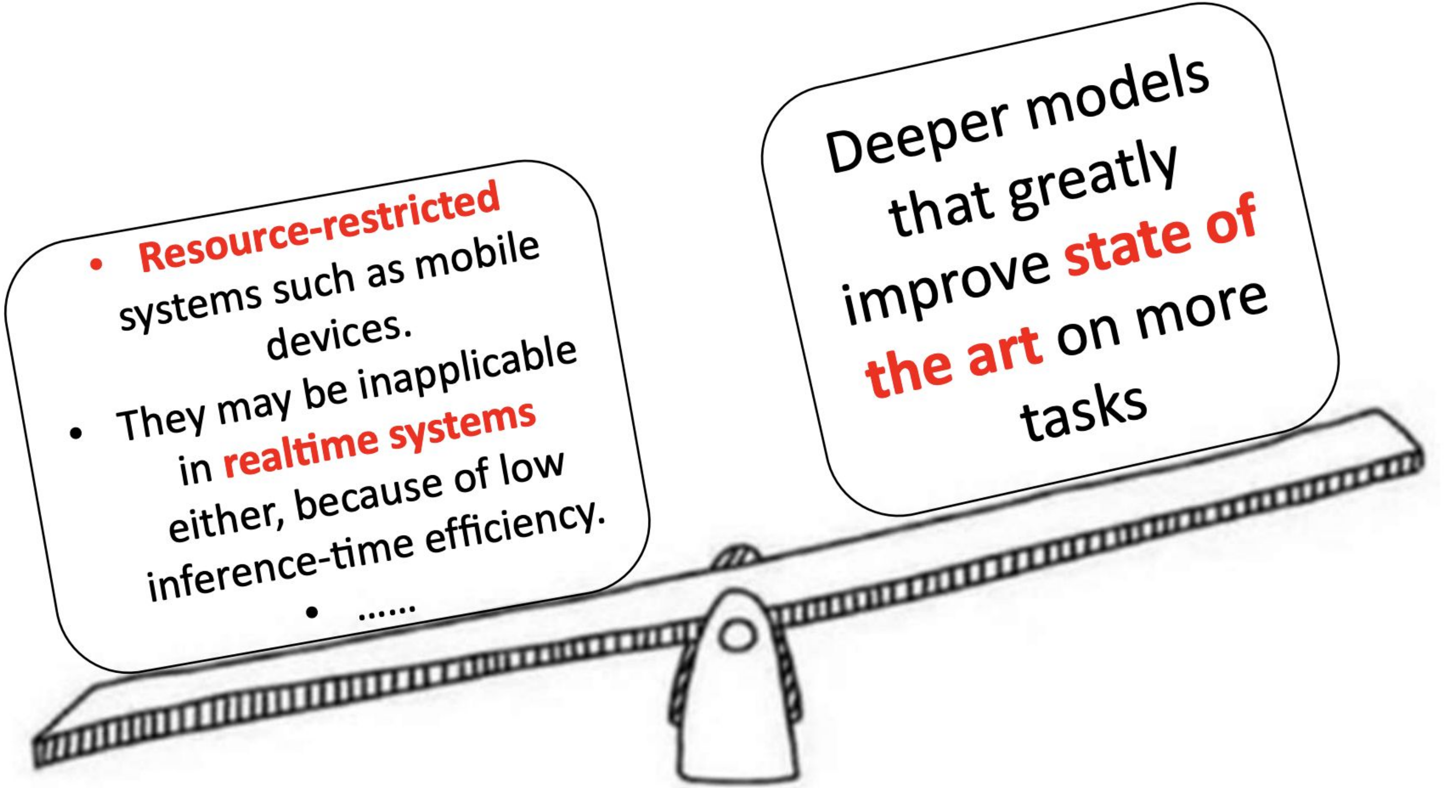
Neural Networks Speed-up and Compression

Lecture 5: Knowledge Distillation

Outline

- Motivation
- Knowledge Distillation concept
- Knowledge Distillation examples
    - Using Soft targets
    - Using features (FitNets)
    - Using attention
- Other details of KD

