






Feature Fusion Distillation

Chao Tan^{1,2}  and Jie Liu^{1,2}  

¹ Science and Parallel and Distributed Processing Laboratory,
National University of Defense Technology, Changsha 410073, China
liujie@nudt.edu.cn

² Laboratory of Software Engineering for Complex Systems,
National University of Defense Technology, Changsha 410073, China

Abstract. Most recent efforts made in knowledge distillation (KD) can be credited to filling the representation gap between the cumbersome teacher and its light student. In general, the soft targets, the intermediate feature representation in hidden layers, or a couple of them from the teacher serve as the supervisory signal to educate the student. However, previous works aligned hidden layers one on one and cannot make full use of rich context knowledge. To this end, we propose a Feature Fusion Module (FFM) to concatenate diverse feature maps from different layers to aggregate knowledge as the to-be-distilled dark knowledge. Moreover, to hedge the adverse effects of the fused feature maps, we devise an Asymmetric Switch Function (ASF) to make the transfer process more reliable. The combination of FFM and ASF is termed Feature Fusion Distillation (FFD). Experiments of image classification, object detection, and semantic segmentation on individual benchmarks show FFD jointly assist the student in achieving encouraging performance. It is worth mentioning that when the teacher is ResNet34, the ultimately educated student ResNet18 achieves 71.40% top-1 accuracy on ImageNet-1K.

Keywords: Neural network compression · Knowledge distillation · Knowledge transfer

1 Introduction

Hinton *et al.* [13] introduced the concept of knowledge distillation (KD) and explored the teacher-student paradigm for network compression. Following this novel idea, KD based approaches directly train a light network (student), which mimics its original cumbersome network (teacher). However, only transferring soft targets as [13] would limit the performance of output distillation. To make full use of teacher's knowledge, as shown in Fig. 1(a), several approaches [23, 32] transfer teacher knowledge by using hidden layers knowledge. So, this framework is used as the basic framework in this paper. Moreover, KD is widely used for specific applications. For example, prior research has been conducted on face recognition [8, 29], image retrieval [17], cross-modal task [9, 31], neural machine translation [26] and speech recognition [1, 3].

However, as shown in Fig. 1(a), teacher transfers its knowledge one on one in most of the previous approaches. As demonstrated in [34], the projections from each level (low/mid/high) show the hierarchical nature of the features in the network. For example, low-level features show the corners and other edge/color conjunctions, while high-level features are more related to entire objects with significant pose. So, traditional transfer pattern limited the interactivity between different levels.

This drawback motivates us to collect both diverse intermediate features in hidden layers from the teacher as dark knowledge via a *Feature Fusion Module* (FFD). For clarity, FFD is shown in Fig. 1(b), where its way to extract the to-be-distilled dark knowledge is very different from the traditional method. Though more knowledge of the teacher is utilized via the FFD, the empirical analysis shows that the student not only needs to learn negative values of feature maps but also needs to avoid the adverse effects from the negative values of the teacher. So, an *Asymmetric Switch Function* (ASF) is developed to reduce the side effect of dark knowledge due to the large negative activations. The combination of FFM and ASF is termed *Feature Fusion Distillation* (FFD).

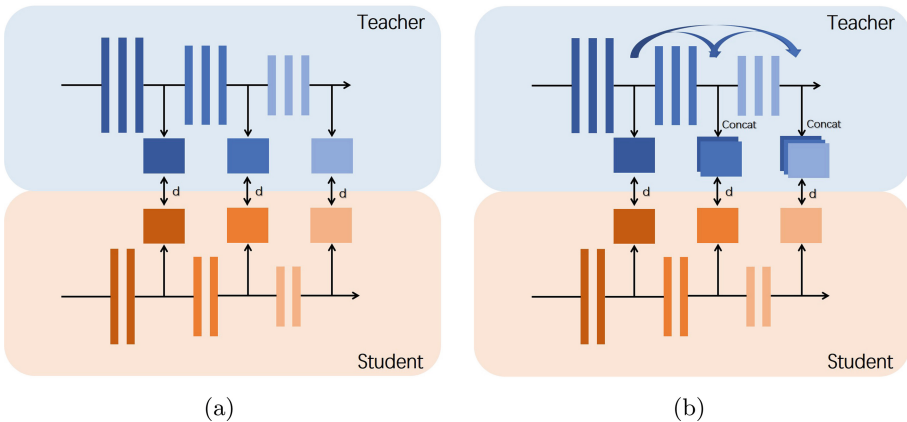


Fig. 1. (a) The general knowledge transfer architecture. (b) The proposed knowledge transfer architecture. Different colors mean the feature maps distilled from different layers.

