Curriculum-based Teacher Ensemble for Robust Neural Network Distillation

Georgios Panagiotatos[1], Nikolaos Passalis[2], Alexandros Iosifidis[3], Moncef Gabbouj[2], and Anastasios Tefas[1]
[1]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2]Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland
[3]Department of Engineering, Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark
Email: panagiotat@csd.auth.gr, nikolaos.passalis@tuni.fi, ai@eng.au.dk, moncef.gabbouj@tuni.fi, tefas@csd.auth.gr

***Abstract*—Neural network distillation is used for transferring the knowledge from a complex teacher network into a lightweight student network, improving in this way the performance of the student network. However, neural distillation does not always lead to consistent results, with several factors affecting the efficiency of the knowledge distillation process. In this paper it is experimentally demonstrated that the selected teacher can indeed have a significant effect on knowledge transfer. To overcome this limitation, we propose a curriculum-based teacher ensemble that allows for performing robust and efficient knowledge distillation. The proposed method is motivated by the way that humans learn through a curriculum, as well as supported by recent findings that hints to the existence of critical learning periods in neural networks. The effectiveness of the proposed approach, compared to various distillation variants, is demonstrated using three image datasets and different network architectures.**

***Index Terms*—neural network distillation, knowledge transfer, curriculum-based distillation, lightweight deep learning**

I. Introduction

Recent advances in Deep Learning (DL) led to state-of-the-art results in various difficult problems, ranging from computer vision to natural language processing and reinforcement learning [1]. However, DL suffers from an important drawback: DL models are becoming increasingly larger requiring vast amounts of computational power and energy, increasing the cost and limiting the potential applications of DL. Several solutions have been proposed in the literature for tackling this problem, e.g., lightweight architectures [2]–[4], and quantization methods [5], [6] to knowledge transfer approaches [7]–[11]. The latter approaches aim to transfer the knowledge encoded in a large and complex neural network, called *teacher*, into a smaller and more efficient one, called *student*. Being orthogonal to the rest of the said approaches, i.e., knowledge transfer can be almost always applied on top of (or before) other methods and further improve the results, has attracted significant attention from the scientific community, as well as from the industry.

Perhaps the most widely used knowledge transfer method is *neural network distillation*, which is usually used to transfer the knowledge between a student and teacher trained for classification tasks. Neural network distillation works by producing soft labels using the teacher network, after appropriately raising the temperature of the softmax activation. Then, the student network is trained to match these soft labels instead of the original hard labels of the training set. These soft labels contain more information regarding the semantic similarities between the classes and the training samples, as well as regarding the actual way that the teacher network works, having a positive regularization effect and allowing for more efficiently training the student, as originally demonstrated in [7].

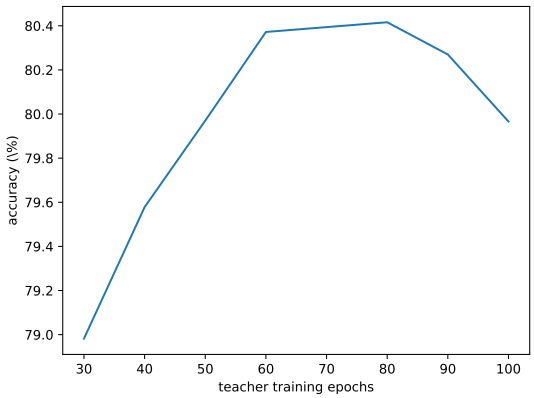

Fig. 1. Student testing accuracy when the knowledge was distilled from teachers which were trained for different number of epochs. Each point in the plot corresponds to training a student for 200 epochs using the corresponding teacher (results smoothed with a moving average filter). The student accuracy increases as further training the teacher, but after a certain point training the teacher starts to have a negative effect on the accuracy of the student. The experimental results are reported using the CIFAR-10 dataset and a student that has more than 16 times less parameters than the teacher. The extent to which this behavior emerges also depends on the complexity of the dataset/classification problem, as demonstrated through this paper.