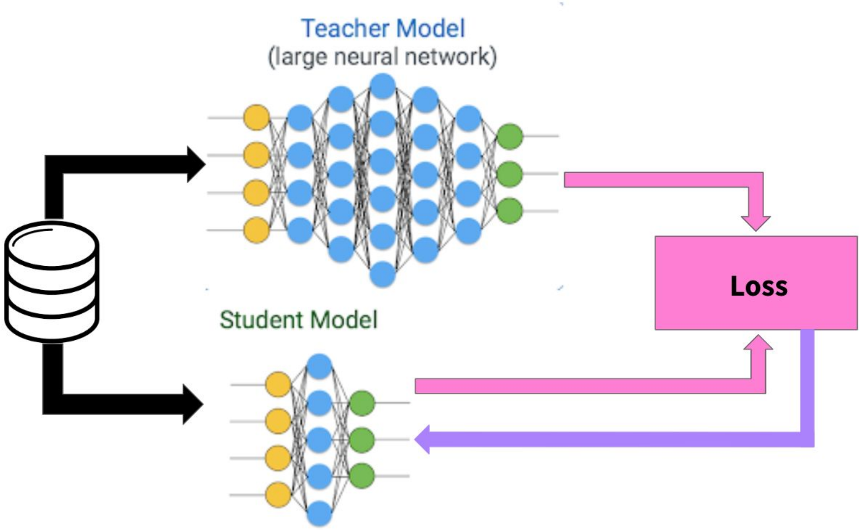
Motivation



Credits: http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf

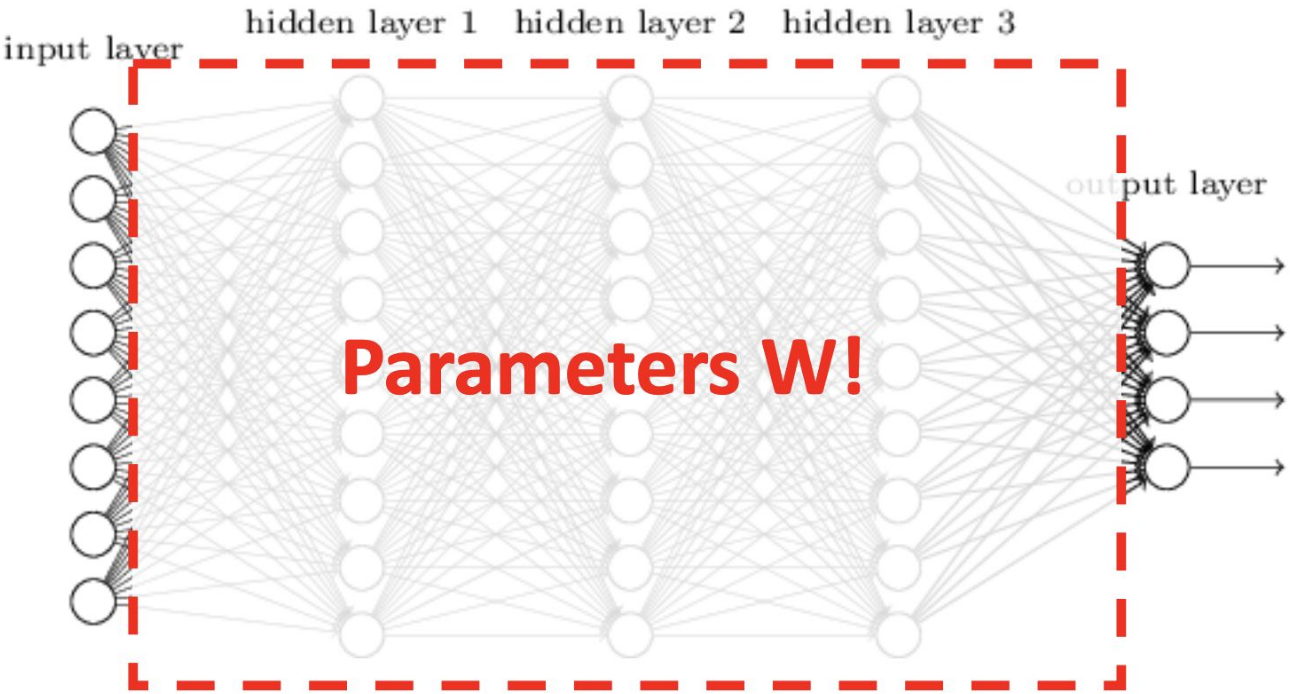
Knowledge Distillation

Knowledge distillation is a process of distilling or transferring the knowledge from a (set of) large, cumbersome model(s) to a lighter, easier-to-deploy single model



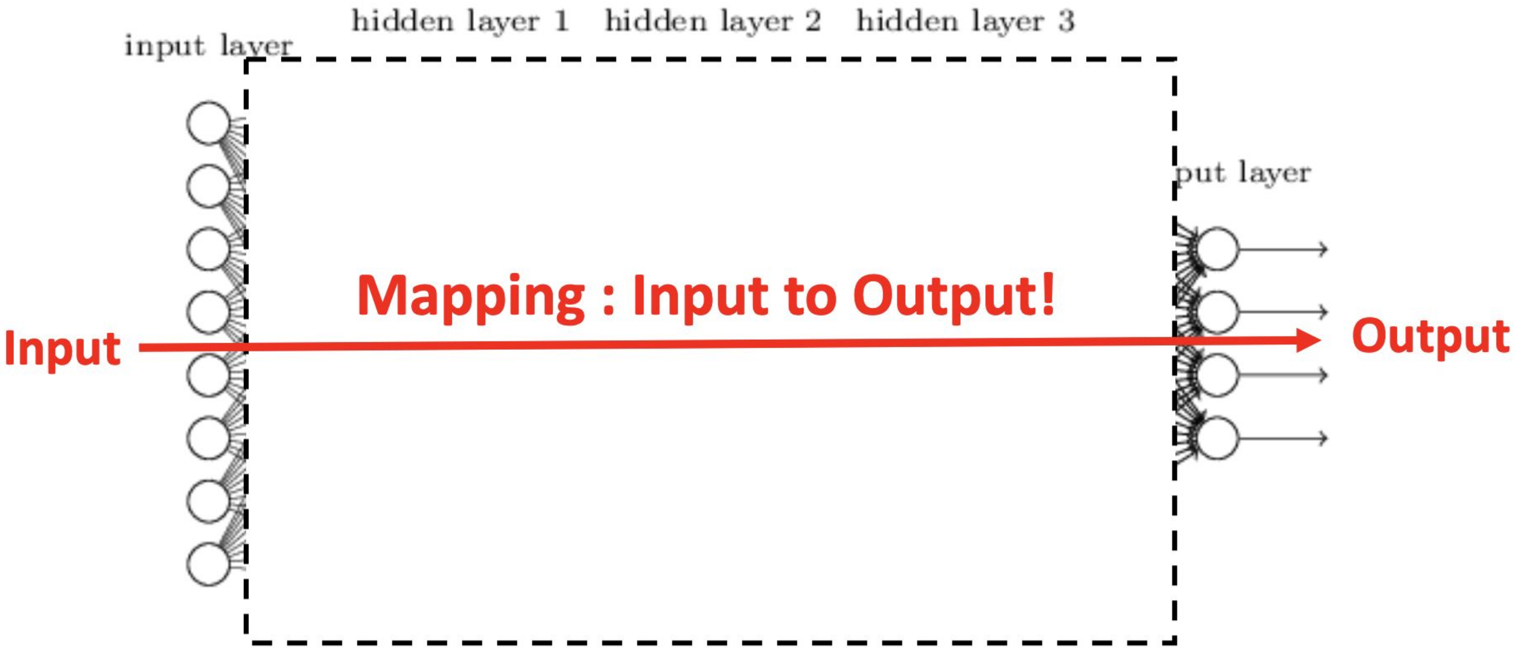
Credits: http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf
https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764