The contributions in this paper can be summarized as follows:

1. A novel feature fusion module is introduced to make the best use of the teacher's hidden layers knowledge;
2. An asymmetric switch function is proposed to eliminate the harmful knowledge from the teacher;
3. Experiments of image classification on both CIFAR-100 [16] and ImageNet-1K [24], object detection on VOC 2007 [7], and semantic segmentation on VOC 2012 Aug [6] demonstrate the efficacy of FFD.

2 Related Work

Hinton *et al.* [13] introduced the concept of KD for model compression and acceleration, where soft targets of the teacher network are used to educate the student network. Compared with hard targets, soft targets contain the information about inter-class correlations. Consequently, the student network can get relatively better performance. Henceforth, KD becomes an important branch of model compression and acceleration, then amounts of efforts have been made. The differences of such methods primarily lie in two aspects: knowledge representation and transfer skill.

For knowledge representation, the core is to aggregate feature representation from the teacher network as dark knowledge to educate the student. Build off [13], Romero *et al.* [23] utilized intermediate representation in hidden layers of the teacher to make the student better mimic. Yim *et al.* [32] intended to leverage the flow of solution procedure (FSP) matrix between the layers from the teacher as knowledge representation. To educate the student effectively, Kim *et al.* [15] exploited convolutional operations to paraphrase teacher knowledge. Afterwards, Huang *et al.* [14] compared the distributions of neuron selectivity patterns between teacher and student. In [20, 22, 28], they investigated the correlation between multiple instances as the interaction way between teacher and student. Tian *et al.* [27] used contrastive learning to help the student capture the teacher’s knowledge.

For transfer skill, the involved techniques primarily consider how to help the student absorb the teacher’s knowledge to the utmost. In general, the basic skill is to employ the squared loss to evaluate the feature similarity between teacher and student. Heo *et al.* [12]. proposed a knowledge transfer method by transferring activation boundaries of hidden neurons. Afterwards, Heo *et al.* [11] moved the distillation position to the front of the ReLU layer and minimized a new distance function, called partial L_2 distance, to realize knowledge transfer. Yue *et al.* [33] tried to match the teacher’s channels with student’s without convolutional operation. Afterwards, Passalis *et al.* [21] and Chen *et al.* [4] automatically assigned proper target layers of the teacher model for each student layer.

3 Method

3.1 Feature Fusion Module

Zagoruyko and Komodakis [34] pointed out the projections from each level (low/mid/high) show the hierarchical nature of the features in the network. For example, low-level features show the corners and other edge/color conjunctions. Besides, high-level features are more related to entire objects with significant pose. Specifically, we show the Grad-cam++ [2] visualization. As Grad-cam++ visualizes the regions where the network has considered important, in Fig. 2, we compare the results of a model on different levels. It can be observed that low-/mid-level feature maps pay much more attention on the edge of objects. While high-level maps concentrate more on entire objects. Based on this, feature maps

on different levels can be mutually complementary. However, previous works just aligned hidden layers one on one and cannot make full use of rich context knowledge. By this insight, we devise FFM to extract diverse intermediate features as rich knowledge in a simple concatenation manner, which is a necessity for a qualified teacher. In this way, context knowledge can be leveraged.

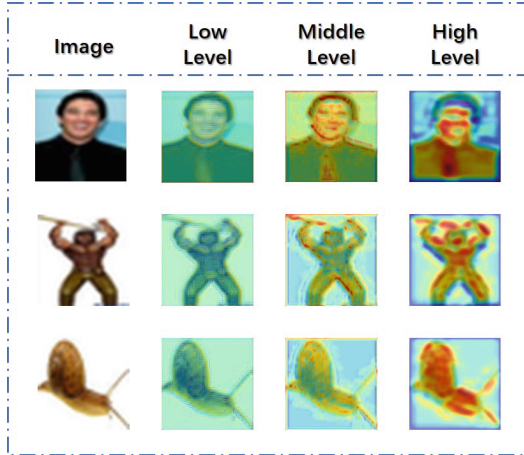


Fig. 2. The Grad-cam++ visualization on different levels.