However, neural network distillation does not always lead to consistent results. In fact, several factors, e.g., teacher model, temperature [12], etc., can severely impact the efficiency of knowledge transfer. We have also experimentally observed that the employed teacher can have an enormous effect on the process of knowledge transfer. This is well understood for less trained teachers, i.e., using under-fitted teachers leads to worse performance. However, we have also observed that training the teachers after a certain point also negatively affected the quality of knowledge transfer, without necessarily over-fitting the teacher. This is illustrated in Fig. 1, where the neural

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

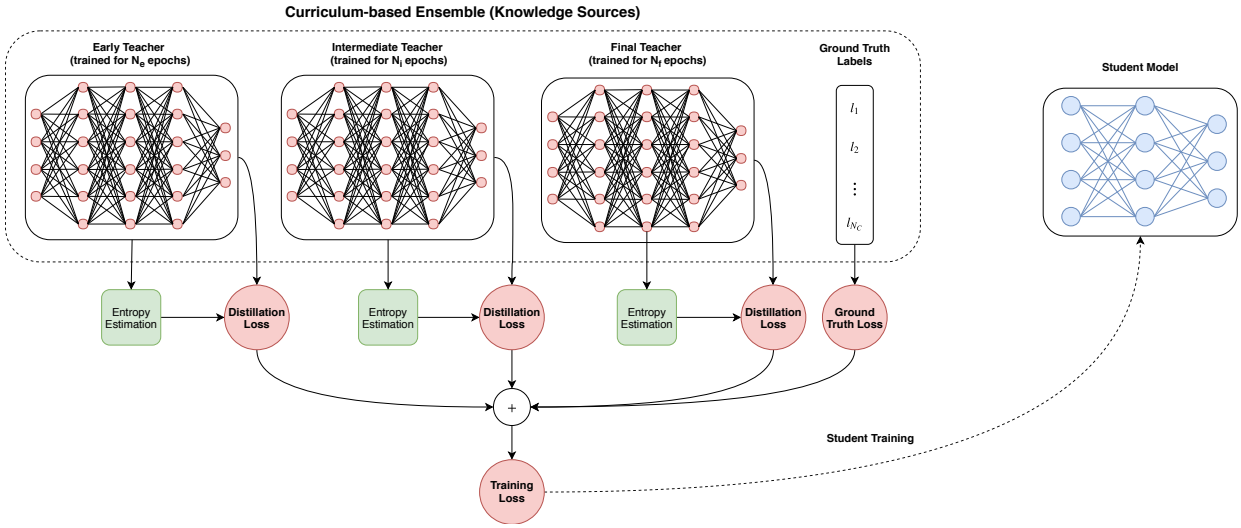

Fig. 2. Proposed Method: Transferring the knowledge from multiple snapshots of the same teacher, trained for N_w, N_i and N_f epochs, to a lightweight student model.

network distillation approach has been applied to transfer the knowledge from teachers that have been trained for different number of epochs. Note that this figure is different from the regular learning curve plots that plot the accuracy of a network vs. training epochs. In this plot the network (student) was trained for a fixed number of epochs and we varied the number of epochs for which the teacher was trained.

This behavior can be also explained if we draw an analogy with the way that humans learn through curriculum-based learning [13]. First, less specialized teachers are used to teach us the basic concepts, e.g., elementary and high-school teachers, and then more specialized teachers, e.g., university professors, are employed to teach us more sophisticated concepts and provide more fine-grained information. It is worth noting that a completely different teaching style is required through these stages and thus different teaching material and teaching approaches are used through our educational system. For example, a university professor would be probably less efficient at teaching elementary math to elementary school students than an elementary school teacher specifically trained for this task. Therefore, using a well-trained and very powerful network at the early stages of student training is probably not the optimal training strategy, as shown in Fig. 1 and validated through the conducted experiments in this paper. Note that this observation is also supported by recent evidence that hints to the existence of critical learning periods in neural networks [14]. Therefore, different training approaches, that are carefully adapted to the needs of each learning stage, should be used through the learning process.