What is knowledge?



Credits: http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf

What is knowledge?

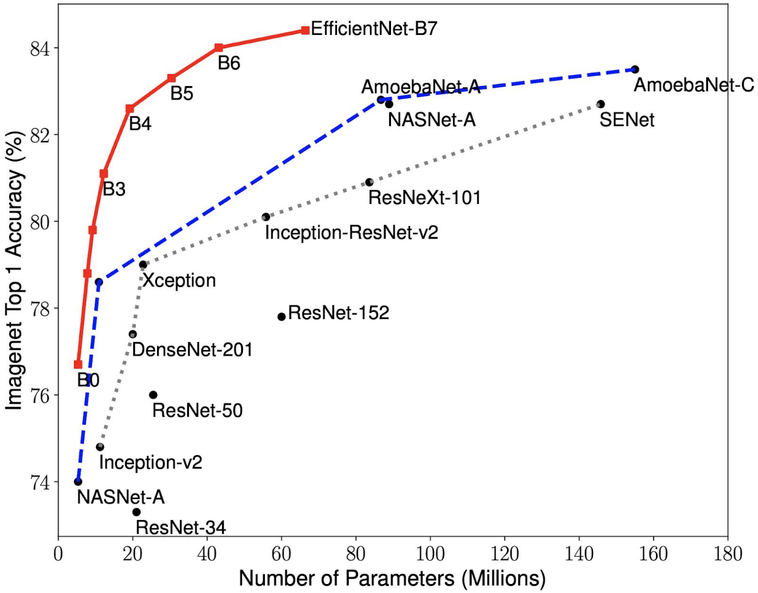


A more abstract view of the **knowledge**, that frees it from any particular instantiation, is that it is a **learned mapping from input vectors to output vectors**.

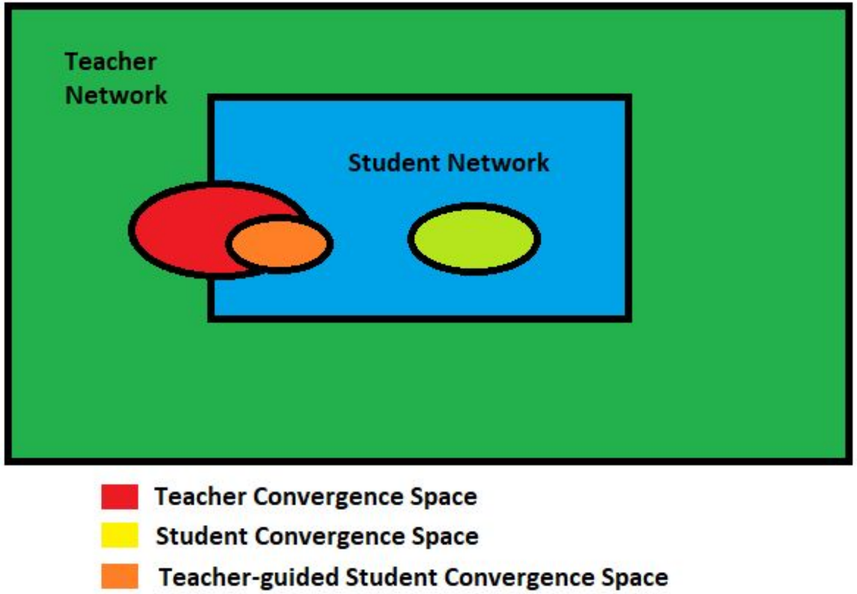
Credits: http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf

Why Knowledge in bigger models is better?