Following [11], the intermediate features are acquired before ReLU. This distillation position enables the student to touch the preserved knowledge of the teacher before it passes through ReLU. The selected hidden layer in the teacher network outputs the corresponding feature map matrix $\mathbf{F} \in \mathbf{R}^{h \times w \times c}$, where h , w and c represent the height, width and the number of channels, respectively. Using the given notations, FFM can be described as:

$$\begin{aligned}
 \mathbf{F}'_1 &= \psi \{ \eta(\mathbf{F}_1) \}, \\
 \mathbf{F}'_2 &= \psi \{ \eta(\mathbf{F}_1), \eta(\mathbf{F}_2) \}, \\
 &\dots, \\
 \mathbf{F}'_l &= \psi \{ \eta(\mathbf{F}_1), \eta(\mathbf{F}_2), \dots, \eta(\mathbf{F}_l) \}.
 \end{aligned} \tag{1}$$

where l represents the number of those selected hidden layers and could vary from different visual tasks. In our empirical studies, $l = 3$ is suited for image classification on CIFAR-10 [16] and CIFAR-100 [16], while $l = 4$ is better on ImageNet [24]. $\eta(\cdot)$ means the adaptive function of each distilled feature map before they are combined. To fuse the feature maps with different dimensions more simply and efficiently, the AdaptiveAvgPool2d function is used to resize them to the same size. Afterwards, the concatenation acts as $\psi(\cdot)$ to aggregate different layer-wise feature maps altogether. The element-wise summation as another generic alternative to fusing layer-wise features is not considered here.

This is because it requires layer-wise feature maps to be the same channels, which might not be best. As in Eq. (1), \mathbf{F}_l' is the fused feature map, which corresponds to the l -th layer feature maps of the student.

3.2 Asymmetric Switch Function

As there is some harmful knowledge lurked in feature maps, especially in feature maps before ReLU, none of the transformations over the teacher’s feature maps might hurt the KD process. So, we investigate ASF used to transform the teacher’s feature maps to further improve the efficacy of FFM. As FFM transfer the knowledge before ReLU, the switch function should be changed considering ReLU. In the feature maps of the teacher, the positive values are actually used for the network which implies that the positive responses of the teacher should be transferred by their exact values. However, since the negative values are filtered out by ReLU, learning from all the negative values could not always be helpful for the student. However, negative values are not. For these values in the teacher are negative, if the student’s value is higher than the target value, it should be reduced, but if the student’s value is lower than the target value, it does not need to be increased since negatives are equally blocked by ReLU regardless of their values. Furthermore, as mentioned in [12], to transfer the activation boundary accurately, it is required to amplify the negligible values near the activation boundary. So, we propose a switch function to suspend negative values and transfer the activation boundary accurately. The concrete form of ASF is

$$\delta(x) = \begin{cases} n & x < 0 \\ \max(x, m) & x \geq 0, \end{cases} \quad (2)$$

where m is a positive value and n is a negative one. In Fig. 3(a), the activation boundaries are determined by the lines $y = 1$ and $y = -1$. However, the teacher’s knowledge can not be well represented in this rigid setting. In Fig. 3(b), only a negative boundary is fixed [12]. In principle, the positive part also needs the counterpart for activation boundaries. As in Fig. 3(c), ASF shares the merits of the previous two switch functions, thus it not only magnifies the tiny values around zero to transfer activation boundaries but also suspends the adverse values. Note that the predefined parameters m and n are defined as the channel-by-channel expectation of the positive and negative values, respectively. Given that the c -th channel of the teacher’s fused feature map is $F_l^{c'}$, the m^c and n^c of a channel can be calculated from the expectation values of all the training data as follows.

$$\begin{aligned} m^c &= \mathbb{E} [F_l^{c'} | F_l^{c'} \geq 0], \\ n^c &= \mathbb{E} [F_l^{c'} | F_l^{c'} < 0]. \end{aligned} \quad (3)$$

The expectation values can be calculated via the parameters of the batch normalization layer before the distillation position. Appendix A contains the process of calculation. ASF obtains channel-wise margin value without sampling and averaging on training process. As a result, ASF $\delta(\cdot)$ generates the target value as the ultimate knowledge representation to educate the student network.

