A survey of Knowledge Distillation

Trung Dao
VinAI Research
14 Thuy Khue, Tay Ho, Ha Noi
v.trungdt21@vinai.io

Abstract

For the last few years, deep learning has proved its ability to solve complicated problems that were too cumbersome to handle. It is undeniable that key factors contributing to this sucess are mainly large-scale data and deep neural networks with billions of parameters. However, to deploy and smoothly run these enormous models on on-edge devices is not always a straightforward process. In order to overcome this, a variety of model compression and acceleration techniques have been developed. As one of the outstand techniques, knowledge distillation (KD) seeks to transfer information learned from a large model (called teacher) to smaller one (called student). This survey firstly gives a brief overview of what KD is then quickly investigates some of the recent works around logits-based and feature-based distillation, then finally review several regularization techniques.

1. Introduction

After achieving extraordinary establishment in various fields of Machine Learning using neural networks, scientists began to move interests into network compression and enhancement. Several worth-mentioning approaches which seek to create smaller model and cost-efficient are parameter pruning/sharing, model quantization, low-rank factorization, and knowledge distillation. A comprehensive overview of these approaches is outside of the scope of this survey, and the focus of this paper is solely about knowledge distillation, which has attracted much attention from the deep learning research community in recent years.

Comparing to other compression methods, KD might be considered superior since it can squeeze down a network disregarding the structural difference between the teacher and student network. The original idea of KD was proposed in [4], where the author focused on transferring the information from a large model or an ensemble of models into a smaller model without significant performance loss. This work is later generalized and popularized by Hinton et al. [15]. In their work, they found out an astounding observation that it is simpler to train a smaller-scale network (classifier) using the soften predictions of another classifier as target value rather than the ground truth (one-hot) labels and later on called this procedure as *distillation* (short for knowledge distillation). They also stated that these probabilities offer richer information than labels alone, and empirically help the student network learn better. Knowledge distillation field since then has been developed and applied in various forms, but the main characteristic of KD has remained the same: Teacher-Student framework, where the teacher model provides useful *knowledge* to the student model to improve its learning performance.

Thanks to its effectiveness on neural network compression and acceleration, KD has been widely applied in different fields of artificial intelligence: visual/speech recognition, natural language processing (NLP). In visual recognition, KD was firstly applied on classification tasks [15, 19, 5, 3, 8] then later on expanded to other applications such as image segmentation [12], lane detection [16], facial landmark detection [39, 11]. In other fields such as NLP, KD also proves its worth as being used to compress complex structures such as BERT [29, 18]. To briefly generalized, KD offers not only lightweight deep models, which allows deploying models to on-edge devices efficiently but recently also competitive performant ones.

The main contributions of this paper consist of:

1. Provide a comprehensive overview of Knowledge Distillation: problem definition, type of knowledge, a family of KD methods with deep learning.

2. Investigate how the research community try to theoretically explain KD.

3. Review several applicable regularization techniques used in distillation process.

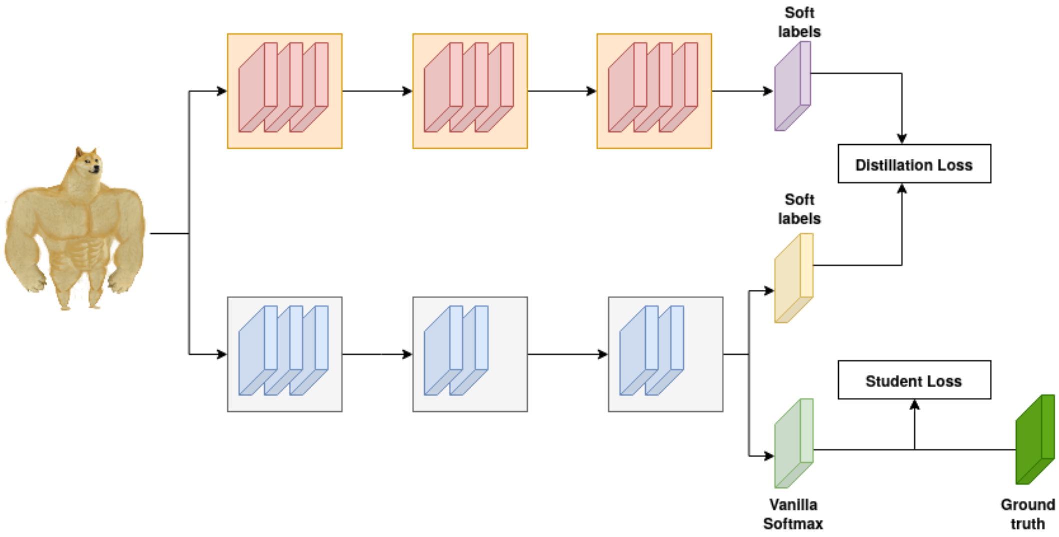

Figure 1. The typical architecture of logits-based knowledge distillation framework.