Higher number of parameters -> better performance



We can train smaller (student) model to mimic a behaviour of bigger model (teacher)

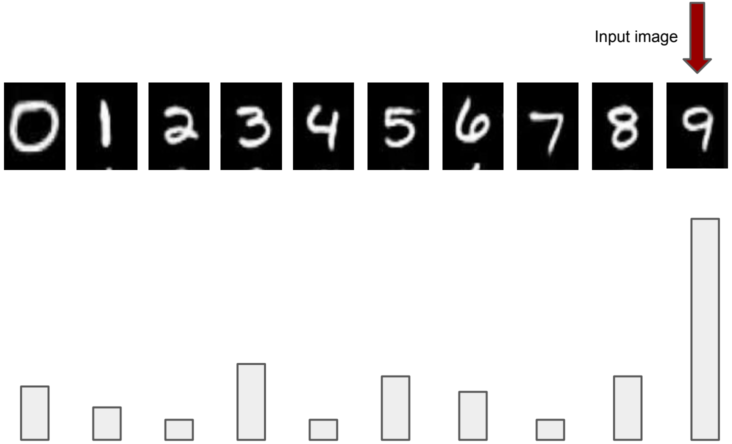


Credits: https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html
https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764

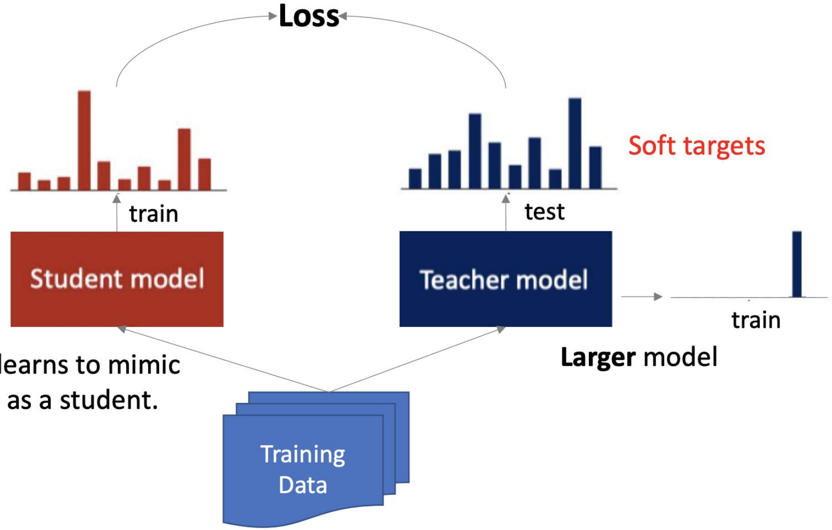
Knowledge Distillation via Soft Targets

Basic concept

Idea: Train less redundant model by using outputs from big models



Output probability distribution



Small model learns to mimic the teacher as a student.

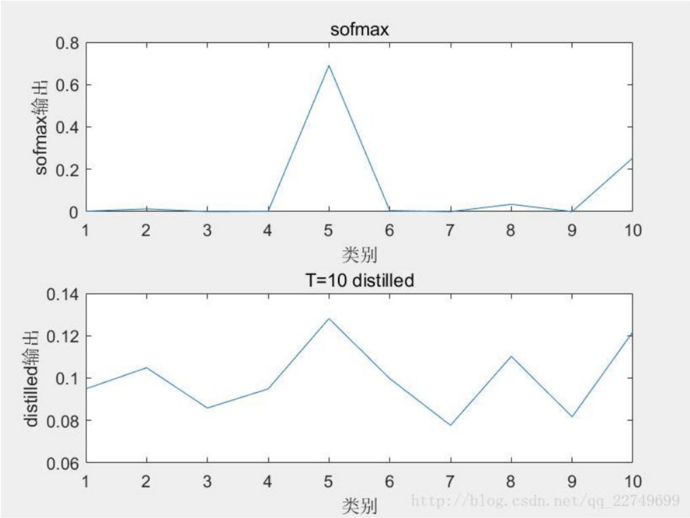
Credits: http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf
https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764