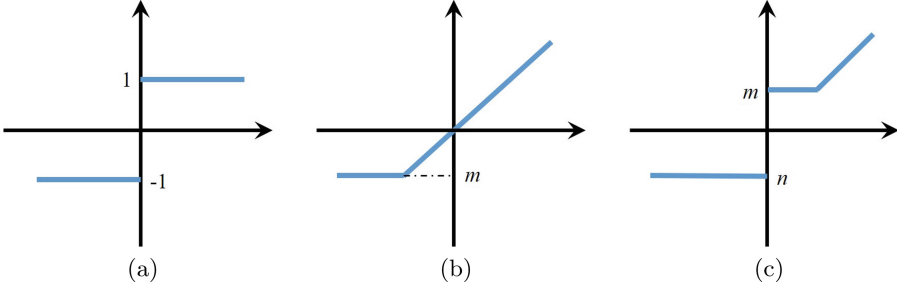


Fig. 3. (a) Switch function of [12]. (b) Switch function of [11]. (c) ASF.

3.3 Total Loss Function

Since the distillation position is before ReLU, the ReLU should be taken into account in designing FFD. For the target value, the positive part should be transferred to the student in an exact way. However, the negative part is not done. If the source value of the student is greater than the negative target value, it should be diminished to amplify the negligible values. But if the source value is less than or equal to the negative target value, it does not need to be added since the negative source value is blocked by the ReLU. Suppose that both source value and target value are \mathbf{T} and $\mathbf{S} \in \mathbf{R}^{h \times w \times c}$, respectively, then the i -th component of \mathbf{T} and \mathbf{S} are T_i and S_i , respectively, and the distance function d is defined as follows.

$$d = \sum_i^{h \times w \times c} \begin{cases} 0 & \text{if } S_i \leq T_i = n, \\ (T_i - S_i)^2 & \text{otherwise.} \end{cases} \quad (4)$$

In terms of FFD, $\eta(\cdot)$ and $\psi(\cdot)$ are used as the adaptive function and FFM. ASF is $\delta(\cdot)$ is a regressor $\gamma(\cdot)$ comprising a convolutional layer and a batch normalization layer to align the student’s feature maps with the teacher’s. Besides, $d(\cdot)$ is the distance function to evaluate the learning effect. Then, the calculating flow of distillation loss is

$$L_{distill} = d(\delta(\psi(\eta(\mathbf{F}_T))), \gamma(\mathbf{F}_S)), \quad (5)$$

where \mathbf{F}_T and \mathbf{F}_S are input feature maps of the teacher and student network, respectively. Yet, several points of the distillation loss are non-differentiable. The non-differentiable points of the function will never appear in practice [36]. L_{task} is task-specific. Consequently, the total loss function can be described as

$$L_{total} = L_{task} + \lambda L_{distill}. \quad (6)$$

4 Experiments

4.1 Image Classification (CIFAR-100)

CIFAR-100 contains 50K training images and 10K test images, both of which have 100 classes. In recent distillation literature, CIFAR-100 is a widely used

benchmark for classification performance evaluation. To make the results convincing, FFD should work well with different network architectures. For training convergence and efficiency, different types of ResNet [10] and WideResNet [35] are chosen in this section. Each experiment is trained around 200 epochs with an initial learning rate of 0.1, which is divided by 10 at epoch 100 and epoch 150, respectively. The hyperparameter λ in Eq. (6) is set to $6e^{-4}$. The top-1 accuracy acts as the evaluation metric. The hyperparameters of the other methods can be referred to [27].

Table 1 report the results of different methods. The results of the KD model with FFD substantially surpass the state-of-the-art counterparts. The margins range from 0.10% to 3.17%, then the average value is 1.39%. Compared with the student’s ‘Baseline’, the average increase reaches 2.94%. Particularly, when both the teacher and the student are built off ResNet110 and ResNet32, respectively, training the student network with FFD can excel the teacher network in many cases. Notably, in two cases the student even outperforms the teacher network. Therefore, FFD can work well to distill either the same architecture network or the different structure network with some expected performance gain.

Table 1. Top-1 accuracy (%) of the student and the corresponding teacher exploited by different KD methods on CIFAR-100 dataset. Of them, ‘Baseline’ represents the result without distillation.

