Distilling the Knowledge in a Neural Network 2015

Geoffrey Hinton, Oriol Vinyals, Jeff Dean

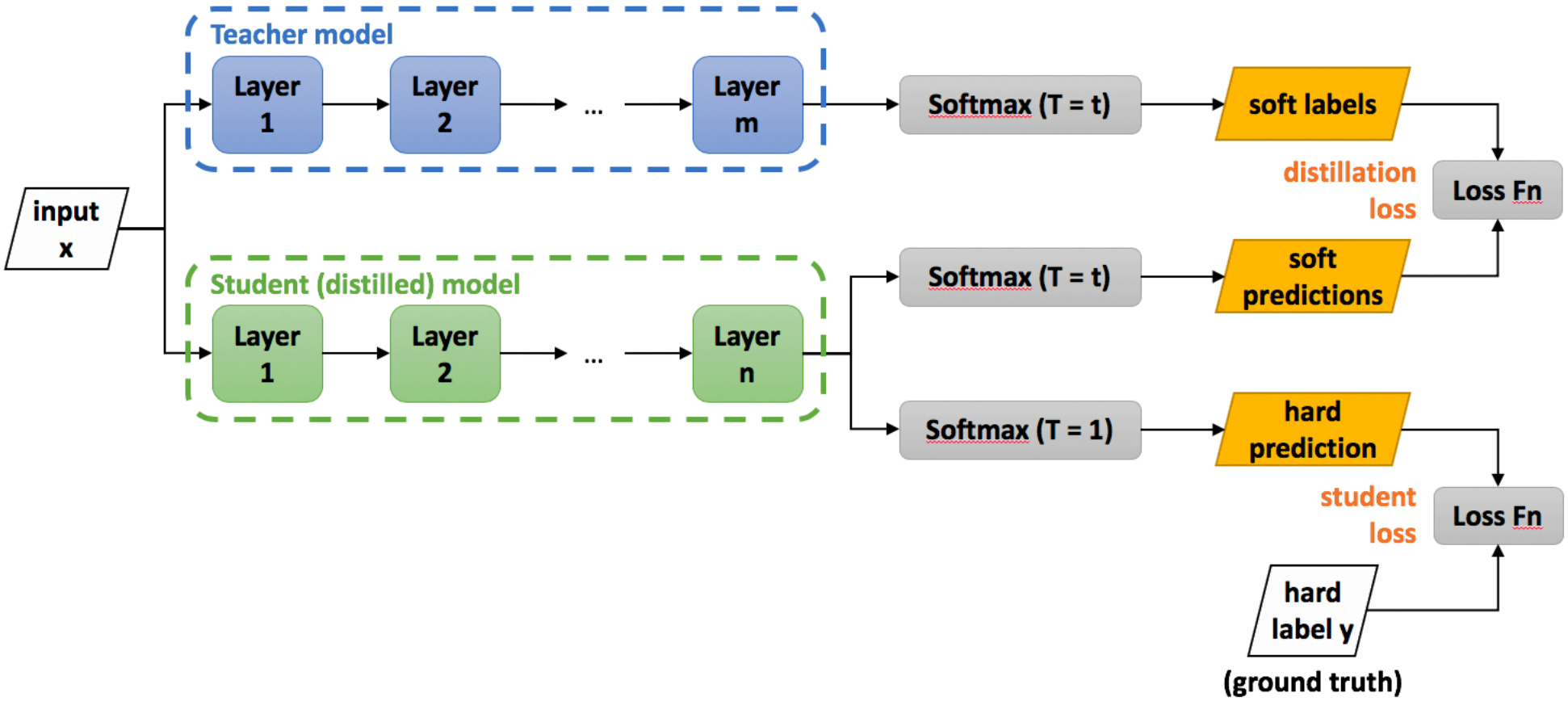
- Transfer the knowledge from the cumbersome model to a small model.
- Soft targets (分类问题):
  - Teacher ensemble of models： arithmetic or geometric mean
  - Neural Network

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

  - T is a temperature controls the softness , using a higher value for T produces a softer probability distribution over classes. [0.99,0.01]-> [0.9, 0.1]

- Loss Function:

$$\alpha * \underbrace{\mathcal{H}(\hat{y}_S, y)}_{Cross-entronpy} + \beta * \mathcal{H}(\sigma(\hat{y}_S, T), \sigma(\hat{y}_T, T))$$



- MNIST
  - (1200,1200) Teacher: 67 errors
  - (800, 800) Student without KD: 146 errors
  - Student with KD 74 errors

- Speech recognition
  - Baseline : 8 hidden layers each containing 2560 relu: 58.9% frame accuracy
  - Teacher: 10xEnsemble: 61.1
  - Student : 60.8

Why Soft Targets Work

- When the soft targets have high entropy, they provide much more information per training case than hard target
  - Similarity between output categories.