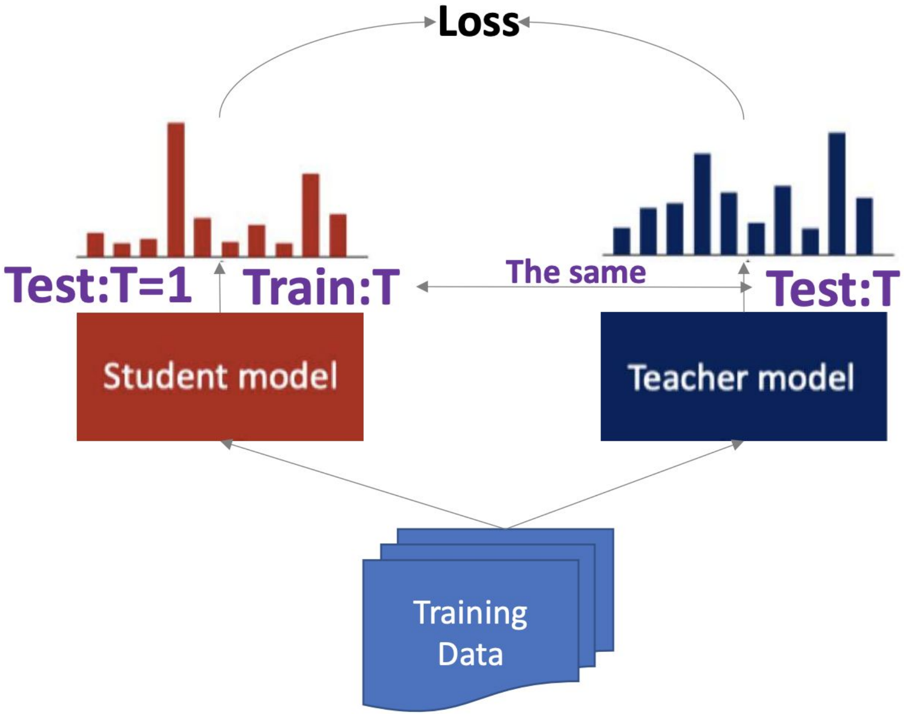
Softmax with Temperature

Logits

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

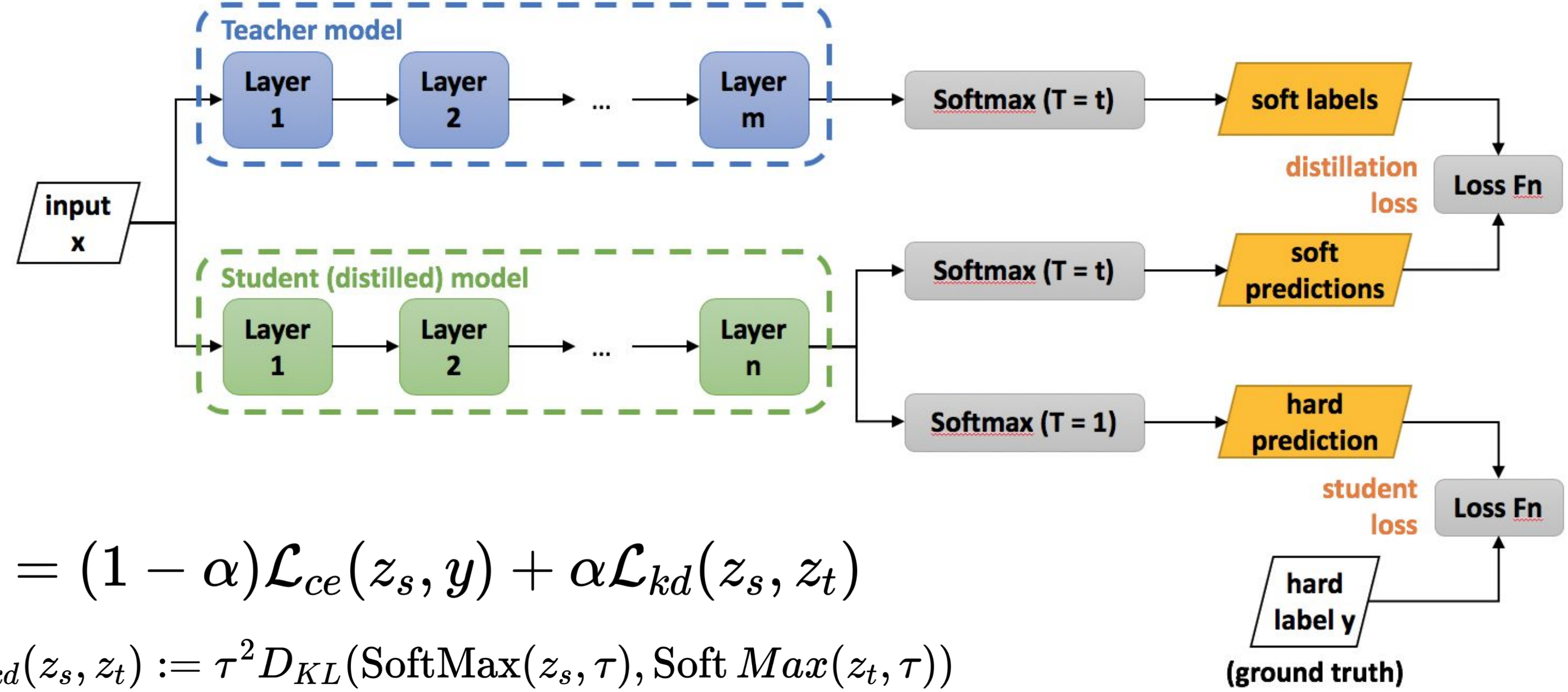
Temperature





Credits: http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf
https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764

Full Method



$$\mathcal{L} = (1-\alpha)\mathcal{L}_{ce}(z_s, y) + \alpha\mathcal{L}_{kd}(z_s, z_t)$$

$$\mathcal{L}_{kd}(z_s, z_t) := \tau^2 D_{KL}(\mathrm{SoftMax}(z_s, \tau), \mathrm{Soft}\,Max(z_t, \tau))$$