Teacher	WRN40-2	WRN40-2	WRN22-4	WRN22-4	ResNet56	ResNet110	ResNet110
Baseline	75.91	75.91	77.56	77.56	72.98	73.79	73.79
Student	WRN16-2	WRN40-1	WRN10-4	WRN22-2	ResNet20	ResNet20	ResNet32
Baseline	73.26	71.68	71.41	73.92	68.76	68.76	70.79
KD [13]	74.81	73.06	73.52	76.35	71.19	70.56	72.79
FitNet [23]	73.50	72.17	72.93	74.12	69.89	69.50	71.19
AT [34]	73.33	72.23	73.20	74.78	70.30	70.08	72.06
FSP [32]	73.54	n/a	72.87	n/a	69.94	70.08	71.39
CC [22]	73.31	71.98	72.93	74.65	69.98	69.90	71.86
SP [28]	73.69	71.91	72.93	74.87	70.00	69.88	71.65
CO [11]	75.23	73.75	73.97	76.77	70.06	70.33	73.28
CRD [27]	75.56	73.95	74.28	76.97	71.06	70.92	73.62
SSKD [30]	75.80	74.12	74.20	77.02	71.11	71.24	73.85
FFD	75.91	74.45	74.59	77.34	71.29	71.65	73.95

4.2 Image Classification (ImageNet-1K)

To convince the ones of the efficacy of FFD, we further conduct our experiments on ImageNet-1K, which is a larger dataset made up of 1.2M training images and 50K test images. ResNet34 is selected as the teacher network and ResNet18 is the student. The parameters ratio of the student to the teacher is 53.63%. For a fair comparison, we directly use the off-shelf pre-trained model as the teacher

network. The experiment performs 100 epochs in total with an initial learning rate of 0.1, which is gradually reduced by dividing itself by 10 at epoch 30, epoch 60 and epoch 90 in order. The hyperparameter λ is set to $6e^{-5}$. Both top-1 and top-5 accuracy are viewed as the evaluation metrics. The settings of the compared other methods are suggested from [27].

To compare with FFD, we select four other counterparts, which have shown sound performance in Sect. 4.1. According to Table 2, the ‘Baseline’ gap between the teacher and student network is 3.55%, in terms of the top-1 accuracy. FFD can reduce this gap by 1.64%, which is obviously superior to the other methods. So is the top-5 accuracy of FFD.

Table 2. Top-1 and top-5 accuracy (%) of the student network ResNet18 on ImageNet dataset. ‘Baseline’ stands for the results without distillation.

	ResNet34 Baseline	ResNet18 Baseline	KD [13]	CC [22]	SP [28]	CO [11]	SSKD [30]	FFD
Top-1	73.31	69.76	70.70	70.61	70.64	70.96	71.27	71.40
Top-5	91.42	89.08	89.85	89.59	89.70	90.05	90.22	90.42

4.3 Object Detection

In this section, another computer vision task, object detection, is used to further verify the effectiveness of our work. We distill a high-speed detector termed Single Shot Detector (SSD) [18] with FFD. PASCAL VOC 2007 trainval and test dataset are used to train and evaluate the smaller SSD. The backbone of SSD in this section is pre-trained by ImageNet-1K. The starting learning rate is set to $1e^{-3}$ then divided by 0.1 at 80K iterations and 100K iterations, a total of 120K iterations with a batch size of 16. We assign $3e^{-5}$ to the hyperparameter λ .

Table 3 shows the mAP of FFD. ResNet is widely used as the backbone of SSD. The teacher network is based on ResNet50. There are two student networks which are based on ResNet34 and ResNet18. The parameters ratios are 38.21% and 63.76%, respectively. In the case of ResNet34, the mAP of FFD is greater than the ‘Baseline’ and [5] by 2.91% and 0.60%. Particularly, it is even greater than the teacher’s mAP. As for the ResNet18, the mAP increases by 1.45% and 0.51%. Regardless of the compression ratio, the student networks in experiments are enhanced. Obviously, when the student’s capacity is closer to the teacher’s, performance improvement tends to be more significant.

4.4 Semantic Segmentation

Similar to the goal of the above object detection experiments, semantic segmentation is another visual task. We select DeepLabV3+ [7] as our model architecture. For the teacher network, we choose ResNet101 as the backbone. And ResNet50 and MobileNetV2 [25] are used as the student network, respectively.