- much less variance in the gradient between training cases

Born-Again Neural Networks ICML 2018

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti , Anima Anandkumar

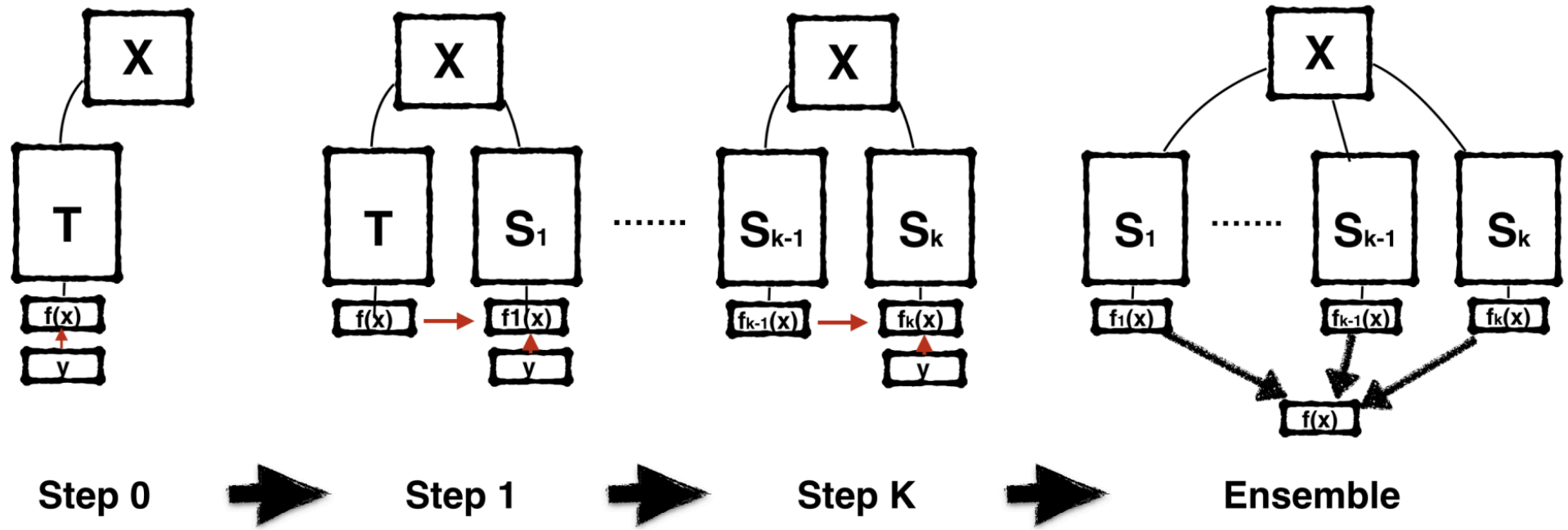


Figure 1. **Graphical representation of the BAN training procedure**: during the first step the teacher model T is trained from the labels Y. Then, at each consecutive step, a new identical model is initialized from a different random seed and trained from the supervision of the earlier generation. At the end of the procedure, additional gains can be achieved with an ensemble of multiple students generations.

- Single sample gradient between student logits zj and teacher logits tj with respect to the ith output ：

$$\frac{\partial L}{\partial z_i} = q_i - p_i = \frac{\exp z_i}{\sum_j \exp z_j} - \frac{\exp t_i}{\sum_j \exp t_j}$$

- b个sample 梯度

$$\sum_s^b \sum_i^n \frac{\partial L_i}{\partial z_{i,s}} = \sum_s^b \sum_i^n \left(\frac{\exp z_{i,s}}{\sum \exp z_{i,s}} - \frac{\exp t_{j,s}}{\sum \exp t_{i,s}} \right)$$

$$= \underbrace{\sum_s^b (q_{*,s} - p_{*,s} \overbrace{y_{*,s}}^{=1})}_{\text{label}} + \underbrace{\sum_s^b \sum_{i \neq *} (q_{i,s} - p_{i,s})}_{\text{wrong prediction}}$$

- 如果teacher is confident i.e. P*,s=1， 那么等价直接训练
- 否则，相当于给与样本b的梯度 一个相对较低权重

KD is performing a kind of importance weighting?

- Confidence Weighted by Teacher Max (CWTM): exclude the effect of all the teacher's output except for the argmax dimension.

$$\sum_{s}^{b} \frac{\max p_{.,s}}{\sum_{u} \max p_{.,u}} (q_{*,s} - y_{*,s})$$

- Dark knowledge with Permuted Predictions (DKPP): randomly permute each output dimension except the argmax one

$$\sum_{s}^{b} \sum_{i}^{n} \frac{\partial L_i}{\partial z_{i,s}} = \sum_{s}^{b} (q_{*,s} - \max p_{.,s}) + \sum_{s}^{b} \sum_{i}^{n-1} q_{i,s} - \phi(p_{j,s})$$

Table 2. **Test error on CIFAR-100** *Left Side:* DenseNet of different depth and growth factor and respective BAN student. BAN models are trained only with the teacher loss, BAN+L with both label and teacher loss. CWTM are trained with sample importance weighted label, the importance of the sample is determined by the max of the teacher's output. DKPP are trained only from teacher outputs with all the dimensions but the argmax permuted. *Right Side:* test error on CIFAR-100 sequence of BAN-DenseNet, and the BAN-ensembles resulting from the sequence. Each BAN in the sequence is trained from cross-entropy with respect to the model at its left. BAN and BAN-1 models are trained from Teacher but have different random seeds. We include the teacher as a member of the ensemble for Ens*3 for 80-120 since we did not train a BAN-3 for this configuration.

Network	Teacher	BAN	BAN+L	CWTM	DKPP	BAN-1	BAN-2	BAN-3	Ens*2	Ens*3
DenseNet-112-33	18.25	**16.95**	17.68	17.84	17.84	17.61	17.22	**16.59**	15.77	15.68
DenseNet-90-60	17.69	**16.69**	16.93	17.42	17.43	16.62	**16.44**	16.72	15.39	15.74
DenseNet-80-80	17.16	**16.36**	16.5	17.16	16.84	16.26	16.30	**15.5**	15.46	15.14
DenseNet-80-120	16.87	**16.00**	16.41	17.12	16.34	**16.13**	16.13	/	**15.13**	**14.9**

- L: both labels and teacher
- KD does not simply contribute information on each specific non-correct output
  - information contained in pre-trained models can be used to rebalance the training set, by giving less weight to training samples

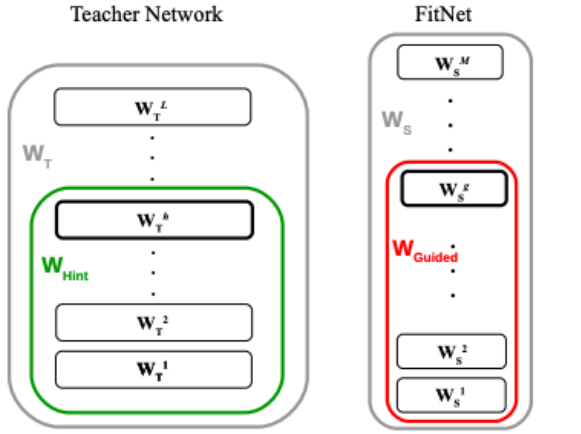
Fitnets: Hints for thin deep nets. 2015 ICLR

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

- r is the regressor function on top of the guided layer with parameters Wr.