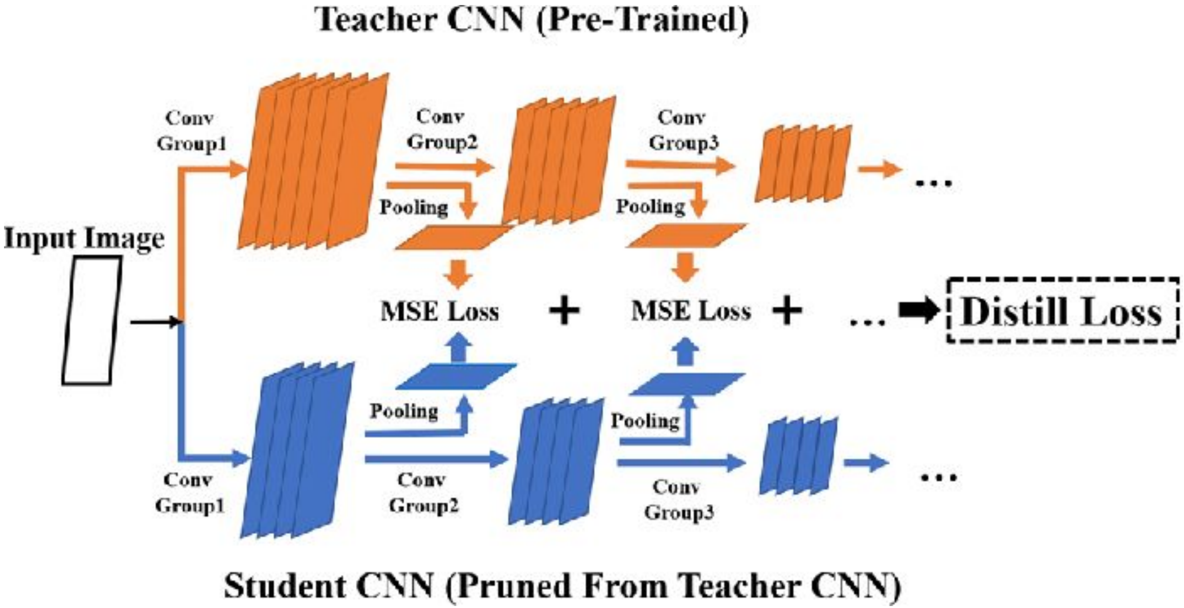
Credits: https://devopedia.org/knowledge-distillation
Hinton et al. **Distilling the Knowledge in a Neural Network**, NIPS (2014)

FitNets: Hits for Thin Deep Nets

FitNets: Concept

Idea: Use intermediate representations to guide training of student:

$$\mathcal{L}_{HT}(\mathbf{W_{Guided}}, \mathbf{W_r}) = \frac{1}{2}\left\|u_h(\mathbf{x}; \mathbf{W_{Hint}}) - r(v_g(\mathbf{x}; \mathbf{W_{Guided}}); \mathbf{W_r})\right\|^2$$



Credits: Romero et al. **FitNets: Hints for Thin Deep Nets**, ICLR 2015

FitNets: Results

Algorithm	# params	Accuracy
Compression		
FitNet	$\sim$2.5M	**91.61%**
Teacher	$\sim$9M	90.18%
Mimic single	$\sim$54M	84.6%
Mimic single	$\sim$70M	84.9%
Mimic ensemble	$\sim$70M	85.8%
State-of-the-art methods		
Maxout		90.65%
Network in Network		91.2%
Deeply-Supervised Networks		**91.78%**
Deeply-Supervised Networks (19)		88.2%

Table 1: Accuracy on CIFAR-10

Algorithm	# params	Accuracy
Compression		
FitNet	$\sim$2.5M	**64.96%**
Teacher	$\sim$9M	63.54%
State-of-the-art methods		
Maxout		61.43%
Network in Network		64.32%
Deeply-Supervised Networks		**65.43%**

Table 2: Accuracy on CIFAR-100

Credits: Romero et al. **FitNets: Hints for Thin Deep Nets**, ICLR 2015

Attention-based Knowledge Distillation

Attention transfer

- Attention plays a critical role in human visual experience
- It also has demonstrated important role in NNs
- We can use it to improve teacher-student training