2. Background: Knowledge Distillation

2.1. Problem statement

Neural network models have been successful in a myriad of fields including extremely complex problems. However, these models are enormous in size and computational hungry, thus cannot be deployed to on-edge devices nor feasibly runnable in some situations. To compensate for this situation, knowledge distillation was first proposed in [4] and properly formalize in [15]. In the work of Bucilu et al. [4], the knowlegde is transferred from the larger model to the smaller model while minimizing the logits difference produced by those two models respectively. This vanilla method opened the first path of knowledge type that can be used to distilled, later on, there started to have some other works using the activations, feature maps of intermediate layers as the guide for the student network [27, 37, 13]. Some other methods to extract knowledge by comparing the relationship between different layers of teacher model [35, 38] are also worth mentioning but won't be discussed in this survey.

2.2. Knowledge

2.2.1 Knowledge from logits

This type of knowledge extraction usually refers to using the neural response of the final output layer, or also known as *logits* of the teacher model. As stated before, the first application of logits in KD is used in [4], but in many situations, given a highly confident teacher model, the output of softmax function gives more or less the same information as the ground truth label (since this is the core idea of softmax function - maximizing the probabilities of one while minimizing the others). Tackling this problem, Hinton et al. [15] proposed the concept of 'soft labels' and declared that this type of label contains informative *dark knowledge*. Given the **logits** z from a network, the 'soft label' p_i of an image is defined as:

$$p_i = \frac{\exp(\frac{z_i}{\rho})}{\sum_j \exp(\frac{z_i}{\rho})}$$

where ρ is the temperature parameter, notice that when $\rho = 1$, we get the normal softmax function. Determine which value is optimal for ρ is still a debatable topics but it is argued that while increasing ρ, the label becomes softer and providing more information about which class is similar to the predicted label. Accordingly, the logits-based distillation loss function is defined as follow:

$$\mathcal{L}_{\text{LB}}(x;\theta) = \alpha * \mathcal{H}(y, \sigma(z_s)) + \beta * \mathcal{H}(\sigma(z_t;\rho), \sigma(z_s;\rho))$$

where x is the input, θ are the student model weights, $\mathcal{H}(.)$ is the cross-entropy loss function, y is the ground truth label, $\sigma(.)$ is the ρ-parameterized softmax function, α, β are the coefficients to tune in order to balance between both cross-entropies, and z_t, z_s are the logits of teacher and student model respectively. Figure 1 illustrate an usual framework while using this type of knowledge.

Logits-based knowledge is fairly straightforward and easy to implement so it is often the first approach used when need to apply knowledge distillation. As far as the survey goes, there are several main motives of using logits-based knowledge: 1) using soft labels with regularization 2) create/use noisy data to train 3) enhancement. A brief table containing its descriptions and several related work can be found at 2.1. Nevertheless, since this type of knowledge only utilizes the final output of the teacher model, it fails to

Paper	Main motives	Description	Year
[33]	Enhance distilling process	Train in generation, control the strictness while training teacher network (adding extra loss term)	2018
[10]	Enhance distilling process	Propose Stage-by-Stage Knowledge Distillation: 2 separated parts: backbone (mimics the output feature of teacher progressively) and task-head (only use ground-truth label).	2018
[34]	Enhance distilling process Novel distillation framework	Propose distillation framework allowing teacher and student having the same architecture, the framework also use the same architecture for both model since it use model from the previous iteration.	2018
[25]	Enhance distilling process	Proposed new loss function while extracting relational information using a relational potential function	2019
[36]	Learn from noisy data Novel distillation framework	Propose Teacher-free distillation framework: including self-training and self-regularization	2020
[23]	Enhance distilling process	Allow student to selectively choose to learn from either the teacher model or the ground truth conditioned on whether the teacher can correctly predict the truth rather than using interpolated labels	2019
[7]	Enhance distilling process	Using early stopping as a regularizer	2019
[24]	Enhance distilling process	In order to minimize the complexity gap, proposed a multi-step distillation.	2019
[22]	Ensemble of distribution	Distill the distribution of the predictions from an ensemble so that student can receive both the enhanced prediction and diversity	2020
[2]	Add Regularization Ensemble of distribution Learn from noisy data	Injecting different type of noises during the distillation process: 1) Dropout while distilling 2) training student with noisy data while teacher had been trained on clean data 3) Randomly change true labels to random target by an uniform distribution	2020
[32]	Enhance distilling process Learn from noisy data	Generate pseudo labels for unlabeled images then train student model on those while injecting noises: dropout, data augmentation	2020
[31]	Add Regularization Enhance distilling process	1) Minimizing genetic error: Using Label Smoothing Regularizer and propose new loss which tries to adjust the teacher's prediction w.r.t. ground truth labels 2) Dynamic temperature distillation: tuning τ sample	2020
[30]	Add Regularization Enhance distilling process	Apply margin-based softmax and normalize input vector/ weight matrix so that student model con focuses on samples with more confident predictions	2020

offer information gathered in the intermediate layers, which later was proved to contain many informative results [27]. Other than that, due to its characteristic, logits-based knowledge is bounded with the supervised learning.