Algorithm 1 FitNet Stage-Wise Training.
The algorithm receives as input the trained parameters $\mathbf{W_T}$ of a teacher, the randomly initializ parameters $\mathbf{W_S}$ of a FitNet, and two indices h and g corresponding to hint/guided layers, respe tively. Let $\mathbf{W_{Hint}}$ be the teacher's parameters up to the hint layer h. Let $\mathbf{W_{Guided}}$ be the FitNe parameters up to the guided layer g. Let $\mathbf{W_r}$ be the regressor's parameters. The first stage consists pre-training the student network up to the guided layer, based on the prediction error of the teache hint layer (line 4). The second stage is a KD training of the whole network (line 6).

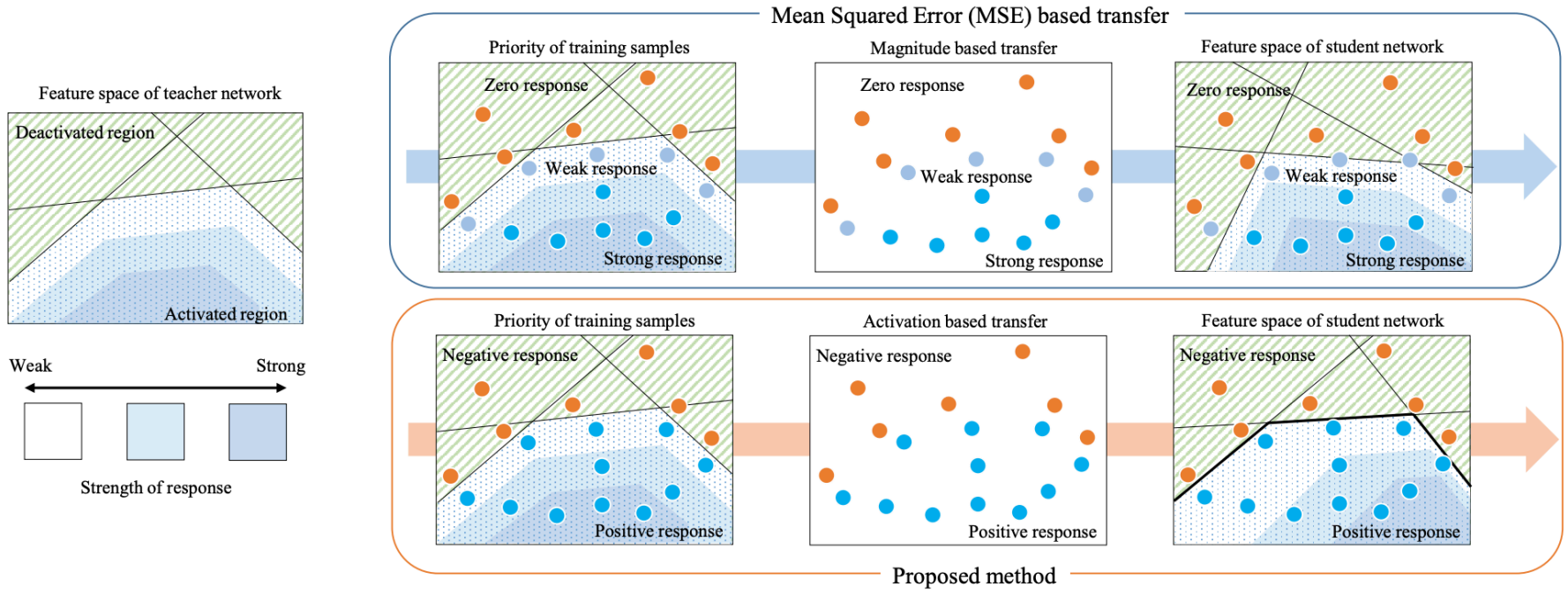
Input: $\mathbf{W_S}, \mathbf{W_T}, g, h$
Output: $\mathbf{W_S^*}$
1: $\mathbf{W_{Hint}} \leftarrow \{\mathbf{W_T}^1, \ldots, \mathbf{W_T}^h\}$
2: $\mathbf{W_{Guided}} \leftarrow \{\mathbf{W_S}^1, \ldots, \mathbf{W_S}^g\}$
3: Intialize $\mathbf{W_r}$ to small random values
4: $\mathbf{W^*_{Guided}} \leftarrow \underset{\mathbf{W_{Guided}}}{\text{argmin}}\ \mathcal{L}_{HT}(\mathbf{W_{Guided}}, \mathbf{W_r})$
5: $\{\mathbf{W_S}^1, \ldots, \mathbf{W_S}^g\} \leftarrow \{\mathbf{W_{Guided}}^{*1}, \ldots, \mathbf{W_{Guided}}^{*g}\}$
6: $\mathbf{W_S^*} \leftarrow \underset{\mathbf{W_S}}{\text{argmin}}\ \mathcal{L}_{KD}(\mathbf{W_S})$