Inspired by the aforementioned observations in this paper we propose a curriculum-based teacher ensemble, which employs different versions of the same teacher, as acquired through the training process, to teach the less capable student model. The capacity of each teacher model is measured through an information-theoretic measure, the entropy, that intuitively measures how certain each teacher is for its decisions and acts as a naive proxy for estimating the teaching capacity of each teacher. Therefore, less trained teachers exhibit a higher degree of entropy (uncertainty) compared to more trained teachers. Then, an ensemble of teachers, weighted by the uncertainty of each teacher, is used for training the students. This allows for training the student model using an ensemble that takes into account the teaching capacity of each teacher, as shown in Fig. 2, improving the classification accuracy compared to directly using a powerful teacher for the training, as analytically demonstrated through the conducted experiments.

The main contribution of this paper is a two-phase knowledge distillation method that employs different snapshots of the same teacher, allowing for overcoming some important limitations of existing distillation approaches. During both training phases a weighted ensemble of different teachers is employed, as described above, for the training process. At the same time, during the first training phase, where the student is less capable of producing correct predictions, the original hard labels are also employed to further steer the training process, as originally proposed in [7], while in the second phase the process continues with a larger emphasis on the distillation from the proposed teacher ensemble. To the best of our knowledge, this is the first multi-teacher distillation approach that takes into account that the same teacher exhibits different capacity for teaching a less capable student, allowing for performing robust neural network distillation. The proposed method is evaluated using three well-known image datasets and several network architectures.

The rest of the paper is structured as follows. First, the proposed method is presented in detail in Section II. Next, the experimental evaluation is provided in Section III. Finally, conclusions are drawn and future research directions are discussed in Section IV.

II. Proposed Method

Let $\mathbf{y}^{t,l} = f(\mathbf{W}^{t,l}, \mathbf{x})$ denote the output of the teacher model $f(\cdot)$ before applying the softmax activation, where $\mathbf{x}$ is a input sample and $\mathbf{W}^{t,l}$ are the weights (parameters) of the teacher after the l-th training epoch. Also, let $\mathbf{y}^s = g(\mathbf{W}^s, \mathbf{x})$ denote the output of the student model $g(\cdot)$ before applying the softmax activation, where $\mathbf{W}^s$ are the weights of the student model. Neural network distillation aims to transfer the knowledge encoded in the teacher model into a less powerful student model. To this end, the teacher model is trained using a *transfer set* to minimize the following loss (assuming that the teacher was trained for l epochs):

$$\mathcal{L}_d^l = -\sum_{i=1}^{N_C} [\mathbf{q}^{t,l}]_i \log[\mathbf{q}^s]_i, \tag{1}$$

where N_C is the number of classes and the soft labels are acquired using a temperature of T:

$$[\mathbf{q}^{t,l}]_i = \frac{\exp([\mathbf{y}^{t,l}]_i/T)}{\sum_{j=1}^{N_C} \exp([\mathbf{y}^{t,l}]_j/T)}, \tag{2}$$

and the notation $[\mathbf{x}]_i$ is used to refer to the i-th element of the vector $\mathbf{x}$. The output of the student model $\mathbf{q}^s$ is similarly defined. The transfer set can be the original training set or a dataset that contains relevant data. Note that no labels are required for the transfer set, except from when hard labels are also used during the training process.

In traditional neural network distillation a fully trained teacher model is used, i.e., $l = N$, where N is the total number of training epochs used for the teacher. However, as previously discussed and briefly demonstrated in Fig. 1, using the last state of the teacher network is not always optimal for training the student model. Therefore, in this paper we propose to use multiple intermediate states of the teacher network, instead of just using the last snapshot of the teacher model. To this end, we propose using a curriculum-based teacher ensemble to define the distillation loss as:

$$\mathcal{L} = \sum_{l \in \mathcal{K}} \alpha_l \mathcal{L}_d^l + \alpha_{labels} \mathcal{L}_{labels}, \tag{3}$$

where the set $\mathcal{K}$ contains the epochs used for forming the teacher ensemble that will be used for the neural network distillation, α_l are the weights used for each teacher snapshot in the ensemble, while $\mathcal{L}^{labels}$ is the typical cross-entropy loss from the hard labels and a_{labels} is the corresponding weight. Altering the parameters α_l and α_{labels} allows for changing the dynamics of the distillation process. For example, setting $\alpha_l = \frac{1}{|\mathcal{K}|}$ leads to multi-teacher neural network distillation, where each teacher is equally important during the knowledge transfer.