Table 3. Object detection results of SSD300 on PASCAL VOC 2007 dataset. The mean average precision (mAP) serves to stand for the results. ‘Baseline’ represents the result without distillation.

	Networks	Methods	mAP (%)
Teacher	ResNet50-SSD	Baseline	79.52
Student	ResNet34-SSD	Baseline	76.93
		[5]	79.24
		FFD	79.84
	ResNet18-SSD	Baseline	70.10
		[5]	71.04
		FFD	71.55

All these backbones are pre-trained on ImageNet. FFD is based on the PASCAL VOC 2012 Aug dataset. Each image is cropped to the same size 513×513 , which are largest in any other tasks and more challenging. The learning rate decreases in a polynomial curve. All the compared models are trained around 30K iterations with a batch size of 16. The hyperparameter λ is set to $1e^{-6}$.

Results are displayed in Table 4. The performance is measured in terms of pixel intersection-over-union averaged with 21 classes (mIOU). No matter whether the student share has the same architectural style as the teacher or not, FFD remarkably enhances the performance of the student. Without surprise, the same network architecture can induce better results. No matter in which cases FFD is a generic approach for improving other KD models.

Table 4. Semantic segmentation results of DeepLabV3+ in PASCAL VOC 2012 Aug dataset. The performance is measured in terms of pixel intersection-over-union averaged (mIOU). ‘Baseline’ represents the result without distillation.

	Backbones	Methods	mIOU (%)
Teacher	ResNet110	Baseline	78.33
Student	ResNet50	Baseline	74.11
		[19]	76.12
		FFD	76.92
	MobileNetV2	Baseline	71.13
		[19]	72.61
		FFD	73.24

5 Ablation Study

In this section, the ablation study is conducted to help further understand FFD. The ablation study is conducted by in order adding the layer-wise feature maps

gradually to observe how the aggregated context knowledge affects performance. We select CIFAR-100 and ImageNet-1K for image classification tasks. Three kinds of teacher-student networks are adopted, which are recorded in Table 5. Implementation details and evaluation metrics are introduced in Sect. 4.1 and Sect. 4.2.

Table 5. Experiments settings with several network architectures on CIFAR-100 and ImageNet-1K.

Setup	Dataset	Teacher	Student	Teacher params	Student params	Parameters ratio
(a)	CIFAR-100	WRN22-4	WRN10-4	4.32M	1.22M	28.24%
(b)		WRN22-4	WRN22-2	4.32M	1.09M	25.23%
(c)		WRN22-4	ResNet20	4.32M	0.28M	6.48%
(d)	ImageNet-1K	ResNet34	ResNet18	21.80M	11.69M	53.62%

The results are shown in Table 6. The ‘Baseline’ is not trained as in [23]. It can be observed that considering all the ablation components can improve the performance of the fundamental KD model to different degrees. Thus, FFM and ASF are indispensable roles for KD.

Table 6. Ablation study of FFD. The results are evaluated by the top-1 accuracy (%). The values in bracket denote the improvement by adding a layer-wise feature maps of FFD.

Setup	Baseline	Feature fusion module	Asymmetric switch function
(a)	73.99	74.23(+0.24)	74.59(+0.36)
(b)	76.81	77.09(+0.28)	77.34(+0.25)
(c)	69.00	69.49(+0.49)	70.77(+1.38)
(d)	70.66	71.05(+0.39)	71.40(+0.35)

6 Conclusion

In this paper, we make an improvement for knowledge distillation (KD) by a Feature Fusion Module (FFM) and an Asymmetric Switch Function (ASF) are proposed. The combination of FFM and ASF is termed Feature Fusion Distillation (FFD). FFD is evaluated on three visual tasks including image classification, object detection and semantic segmentation. In particular for image classification, FFD shows its performance superior to the state-of-the-art methods and even better than the teacher model on standard benchmark datasets. Particularly, when the teacher is ResNet34, the top-1 accuracy of the student ResNet18

is greater than that of baseline by 1.64% on ImageNet-1K. Meanwhile, in the case of object detection and semantic segmentation, the efficacy of the educated student significantly excels its baseline. These results imply that FFD including knowledge representation and transfer skill can boost the KD model and also be applied for many fields in practice.