(a) Teacher and Student Networks

Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons 2019 AAAI

- for initializing the student network before the classification training



- Activation transfer loss

whether a neuron is activated or not, i.e.,

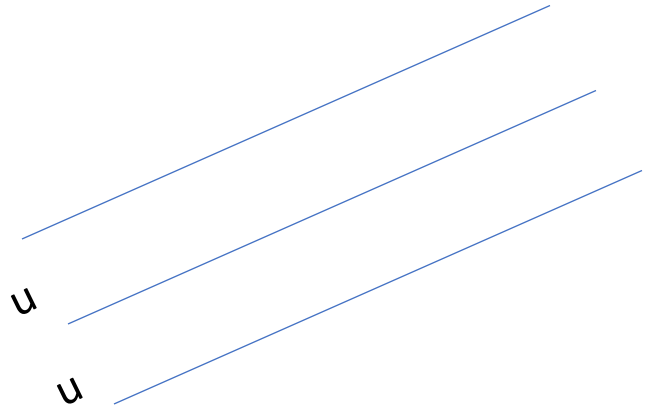
$$\rho(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The new loss to transfer the neuron activation is given by

$$\mathcal{L}(\boldsymbol{I}) = \|\rho(\mathcal{T}(\boldsymbol{I})) - \rho(\mathcal{S}(\boldsymbol{I}))\|_1, \tag{3}$$

- ρ() is a discrete function, the activation transfer loss can not be minimized by SGD, they use an alternative loss for SVM

$$\mathcal{L}(\boldsymbol{I}) = \|\rho(\mathcal{T}(\boldsymbol{I})) \odot \sigma(\mu\mathbf{1} - \mathcal{S}(\boldsymbol{I})) + (\mathbf{1} - \rho(\mathcal{T}(\boldsymbol{I}))) \odot \sigma(\mu\mathbf{1} + \mathcal{S}(\boldsymbol{I}))\|_2^2$$



Scheme

- Transfer the neuron responses for initializing the student network before the classification training
- After the network is initialized based on the transfer loss, it is trained for classification based on the cross-entropy loss.

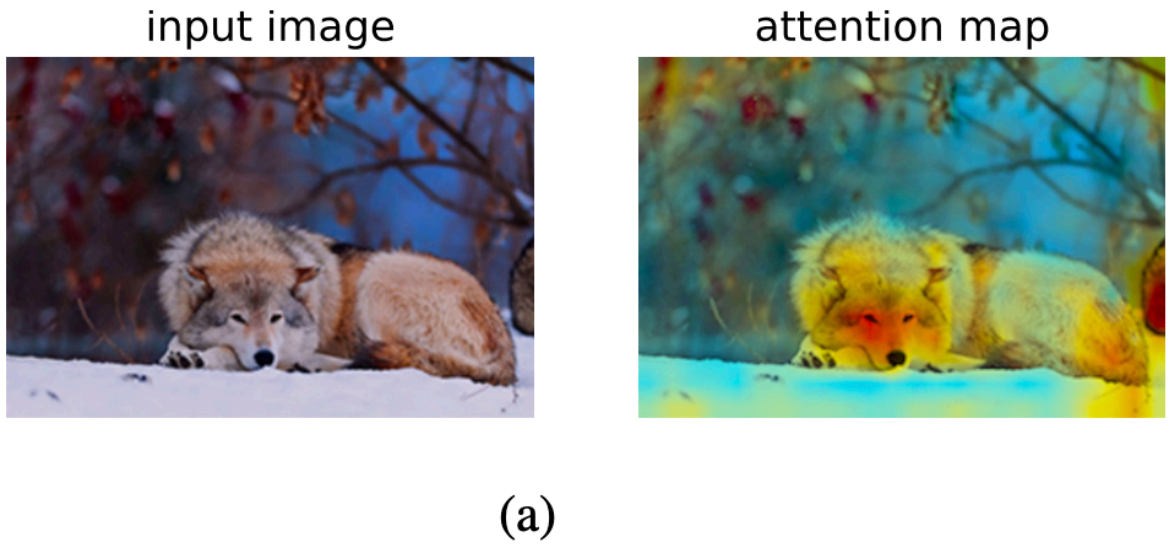
Cifar 10 training was performed **for 120 epochs using 10% of the training data.**

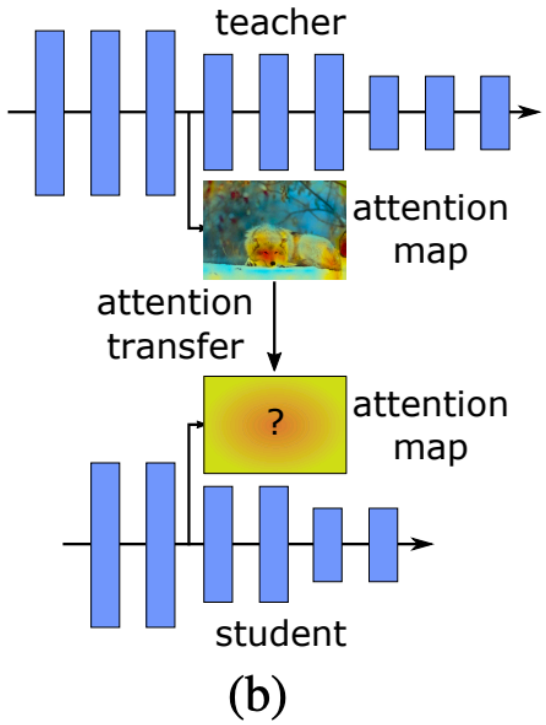
Table 3: Performance for various sizes of networks. Table shows error rate(%) on test set.