However, in this work, different weights α_l are assigned to the teachers in order to form a curriculum-based ensemble, where the less "sophisticated" teachers will be allowed to guide the training process allowing for more efficiently transferring the knowledge to the less complex student network. Perhaps the easiest approach to estimate the *confidence* level of the teacher models is to assume that teachers that were trained for less epochs would be less confident and should be given a higher weight. Even though this is - to some extent - true, it does not provide an efficient way for calculating the appropriate values for α_l. To overcome this limitation, in this paper a simple information-theoretic measure, the Shannon's entropy, is employed to estimate the *confidence level* of the l-th teacher as follows:

$$H_l = -\sum_{i=1}^{N_C} [\mathbf{q}^{t,l}]_i \log[\mathbf{q}^{t,l}]_i. \tag{4}$$

The entropy of each model is estimated as the average entropy over the transfer set. Higher values of entropy indicate that the teacher is generally less confident and is able to capture more rough concepts, that would be easier to transfer to the student model, compared to more confident and sophisticated teachers that are able to perfectly discriminate the classes of the input object. Therefore, the parameters α_l should be defined according to this observation, i.e., the weight given to a teacher should be inversely correlated with its normalized confidence (taking into account the rest of the teachers of the ensemble):

$$\frac{(H_l)^\alpha}{\sum_{l' \in \mathcal{K}} (H_{l'})^\alpha}, \tag{5}$$

where the hyper-parameter α controls the fuzziness of this process, i.e., α allows for suppressing or increasing the differences between the teachers. The hyper-parameter α is typically set to 1, unless otherwise stated.

Apart from the hyper-parameter α, we also define the hyper-parameter $0 \leq \beta < 1$ that controls the weight of the regular cross-entropy loss (between the output of the teacher and the hard labels). Therefore, the parameters used in (3) are defined with respect to only two other hyper-parameter (α and β) as:

$$\alpha_l = (1-\beta)\frac{(H_l)^\alpha}{\sum_{l' \in \mathcal{K}} (H_{l'})^\alpha}, \tag{6}$$

and

$$\alpha_{labels} = \beta, \tag{7}$$

allowing for controlling the various distillation phases of the proposed method just by altering the two hyper-parameters α and β. The proposed distillation approach is summarized in Fig. 2, where an ensemble with three different teacher models trained for three different number of epochs is used, i.e., $\mathcal{K} = \{N_e, N_i, N_f\}$.

Two-stage training: Apart from defining and using curriculum-based teacher ensembles for the neural network distillation, we also propose employing a two-stage distillation process, inspired by the observations reported in [14]. Therefore, in the initial training stage a higher weight is given to the cross entropy loss from the hard labels (higher β value) ensuring that the critical weight connections will form during the initial critical learning period. After this stage is completed, the value of β is reduced and the distillation process continues setting as main objective to transfer the knowledge from the curriculum-based teacher ensemble. For all the experiments

TABLE I
TEACHER AND STUDENT MODELS USED FOR KNOWLEDGE DISTILLATION FOR VARIOUS DATASETS

Model	Dataset	Test acc.	# Parameters
Teacher	MNIST	99.11%	844k
Student	MNIST	97.24%	14k
Teacher	Fashion MNIST	92.45%	890k
Student	Fashion MNIST	91.35%	212k
Teacher	CIFAR-10	84.80%	1,250k
Student	CIFAR-10	78.50%	303k

TABLE II
MNIST EVALUATION (THE CLASSIFICATION ACCURACY (%) IS REPORTED)

Model	Test acc. (final)	Test acc. (avg)
Teacher 1	97.39	97.32
Teacher 2	97.35	97.39
Teacher 3	97.44	97.39
Teacher 1+2+3	97.05	97.34
Proposed (single stage)	**97.48**	97.37
Proposed	97.47	**97.44**

TABLE III
FASHION MNIST EVALUATION (THE CLASSIFICATION ACCURACY (%) IS REPORTED)