Appendix

A Margin Value

When the feature maps are before ReLU, the batch-norm layer determine the distribution of feature $F_l^{c'}$ in a batch. Batch norm layer normalizes the feature for each channel to a gaussian distribution with a specific mean μ and variance σ . In other words,

$$F_l^{c'} \sim \mathcal{N}(\mu, \sigma). \quad (7)$$

The value of mean and variance (μ, σ) of each channel correspond to the parameters (β, γ) of the batch-norm layer. So, it can be obtained by analyzing the teacher network. Using the distribution of $F_l^{c'}$, we can directly calculate the margin value.

$$m = \frac{1}{Z} \int_0^\infty \frac{x}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx. \quad (8)$$

The expectation can be obtained from integration using pdf of gaussian distribution, where the range is smaller than zero. The result of the integration can be expressed in simple form using the cdf function $\Phi(\cdot)$ of normal distribution.

$$m = \frac{\sigma e^{-\mu^2/2\sigma^2}}{\sqrt{2\pi}\Phi(-\mu/\sigma)}. \quad (9)$$

As $m + n = \mu$, then, n can be calculated as follows.

$$n = \mu - \frac{\sigma e^{-\mu^2/2\sigma^2}}{\sqrt{2\pi}\Phi(-\mu/\sigma)}. \quad (10)$$

References

1. Aguilar, G., Ling, Y., Zhang, Y., Yao, B., Fan, X., Guo, C.: Knowledge distillation from internal representations. In: AAAI, pp. 7350–7357 (2020)
2. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE (2018)
3. Chebotar, Y., Waters, A.: Distilling knowledge from ensembles of neural networks for speech recognition. In: Interspeech, pp. 3439–3443 (2016)
4. Chen, D., Mei, J.P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., Chen, C.: Cross-layer distillation with semantic calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 7028–7036 (2021)

5. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. In: *Advances in Neural Information Processing Systems*, pp. 742–751 (2017)
6. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018*. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
7. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. *Int. J. Comput. Vision* **111**(1), 98–136 (2015)
8. Ge, S., Zhao, S., Li, C., Li, J.: Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Trans. Image Process.* **28**(4), 2051–2062 (2018)
9. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2827–2836 (2016)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
11. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1921–1930 (2019)
12. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3779–3787 (2019)
13. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint [arXiv:1503.02531](https://arxiv.org/abs/1503.02531) (2015)
14. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint [arXiv:1707.01219](https://arxiv.org/abs/1707.01219) (2017)
15. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: network compression via factor transfer. In: *Advances in Neural Information Processing Systems*, pp. 2760–2769 (2018)
16. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images. Technical report (2009)
17. Liu, Q., Xie, L., Wang, H., Yuille, A.L.: Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3662–3671 (2019)
18. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
19. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2604–2613 (2019)
20. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976 (2019)
21. Passalis, N., Tzelepi, M., Tefas, A.: Heterogeneous knowledge distillation using information flow modeling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2339–2348 (2020)
22. Peng, B., et al.: Correlation congruence for knowledge distillation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5007–5016 (2019)

23. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: hints for thin deep nets. arXiv preprint [arXiv:1412.6550](https://arxiv.org/abs/1412.6550) (2014)
24. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**(3), 211–252 (2015)
25. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv 2: inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520 (2018)
26. Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z., Liu, T.Y.: Multilingual neural machine translation with knowledge distillation. arXiv preprint [arXiv:1902.10461](https://arxiv.org/abs/1902.10461) (2019)
27. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint [arXiv:1910.10699](https://arxiv.org/abs/1910.10699) (2019)
28. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1365–1374 (2019)
29. Wang, M., Liu, R., Hajime, N., Narishige, A., Uchida, H., Matsunami, T.: Improved knowledge distillation for training fast low resolution face recognition model. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2019)
30. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12354, pp. 588–604. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_34
31. Ye, H.J., Lu, S., Zhan, D.C.: Distilling cross-task knowledge via relationship matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12396–12405 (2020)
32. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141 (2017)
33. Yue, K., Deng, J., Zhou, F.: Matching guided distillation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12360, pp. 312–328. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_19
34. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. arXiv preprint [arXiv:1612.03928](https://arxiv.org/abs/1612.03928) (2016)
35. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint [arXiv:1605.07146](https://arxiv.org/abs/1605.07146) (2016)
36. Zhou, H., Alvarez, J.M., Porikli, F.: Less is more: towards compact CNNs. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9908, pp. 662–677. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_40