Compression type	Teacher	Student	Size ratio	KD	FITNET	FSP	AT	Jacobian	Proposed
Depth	WRN 22-4	WRN 10-4	27.9%	22.98%	23.34%	22.99%	18.06%	18.28%	**14.05%**
Channel	WRN 16-4	WRN 16-2	25.2%	20.48%	19.98%	19.78%	14.81%	14.41%	**11.62%**
Depth & Channel	WRN 22-4	WRN 16-2	16.1%	21.21%	21.42%	20.80%	15.61%	15.03%	**13.09%**
Tiny network	WRN 22-4	WRN 10-1	1.8%	29.57%	29.18%	28.70%	29.44%	28.70%	**23.27%**
Same network	WRN 16-4	WRN 16-4	100%	18.29%	17.91%	17.81%	12.03%	11.28%	**6.63%**

PAYING MORE ATTENTION TO ATTENTION: IMPROVING THE PERFORMANCE OF CONVOLUTIONAL NEURAL NETWORKS VIA ATTENTION TRANSFER ICLR 2017

spatial areas of the input the network focuses most for taking its output decision t contain valuable information about the network





Attention maps of CNNs

- Activation-based Activation tensor $A^{C\times W\times H}$
- mapping function: F: $R^{C\times W\times H} \rightarrow R^{W\times H}$

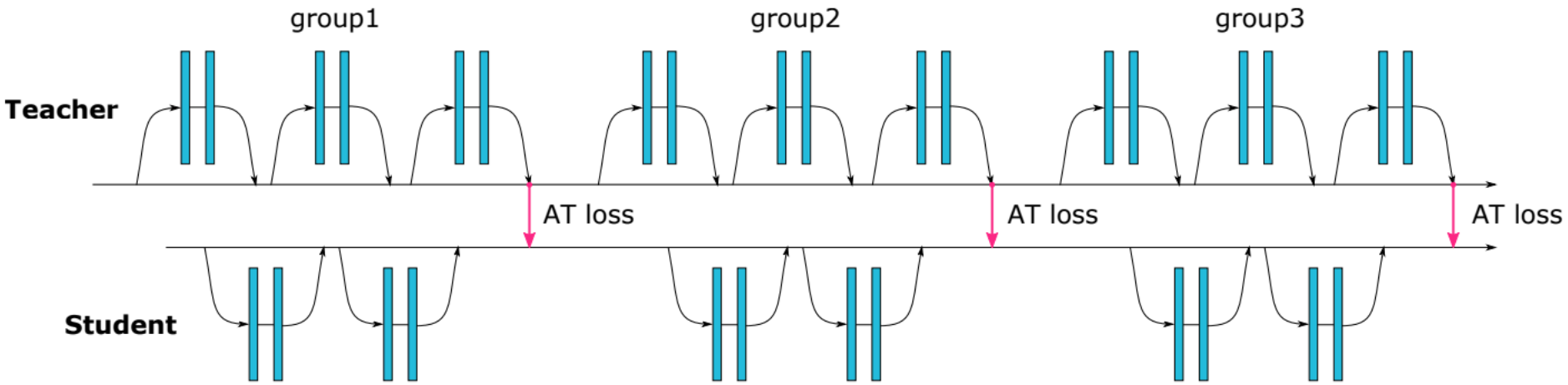
Activation-based spatial attention maps:

- 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 $p > 1$): $F_{\text{max}}^{p}(A) = \max_{i=1,C} |A_i|^p$

Attention maps of CNNs

Loss Function

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



Spatial attention maps of CNNs

GRADIENT-BASED ATTENTION TRANSFER:

- gradient of the loss w.r.t input:

$$J_S = \frac{\partial}{\partial x}\mathcal{L}(\mathbf{W_S}, x), J_T = \frac{\partial}{\partial x}\mathcal{L}(\mathbf{W_T}, x)$$

- Loss

$$\mathcal{L}_{AT}(\mathbf{W_S}, \mathbf{W_T}, x) = \mathcal{L}(\mathbf{W_S}, x) + \frac{\beta}{2}||J_S - J_T||_2$$

Results

- Cirfar10+ activation transfer F^2_sum()

F-ActT means full-activation transfer

student	teacher	student	AT	F-ActT	KD	AT+KD	teacher
NIN-thin, 0.2M	NIN-wide, 1M	9.38	8.93	9.05	8.55	8.33	7.28
WRN-16-1, 0.2M	WRN-16-2, 0.7M	8.77	7.93	8.51	7.41	7.51	6.31
WRN-16-1, 0.2M	WRN-40-1, 0.6M	8.77	8.25	8.62	8.39	8.01	6.58
WRN-16-2, 0.7M	WRN-40-2, 2.2M	6.31	5.85	6.24	6.08	5.71	5.23