Model	Test acc. (final)	Test acc. (avg)
Teacher 1	91.30	91.19
Teacher 2	90.90	91.02
Teacher 3	90.74	90.70
Teacher 1+2+3	91.31	91.33
Proposed (single stage)	91.46	91.46
Proposed	**91.60**	**91.60**

TABLE IV
CIFAR-10 EVALUATION (THE CLASSIFICATION ACCURACY (%) IS REPORTED)

Model	Test acc. (final)	Test acc. (avg)
Teacher 1	81.68	81.76
Teacher 2	81.94	81.94
Teacher 3	81.03	81.26
Teacher 1+2+3	82.79	82.50
Proposed (single stage)	82.81	**82.64**
Proposed	**83.21**	**82.64**

conducted in this paper a value of $\beta = 0.3$ was used during the first training stage, which was then reduced to $\beta = 0.1$ during the second training stage. Note that the aforementioned process can be also extended to have multiple stages allowing for greater granularity, while also different ensembles $\mathcal{K}$ can be used for these stages. In this work, three different teachers were used to form the ensemble and two stages were used. This setup worked quite well for all the conducted experiments, while being relatively straightforward to implement.

III. EXPERIMENTAL EVALUATION

Three image classification datasets were used for evaluating the proposed method: a) the MNIST dataset [15], which contains 60,000 training and 10,000 testing images of handwritten digits, b) the Fashion-MNIST dataset [16], which is a more challenging extension of the MNIST dataset containing 60,000 training images and 10,000 testing images of fashion-related objects, e.g., trousers, t-shirts, etc., instead of handwritten digits, and c) the CIFAR-10 dataset [17], which is composed of 60,000 32×32 color images (50,000 training images and 10,000 testing images) that belongs to 10 different categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. The training set was used for training the teacher networks and transferring the knowledge to the student network (transfer set), while the test set was used to evaluate the trained student models.

The teacher network used for the MNIST dataset was composed of a convolutional layer with 32 3×3 filters, followed by a 2×2 max pooling layer, another convolutional layer with 64 3×3 filters, a 2×2 max pooling layer, and two fully connected layers with 512 and 10 neurons respectively. The ReLU activation function was used for all the layers, except from the final one where the softmax activation was used. Dropout with rate 0.5 was used during the training before and after the first fully connected layer. A lightweight student network was defined by reducing the number of convolutional filter to 4 and 8 (respectively) and the number of neurons in the first fully connected layer to 64.

For the Fashion MNIST and CIFAR-10 the teacher network was composed of two convolutional layers with 32 3×3 filters, followed by a 2×2 max pooling layer (dropout with rate 0.25 was also used during the training), additional two convolutional layers with 64 3×3 filters, followed by a 2×2 max pooling layer (dropout with rate 0.25 was used), and two fully connected layers with 512 and 10 neurons. The student network was composed of a convolutional layer with 16 3×3 filters, followed by a 2×2 max pooling layer, another convolutional layer with 32 3×3 filters and two fully connected layers with 256 and 10 neurons. Again, the ReLU activation function was used for all the layers, except from the final one.

The test accuracy and number of parameters of the teacher and student networks are summarized in Table I. The employed student models are 10 to 60 times smaller than the teacher models. This also has a significant effect on the classification accuracy, which is reduced in all cases, highlighting the need for methods that are able to efficiently train lightweight network architectures.

All the teacher models were trained for 200 (250 for the CIFAR-10 dataset) epochs using the Adam algorithm (the learning rate was set to 0.01 for all the conducted experiments) [18]. Two stages, each composed of 100 training epochs, were used for transferring the knowledge to the student networks. The hyper-parameter α was set to 1 for the MNIST and Fashion MNIST datasets and to 2 for the CIFAR-10 dataset, since for this dataset the entropy values were closer together (increasing the α hyper-parameter allowed for

increasing the differences between the values of α_l). The set $\mathcal{K} = \{N_e, N_i, N_f\}$ contains the epochs that correspond to the teacher models from the $N_e = 40$, $N_i = 90$ and $N_f = 120$ epochs for the MNIST dataset, from the $N_e = 40$, $N_i = 90$ and $N_f = 140$ epochs for the Fashion MNIST dataset and the $N_e = 150$, $N_i = 200$ and $N_f = 250$ epochs for the CIFAR-10 dataset. For all the conducted distillation experiments the temperature was set to $T = 1$, following the recent findings of [12], in which was reported that $T = 1$ usually works better for more complex tasks.