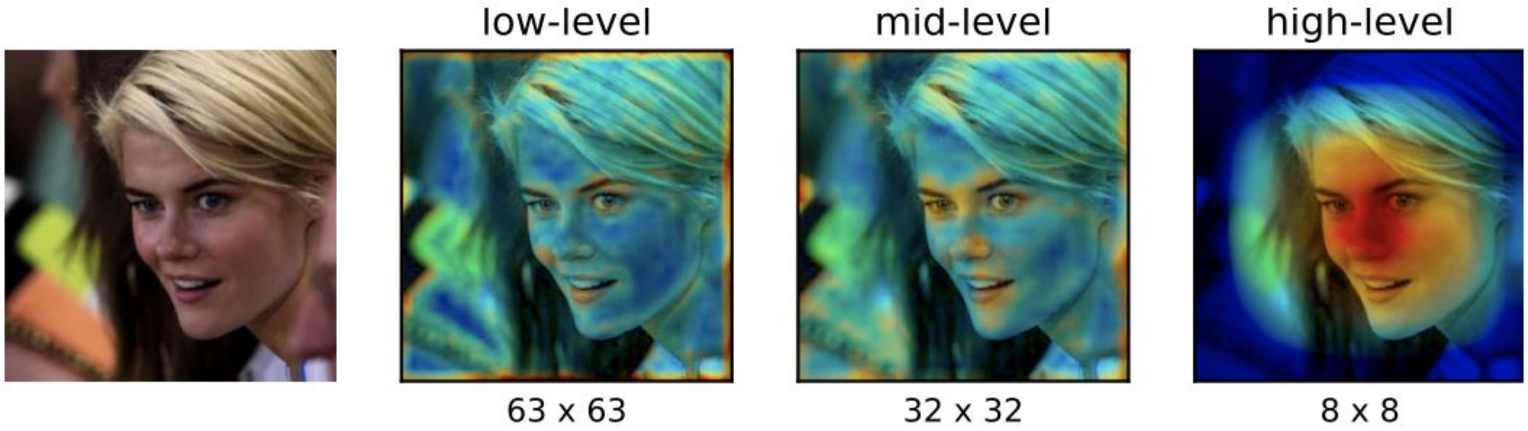


Figure: Sum of absolute values attention maps over different levels of a network trained for face recognition. Mid-level attention maps have higher activation level around eyes, nose and lips, high-level activations correspond to the whole

Credits:
Zagoruyko et al. **Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer**, ICLR 20117

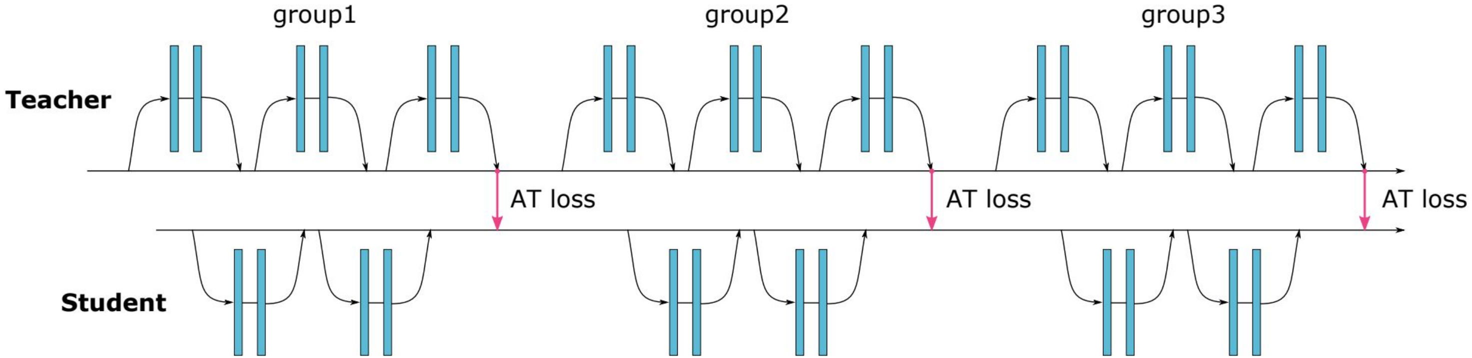
Attention transfer

Attention loss: $$\mathcal{L}_{AT} = \mathcal{L}(\mathbf{W}_S, x) + \frac{\beta}{2} \sum_{j \in \mathcal{I}} \| \frac{Q_S^j}{\|Q_S^j\|_2} - \frac{Q_T^j}{\|Q_T^j\|_2} \|_p$$

where $Q_S^j = vec(F(A_S^j))$ and $Q_T^j = vec(F(A_T^j))$ are vectorized j-th pair of teacher - student attention maps

The work considers several types of attention:

- sum of absolute values: $F_{\text{sum}}(A) = \sum_{i=1}^{C} |A_i|$
- sum of absolute values raised to the power of p (where $p > 1$): $F_{\text{sum}}^p(A) = \sum_{i=1}^{C} |A_i|^p$
- max of absolute values raised to the power of p (where $p > 1$): $F_{\max}^p(A) = \max_{i=1,C} |A_i|^p$



Methods comparison

KD between similar architectures (CIFAR100):

Teacher Student	wrn40-2 wrn16-2	wrn40-2 wrn40-1	resnet56 resnet20	resnet32×4 resnet8×4	vgg13 vgg8
Teacher	76.46	76.46	73.44	79.63	75.38
Student	73.64	72.24	69.63	72.51	70.68
KD [16]	74.92	73.54	70.66	73.33	72.98
FitNet [37]	75.75	74.12	71.60	74.31	73.54
AT [46]	75.28	74.45	**71.78**	74.26	73.62

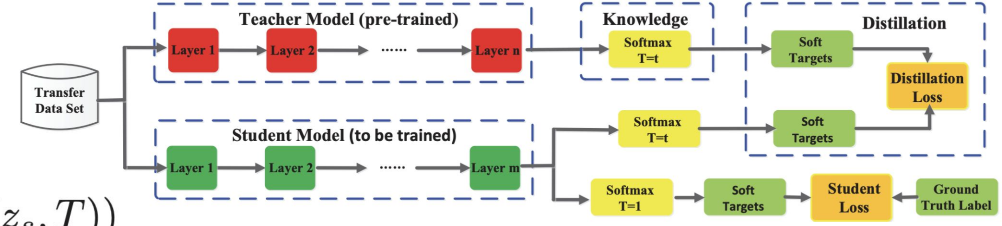
KD between different architectures:

Teacher Student	vgg13 MobileNetV2	ResNet50 MobileNetV2	ResNet50 vgg8	resnet32×4 ShuffleV1	resnet32×4 ShuffleV2	wrn40-2 ShuffleV1
Teacher	75.38	79.10	79.10	79.63	79.63	76.46
Student	65.79	65.79	70.68	70.77	73.12	70.77
KD [16]	67.37	67.35	73.81	74.07	74.45	74.83
FitNet [37]	68.58	68.54	73.84	74.82	75.11	75.55
AT [46]	<u>69.34</u>	69.28	73.45	74.76	75.30	75.61

Credits: Xu et al. **Knowledge Distillation Meets Self-Supervision**, ECCV 2020

Knowledge Distillation Types

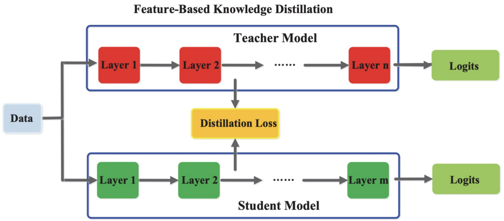
- Response-Based Knowledge



$$L_{ResD}(p(z_t,T),p(z_s,T)) = \mathcal{L}_R(p(z_t,T),p(z_s,T))$$

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

- Feature-Based Knowledge

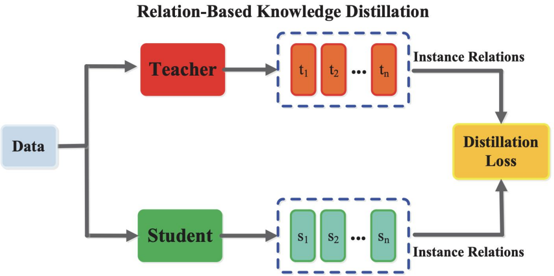


$$L_{FeaD}\big(f_t(x), f_s(x)\big) = \mathcal{L}_F\big(\Phi_t(f_t(x)), \Phi_s(f_s(x))\big)$$

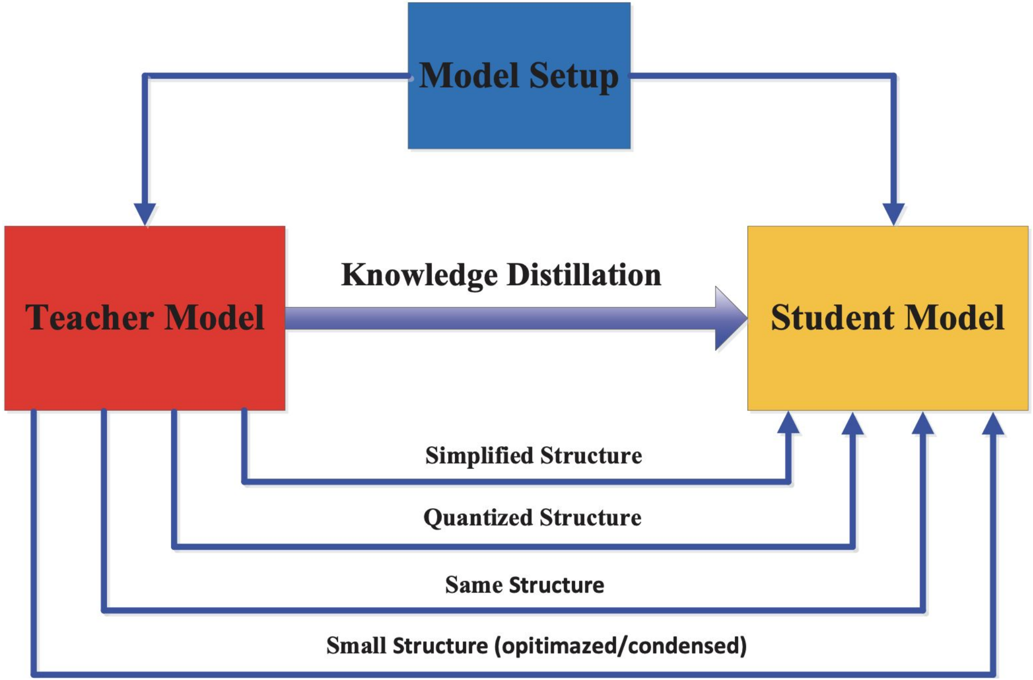
- Relation-Based Knowledge

$$L_{RelD}(f_t, f_s) = \mathcal{L}_{R^1}\big(\Psi_t(\hat{f}_t, \check{f}_t), \Psi_s(\hat{f}_s, \check{f}_s)\big)$$

$$L_{RelD}(F_t, F_s) = \mathcal{L}_{R^2}\big(\psi_t(t_i, t_j), \psi_s(s_i, s_j)\big)$$



Student Setup



Summary

Knowledge Distillation is a powerful technique to train small neural network using information from big pretrained network

Pros:

- Improves model performance

Cons:

- Capacity gap
- Controversy of the method

Sources

1. Hinton et al. **Distilling the Knowledge in a Neural Network**, NIPS 2014
2. Romero et al. **FitNets: Hints for Thin Deep Nets**, ICLR 2015
3. Zagoruyko et al. **Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer**, ICLR 2017
4. https://devopedia.org/knowledge-distillation
5. https://medium.com/analytics-vidhya/knowledge-distillation-in-a-deep-neural-network-c9dd59aff89b
6. https://neptune.ai/blog/knowledge-distillation