- IMAGENET

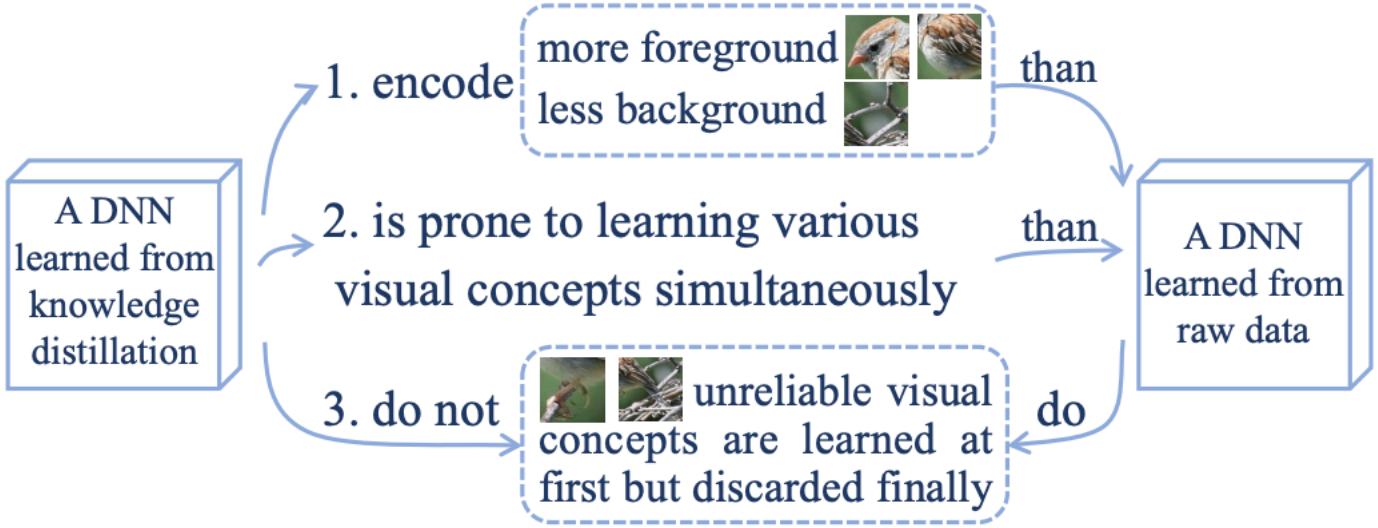
Teacher: resnet34

Student : resnet18

Attention transfer achieved 1.1% top-1 and 0.8% top-5 better validation accuracy

Explaining Knowledge Distillation by Quantifying the Knowledge CVPR 2020

- Success of knowledge distillation (中间层feature)



- Visual concepts: object parts like tails, heads

- intermediate-layer feature f*=f(x)

$$H(X') \quad s.t. \ \forall x' \in X', \quad \| f(x') - f^* \|^2 \leq \tau$$

- 假设 x’ follows an i.i.d. Gaussian distribution

$$x' \sim \mathcal{N}(x, diag(\sigma_1^2 ..., \sigma_n^2))$$

where σi controls the magnitude of the perturbation at each i-th pixel

- entropy H(X’) of the entire image can be decomposed into pixel-level entropies {Hi} as follows. (16 × 16 grid)

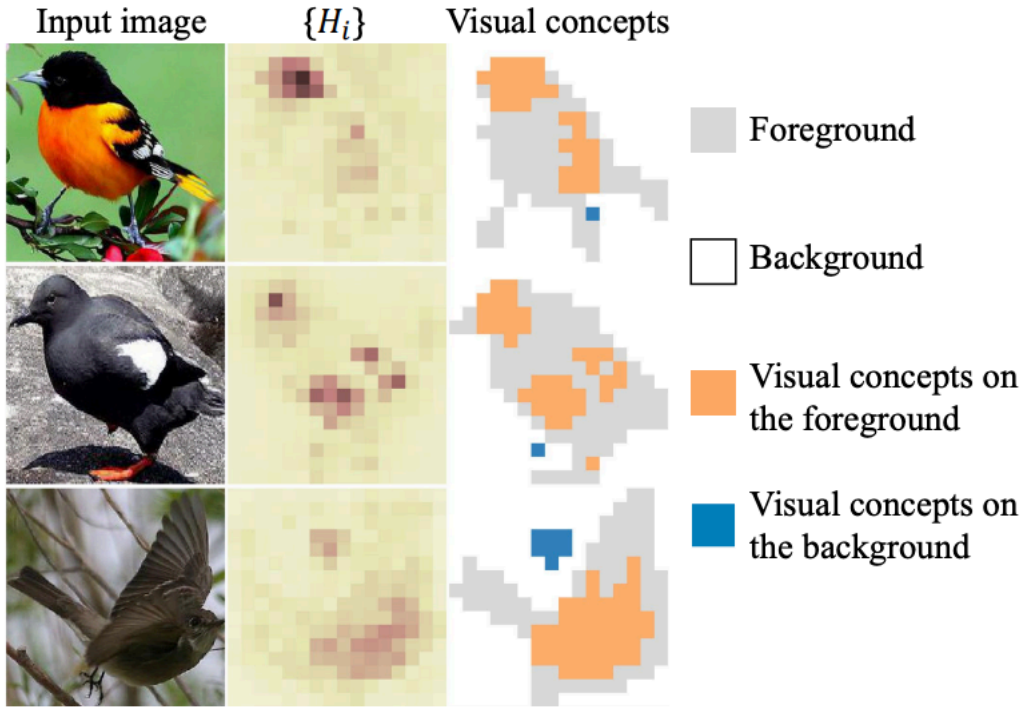
$$H(X') = \sum_{i}^{n} H_i, \ \ H_i = \log \sigma_i + 0.5 \log 2\pi e$$

- image regions with low pixel-wise entropies {Hi} can be considered to represent relatively valid visual concepts.

$$N_{\text{concept}}^{\text{bg}}(x) = \sum_{i \in \Lambda_{\text{bg}} \ w.r.t. \ x} \mathbb{1}(\overline{H} - H_i > b),$$

$$N_{\text{concept}}^{\text{fg}}(x) = \sum_{i \in \Lambda_{\text{fg}} \ w.r.t. \ x} \mathbb{1}(\overline{H} - H_i > b),$$

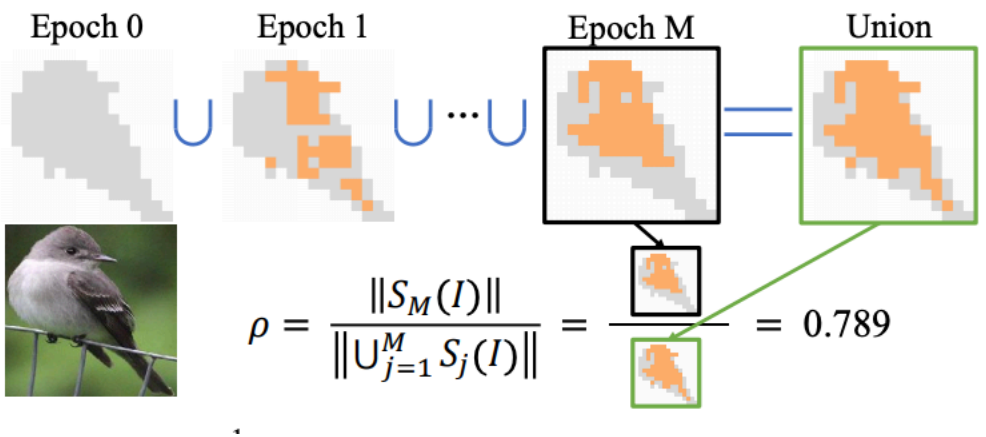
$$\lambda = \mathbb{E}_{x \in \mathbf{I}}[N_{\text{concept}}^{\text{fg}}(x) / (N_{\text{concept}}^{\text{fg}}(x) + N_{\text{concept}}^{\text{bg}}(x))]$$



