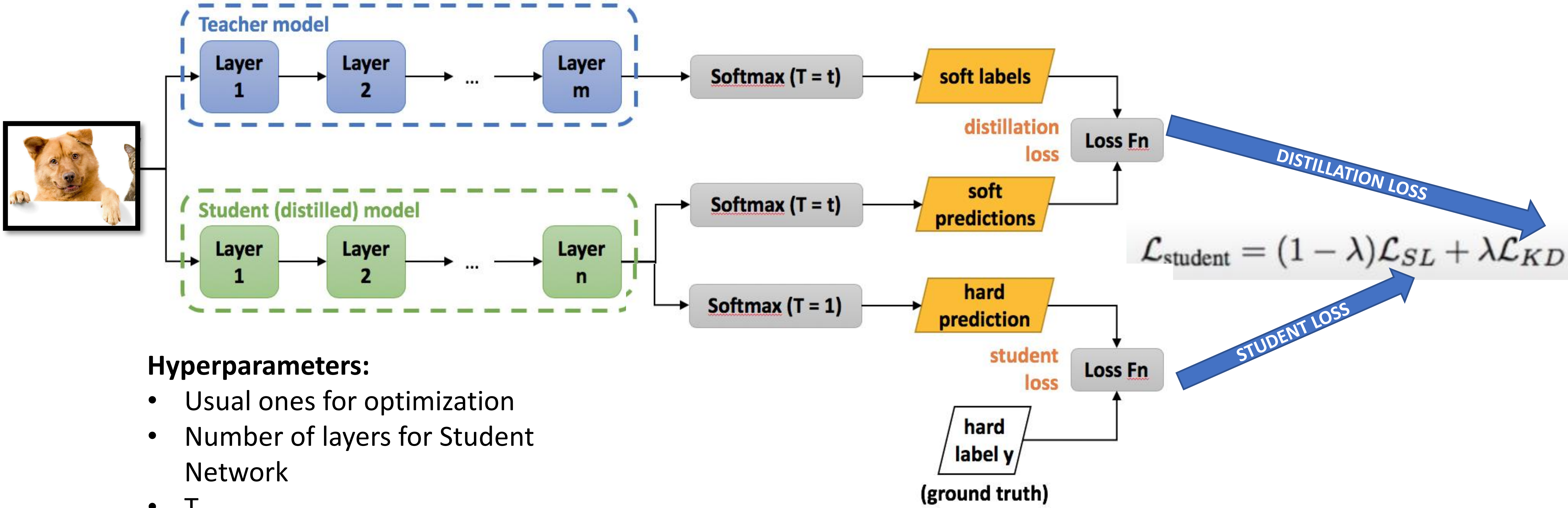


Hyperparameters:

- Usual ones for optimization
- Number of layers for Student Network
- T

Knowledge Distillation

Standard approach (Hinton et al. 2015)[2] - Training the Student Network



$$\mathcal{L}_{\text{student}} = (1-\lambda)\mathcal{L}_{SL} + \lambda\mathcal{L}_{KD}$$

Hyperparameters:

- Usual ones for optimization
- Number of layers for Student Network
- T
- λ

Knowledge Distillation

Results from Mirzadeh S.I. et al. 2019[3]

- Teacher Network: 10 Convolutional Layers
- Student Network: 2 Convolutional Layers

Table 1. Comparison on evaluation accuracy between training a student model with No Knowledge Distillation (**NOKD**) and a Baseline with Knowledge Distillation (**BLKD**)

Model	Dataset	NOKD	BLKD
CNN	CIFAR-10	70.16	72.57
	CIFAR-100	41.09	44.57

3 Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. 2019. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. CoRR, https://arxiv.org/abs/1902.03393

Knowledge Distillation

Variations (Gou J. et al. 2020)[4]