The experimental results using the MNIST, Fashion MNIST and CIFAR-10 datasets are reported in Tables II, III, and IV respectively. For all the datasets we report the average test accuracy (during the last 10 training epochs) and the final test accuracy for: a) students trained using the three different teachers using the distillation process ("Teacher 1", "Teacher 2", and "Teacher 3" respectively, no hard labels used during the training), b) using the average output of the three teachers for simultaneously distilling the knowledge from multiple teachers into one student ("Teacher 1+2+3"), c) using only the first stage of the proposed method for 200 epochs without reducing the β to 0.1, but keeping constant β = 0.25 for the whole process ("Proposed. (single stage)"), and d) using the proposed two stage training process ("Proposed").

Note that the effectiveness of the distillation process increases as more challenging datasets are used. Therefore, the distillation process only leads to marginal improvements for the easier MNIST dataset, while the increase for the more challenging Fashion MNIST and CIFAR-10 datasets are larger. Several conclusions can be drawn from the reported results. First, using the distillation process indeed improves the classification accuracy over directly training with the hard labels (dataset annotations), as reported in Table I. Also, note that the efficiency of the distillation process drops for the Fashion MNIST and CIFAR-10 datasets when the teacher models are trained after a certain point, confirming the initial hypothesis, as demonstrated in Fig. 1. Using the average output of the three teachers (knowledge transfer method using a multi-teacher ensemble) does not seem to be capable of increasing the distillation efficiency for the MNIST dataset, while only leading to marginal improvements for the Fashion MNIST dataset. On the other hand, using the proposed two stage distillation method always improves the distillation results. It is worth noting that the proposed approach leads to an improvement of more than 5% over a student trained with hard labels for the CIFAR-10 dataset. Finally, also note that the second stage of the proposed approach, where the knowledge is actually transferred in an almost fully unsupervised way ($\beta = 0.1$) from the entropy-weighted teacher ensemble, greatly contributes to the effectiveness of the proposed approach, since removing this state, i.e., using the "Proposed (single stage)" method, almost always reduces the distillation efficiency.

IV. Conclusions

In this paper we proposed a two-phase knowledge distillation method which employs different snapshots of the same teacher to improve the efficiency and robustness of knowledge distillation. To this end, a weighted ensemble of different teachers is employed, allowing for overcoming the limitations of existing distillation approaches. Also, during the first training phase the original hard labels were employed to steer the training process during the initial critical learning period of the student model, while in the second phase the process continues with a larger emphasis on the distillation from the proposed teacher ensemble. The efficiency of the proposed approach was experimentally confirmed using three well-known image datasets and several network architectures.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436, 2015.

[2] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[3] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[4] Nikolaos Passalis and Anastasios Tefas, "Training lightweight deep convolutional neural networks using bag-of-features pooling," *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[5] Song Han, Huizi Mao, and William J Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[6] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning Workshop*, 2014.

[8] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," *Proceedings of the International Conference on Learning Representations*, 2015.

[9] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133–4141.

[10] Nikolaos Passalis and Anastasios Tefas, "Unsupervised knowledge transfer using similarity embeddings," *IEEE Transactions on Neural Networks and Learning Systems*.

[11] Nikolaos Passalis and Anastasios Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 268–284.

[12] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 742–751.

[13] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *Proceedings of the International Conference on Machine Learning*, 2009, pp. 41–48.

[14] Alessandro Achille, Matteo Rovere, and Stefano Soatto, "Critical learning periods in deep neural networks," *arXiv preprint arXiv:1711.08856*, 2017.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[17] Alex Krizhevsky, "Learning multiple layers of features from tiny images," *Technical Report*, 2009.

[18] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference for Learning Representations*, 2015.