- $\bar{H}$ average entropy value of the background
- larger λ value denotes the DNN is more discriminative

- Let S1(I), S2(I), . . . , SM(I) denote the set of visual concepts on the foreground of image I in different epochs

$$\rho = \frac{\|S_M(I)\|}{\|\bigcup_{j=1}^{M} S_j(I)\|}$$



- A high value of ρ indicates that the DNN is optimized more stably; vice versa.

- $\hat{m} = \arg\max_k N_k^{fg}(I)$: 图片I 的task relevant visual concept 最多epoches

$$D_{\text{mean}} = \mathop{\mathbb{E}}_{I\in\mathbf{I}}\left[\sum_{k=1}^{\hat{m}} \frac{\|w_k - w_{k-1}\|}{\|w_0\|}\right],$$

$$D_{\text{std}} = \mathop{Var}_{I\in\mathbf{I}}\left[\sum_{k=1}^{\hat{m}} \frac{\|w_k - w_{k-1}\|}{\|w_0\|}\right]$$

- Dmean indicates whether a DNN learns visual concepts quickly. Dstd whether a DNN learns various visual concepts simultaneously.

Network	Layer		CUB200-2011 dataset						ILSVRC-2013 DET dataset						Pascal VOC 2012 dataset					
			$N^{fg}_{concept}\uparrow$	$N^{bg}_{concept}\downarrow$	$\lambda\uparrow$	$D_{mean}\downarrow$	$D_{std}\downarrow$	$\rho\uparrow$	$N^{fg}_{concept}\uparrow$	$N^{bg}_{concept}\downarrow$	$\lambda\uparrow$	$D_{mean}\downarrow$	$D_{std}\downarrow$	$\rho\uparrow$	$N^{fg}_{concept}\uparrow$	$N^{bg}_{concept}\downarrow$	$\lambda\uparrow$	$D_{mean}\downarrow$	$D_{std}\downarrow$	$\rho\uparrow$
AlexNet	FC_1	S	**36.60**	**4.00**	**0.90**	8.35	25.09	**0.57**	**49.46**	**0.66**	**0.99**	**0.48**	**0.10**	**0.62**	**25.84**	**5.86**	**0.79**	**1.14**	**0.56**	0.43
		B	24.13	5.65	0.81	**4.81**	**14.54**	0.52	41.00	0.92	0.98	1.32	0.31	0.61	20.30	6.08	0.77	2.00	2.21	**0.44**
	FC_2	S	**38.13**	**3.50**	**0.92**	**3.77**	**3.97**	0.49	**57.86**	1.70	0.98	**0.28**	**0.01**	0.60	**31.81**	7.29	**0.81**	**0.62**	**0.07**	**0.47**
		B	23.33	5.48	0.80	5.36	20.79	**0.49**	42.24	**0.96**	**0.98**	1.15	0.15	**0.60**	21.85	**6.56**	0.77	2.04	1.46	0.44
	FC_3	S	**33.20**	**4.31**	**0.89**	**8.13**	**39.79**	**0.51**	–	–	–	–	–	–	–	–	–	–	–	–
		B	22.73	4.94	0.83	13.57	137.74	0.42	–	–	–	–	–	–	–	–	–	–	–	–
VGG-11	FC_1	S	**30.69**	**10.65**	**0.75**	**1.21**	**0.61**	**0.56**	**44.48**	**4.68**	**0.91**	**0.26**	**0.06**	0.50	**30.56**	8.36	**0.78**	**1.09**	**0.30**	0.38
		B	24.26	10.77	0.70	2.01	3.18	0.55	28.27	7.80	0.80	0.93	0.08	**0.53**	20.31	**7.28**	0.73	1.41	0.54	**0.44**
	FC_2	S	**36.51**	**10.66**	**0.78**	**5.22**	19.32	0.49	**54.20**	**6.98**	**0.89**	**0.18**	**0.02**	**0.48**	**38.08**	10.34	**0.79**	**0.70**	**0.29**	**0.45**
		B	26.86	10.71	0.72	6.62	**16.21**	**0.54**	29.68	8.64	0.79	1.19	0.52	0.47	20.03	**7.42**	0.72	1.65	1.80	0.36
	FC_3	S	**34.53**	14.21	**0.72**	**4.15**	**4.55**	**0.50**	–	–	–	–	–	–	–	–	–	–	–	–
		B	24.53	**10.95**	0.69	20.66	95.29	0.49	–	–	–	–	–	–	–	–	–	–	–	–
VGG-16	FC_1	S	**43.77**	**8.73**	**0.84**	**0.64**	**0.06**	**0.66**	**56.29**	**3.13**	**0.95**	**0.02**	**0.0001**	**0.47**	**42.26**	11.54	**0.80**	**0.33**	**0.09**	**0.52**
		B	22.50	11.27	0.68	2.38	4.98	0.50	36.06	7.71	0.83	0.40	0.13	0.44	26.87	**8.26**	0.76	1.65	0.61	0.48
	FC_2	S	**36.83**	**11.03**	**0.77**	**0.80**	**0.37**	**0.54**	37.79	**4.31**	**0.90**	**0.17**	**0.02**	0.32	**31.19**	8.70	0.78	**0.83**	**0.45**	0.35