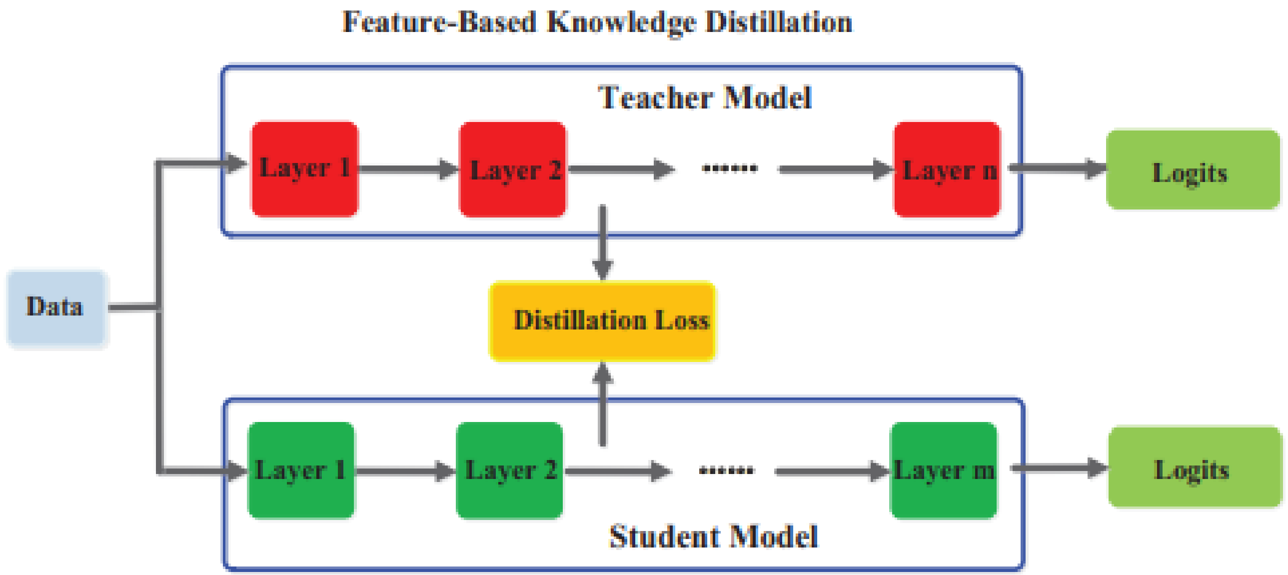


Fig. 7 The generic feature-based knowledge distillation.

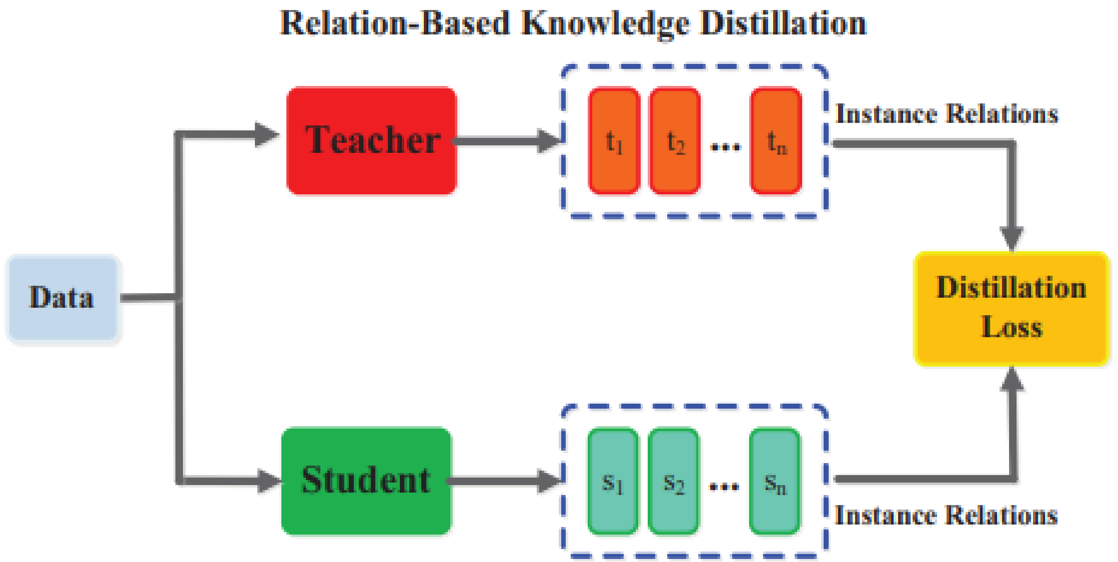


Fig. 8 The generic instance relation-based knowledge distillation.

4 Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey, 2020. https://arxiv.org/abs/2006.05525

Knowledge Distillation

Results from Cho J.H. et al. 2019[5]

- Teacher Network: ResNet18, ResNet34, ResNet50
- Student Network: ResNet18

Table 1. Top-1 error rate for various teachers for a ResNet18 student on ImageNet. The first row corresponds to training from scratch.

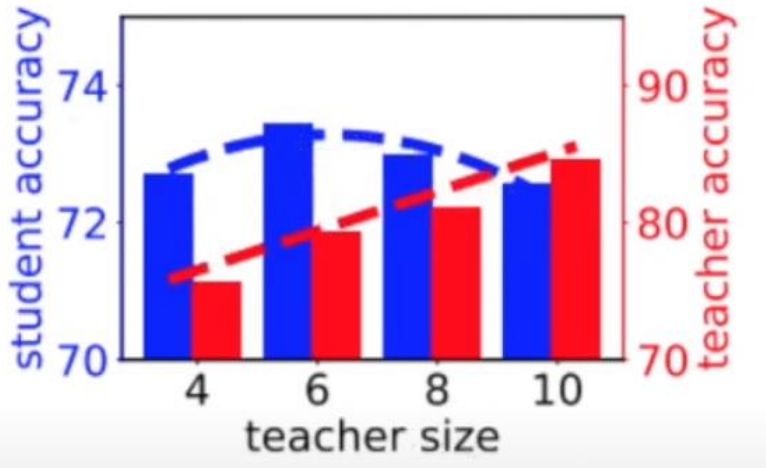
Teacher	Teacher Error (%)	Student Error (%)
-	-	30.24
ResNet18	30.24	30.57
ResNet34	26.70	30.79
ResNet50	23.85	30.95

5 https://openaccess.thecvf.com/content_ICCV_2019/papers/Cho_On_the_Efficacy_of_Knowledge_Distillation_ICCV_2019_paper.pdf

Knowledge Distillation

Results from Mirzadeh S.I. et al. 2019[3]

- Teacher Network: 4,6,8,10 Convolutional Layers
- Student Network: 2 Convolutional Layers