		B	23.31	11.56	0.67	5.43	22.96	0.50	**38.41**	9.66	0.80	0.79	0.52	**0.43**	29.37	**8.04**	**0.78**	2.65	1.90	**0.46**
	FC_3	S	**32.32**	10.21	**0.77**	**6.17**	**32.63**	**0.47**	–	–	–	–	–	–	–	–	–	–	–	–
		B	23.26	**9.97**	0.71	17.53	216.05	0.46	–	–	–	–	–	–	–	–	–	–	–	–
VGG-19	FC_1	S	**40.74**	**10.42**	**0.80**	**0.66**	**0.15**	**0.60**	**46.50**	**2.52**	**0.95**	**0.16**	**0.0002**	**0.39**	**46.38**	14.05	0.77	**0.25**	**0.07**	**0.45**
		B	22.42	11.19	0.67	2.33	3.67	0.47	29.71	5.83	0.84	0.33	0.12	0.39	28.65	**7.93**	**0.78**	1.10	0.80	0.41
	FC_2	S	**40.20**	**9.03**	**0.82**	**1.16**	**0.63**	**0.56**	**50.90**	**5.96**	**0.91**	**0.06**	**0.0006**	**0.37**	**47.03**	13.66	**0.78**	**0.10**	**0.03**	**0.45**
		B	24.00	10.40	0.70	4.64	19.07	0.47	30.31	6.15	0.84	0.45	0.18	0.37	28.46	**8.20**	0.78	2.14	1.92	0.41
	FC_3	S	**28.60**	**6.37**	**0.82**	**4.89**	**11.57**	**0.48**	–	–	–	–	–	–	–	–	–	–	–	–
		B	21.29	7.77	0.74	20.61	143.61	0.46	–	–	–	–	–	–	–	–	–	–	–	–
ResNet-50	FC_1	S	**43.02**	**10.15**	**0.81**	24.43	166.76	0.48	**56.00**	6.50	**0.90**	3.45	**4.74**	**0.45**	**42.54**	10.76	0.80	3.43	19.60	**0.40**
		B	42.15	11.83	0.79	**20.78**	**122.79**	**0.53**	43.80	**5.75**	0.89	**2.73**	6.82	0.36	39.65	**9.81**	**0.81**	**1.64**	**15.20**	0.39
	FC_2	S	**48.58**	**9.75**	**0.83**	37.62	**206.22**	**0.55**	**52.57**	**6.54**	**0.90**	0.25	1.45	**0.40**	**41.03**	12.37	0.77	**1.85**	**13.03**	**0.41**
		B	42.06	11.88	0.79	**29.28**	248.03	0.52	43.63	6.93	0.87	**0.02**	**0.02**	0.35	38.00	**10.00**	**0.80**	2.68	30.91	0.38
	FC_3	S	41.38	11.73	0.77	926.61	142807.00	0.43	–	–	–	–	–	–	–	–	–	–	–	–
		B	**42.03**	**11.48**	**0.79**	**111.18**	**3299.20**	**0.53**	–	–	–	–	–	–	–	–	–	–	–	–
ResNet-101	FC_1	S	**45.93**	**11.14**	**0.81**	**23.32**	**236.76**	0.51	**48.59**	**5.06**	**0.91**	**1.99**	**2.20**	**0.39**	42.54	9.37	0.82	**1.39**	**32.87**	0.35
		B	44.18	12.55	0.78	40.41	828.72	**0.52**	42.94	8.16	0.84	5.41	10.39	0.35	**43.33**	**9.30**	**0.83**	15.28	48.71	**0.39**
	FC_2	S	**51.59**	**9.02**	**0.85**	67.60	**947.85**	**0.54**	**49.27**	**6.39**	**0.89**	**0.98**	**0.65**	**0.37**	**41.71**	9.16	0.82	3.30	100.97	0.38
		B	43.22	12.32	0.78	**43.40**	1155.22	0.50	41.79	7.30	0.85	6.58	17.16	0.34	41.35	**8.32**	**0.84**	**2.26**	**48.61**	**0.39**
	FC_3	S	**47.71**	**10.24**	**0.82**	**73.33**	**2797.15**	**0.53**	–	–	–	–	–	–	–	–	–	–	–	–
		B	42.40	10.53	0.80	162.68	16481.93	0.49	–	–	–	–	–	–	–	–	–	–	–	–
ResNet-152	FC_1	S	44.81	12.09	0.79	**26.35**	**289.59**	0.48	**44.90**	5.63	**0.89**	6.25	**5.86**	**0.36**	**41.09**	**10.09**	**0.81**	**0.33**	**3.59**	**0.39**
		B	**45.62**	**10.68**	**0.81**	36.92	767.58	**0.54**	39.93	**5.40**	0.89	**6.08**	6.74	0.33	40.15	10.82	0.79	0.59	11.39	0.37
	FC_2	S	43.79	**10.04**	**0.81**	**7.13**	**42.77**	0.52	**40.98**	6.90	0.86	**4.64**	**5.71**	0.32	**41.36**	**12.04**	**0.78**	**14.29**	**17.33**	**0.38**
		B	**45.08**	10.85	0.81	44.59	1200.97	**0.52**	40.29	**5.56**	**0.89**	7.86	12.24	**0.33**	38.57	12.07	0.77	18.03	67.52	0.36
	FC_3	S	44.21	11.89	0.79	**47.28**	**1463.55**	0.50	–	–	–	–	–	–	–	–	–	–	–	–
		B	**44.89**	**10.77**	**0.81**	167.41	16331.28	**0.52**	–	–	–	–	–	–	–	–	–	–	–	–