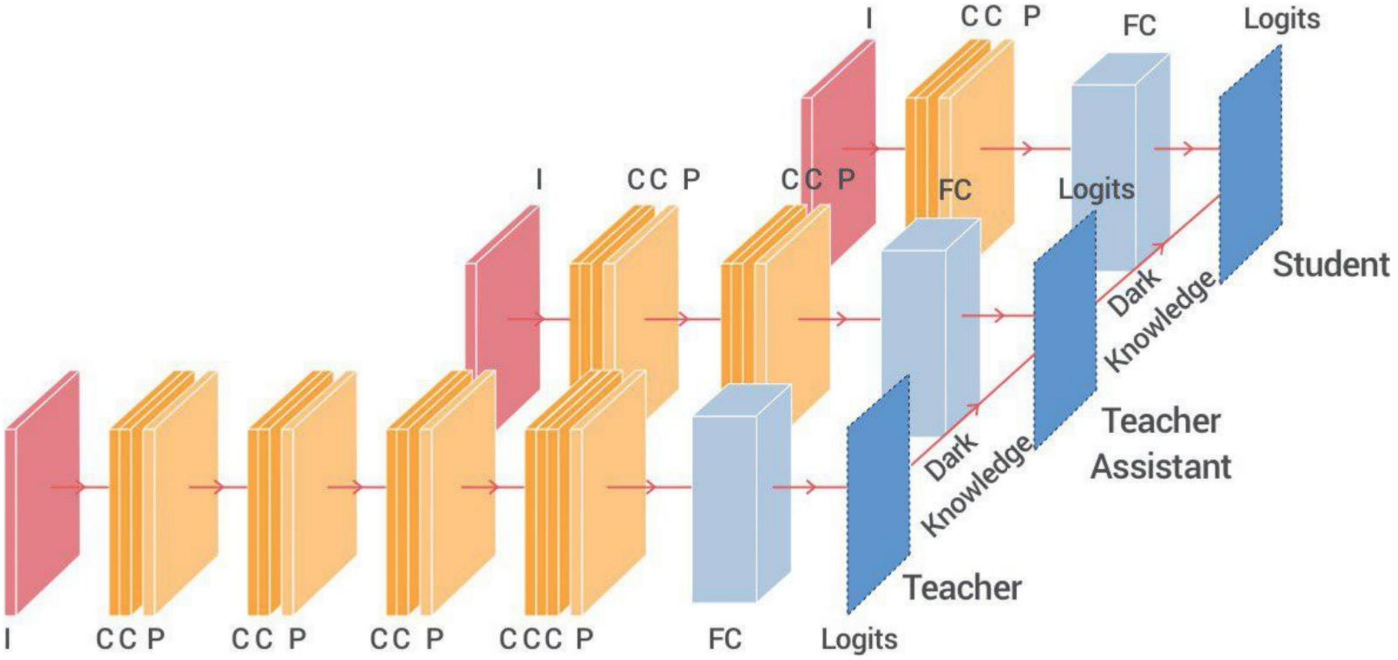


a) CIFAR-10

- **The teacher is becoming so complex that the student does not have the sufficient capacity or mechanics to mimic the teacher's behavior despite receiving hints.**
- **Teacher's certainty about data increases, thus making its soft targets less soft. This weakens the knowledge transfer which is done via matching the soft targets**

Knowledge Distillation

Improved Knowledge Distillation via Teacher Assistant (Mirzadeh S.I. et al. 2019)[3]



Teacher

↓

Teacher Assistant 1

↓

…

↓

Teacher Assistant n

↓

Student

Knowledge Distillation

Results from Mirzadeh S.I. et al. 2019[3]

- CNN layers: TN: 10 ; TA: 4; SN: 2
- ResNet layers: TN: 110; TA: 20; SN: 8

Table 1. Comparison on evaluation accuracy between training a student model with No Knowledge Distillation (**NOKD**) and a Baseline with Knowledge Distillation (**BLKD**) and Knowledge Distillation with Teacher Assistant (**TAKD**)

Model	Dataset	NOKD	BLKD	TAKD
CNN	CIFAR-10	70.16	72.57	**73.51**
	CIFAR-100	41.09	44.57	**44.92**
ResNet	CIFAR-10	88.52	88.65	**88.98**
	CIFAR-100	61.37	61.41	**61.82**

Conclusion

- Knowledge Distillation is a compression technique that transfers knowledge from a big Teacher Network to a small Student Network

- The transfer can be via the output soft labels or the hidden feature maps of the Teacher Network

- Adding intermediate Teacher Assistants can make the learning of the Student Network more effective

References

1. https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764
2. Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the knowledge in a neural network. https://arxiv.org/abs/1503.02531
3. Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. 2019. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. CoRR, https://arxiv.org/abs/1902.03393
4. Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey, 2020. https://arxiv.org/abs/2006.05525
5. https://openaccess.thecvf.com/content_ICCV_2019/papers/Cho_On_the_Efficacy_of_Knowledge_Distillation_ICCV_2019_paper.pdf
6. https://www.youtube.com/watch?v=lSjBc1wSJMI
7. https://www.youtube.com/watch?v=b3zf-JylUus&t=707s
8. KERAS IMPLEMENTATION: https://keras.io/examples/vision/knowledge_distillation/