Paper	Φ_t	Φ_s	**Layer for Distillation**	**Distance metrics**	**Knowledge lost risk**	**Year**
[27]	Identity	1D Conv	Arbitrary middle layer	L_1	None	2015
[37]	Attention map	Attention map	End of layer block	L_p	Yes (depth-wise)	2017
[14]	Activation gate	1D Conv	Before activation	Margin L_2	Teacher class distribution	2018
[28]	Adaptive pooling	Adaptive pooling	End of layer block	L_1/L_2/KL/L_{GAN}	Yes (spatial due to adaptive pooling)	2019
[13]	Margin ReLU	1D Conv	End of layer block	L_2	None	2019
[9]	Identity	1D Conv	End of layer block	L_2	None	2020

2.2.2 Knowledge from intermediate layers

What's extraordinary about deep neural networks is how they extract and abstractly represent features through the feature maps generated between each layer. Hence, feature-based knowledge would theoretically contain richer information than logits-based knowledge. Not only so, by combining both intermediate and last layer's output, we can consider feature-based knowledge is an extension of the previous knowledge type.

Romero et al. [27] introduced the term *hint* as the outputs of a teacher hidden layer, which are used to guide the student learning process. The core idea is to choose some intermediate layers in both teacher and student layer and force the student to mimics the result of the teacher's feature maps. Motivated by this work, there has been many studies to investigate which hint layer/ guided layer to choose and how to measure the distance between them. The feature-based distillation loss function is often defined as follow:

$$\mathcal{L}_{FB}(f_t, f_s) = \mathcal{D}(\Phi_t(f_t), \Phi_s(f_s))$$

where f_t, f_s are the selected hint and guided layers, Φ_t, Φ_s are the transformation functions for the mentioned layers respectively, which often are applied when the result of two layers are mismatched, $\mathcal{D}$ is the distance function measuring the similarity between the hint and guided layer. L1, L2 is often used for this distance function, but for some works, more complex function might be used such as the Maximum Mean Discrepancy (MMD) metric [17] or KullbackLeibler divergence [1]. Figure 2 describe the standard feature-based distillation progress.

Eventhough this knowledge type offers more favorable information for the student learning process, where to pick the hint/ guided layers or the transformation functions still need more investigations. For examples, the teacher's layer transformation function are directly critical for this process since there are risks of losing information in the process of transforming. Several past work has already encountered this problem since the feature map's dimension got downscaled, causing loss of knowledge [37, 28]. To counter this situation, one could either come up with novel transformation function such as margin RELU[13] or not use any transformation at all [27]. But as investigated by Heo et al. [13], the hints consist of both beneficial and adverse information so we should try to only gather the positive ones rather than distilling all of it. The same goes for the guided layers, sometimes researchers use the same transformation [37], which might lead to the same information loss. Some other works use 1×1 convolutional layer as a student transform [27, 13] in order to match the depth dimension with the hints, so no information would be lossed during the process. A more detail of progressive work using feature-based knowledge can be found at figure 2.2.1.

3. Explaining Knowledge Distillation

As discussed in the previous sections, knowledge distillation has been successfully applied in various fields and applications. However, clarying how and why KD works in general and its models are usually superior than those trained from raw data still remains an work in progress. Not only so, there is no universally accepted hypothesis as to how knowledge is transmitted, making it impossible to accurately test emperically scientific results and to design new approaches in a more structured fashion. Most of past works seldom go beyond qualitative assertions, such as suggesting that learning from soft labels may be better than learning from hard labels, or that the teacher's performance in a multi-class environment offers details on how close different classes are to one another [4, 15, 13].

Recently, some scientific researchers start to focus on finding out what are the real factors making KD works. In the case of deep linear classifiers, Phuong et al. [26] came up with a theoretical explaination for a generalization bound for fast convergence of learning distilled student networks. This justification clarifies what the student would learn and how quickly they do it, as well as the factors that influence distillation performance, to be more precise, the authors stated there are three main factors: data geometry, optimization bias and srtong monotocity of the student classifier.

This survey is going to focus on another work, where the author - Chen et al. quantified the distilled knowledge using the term visual concepts from the intermediate lay-

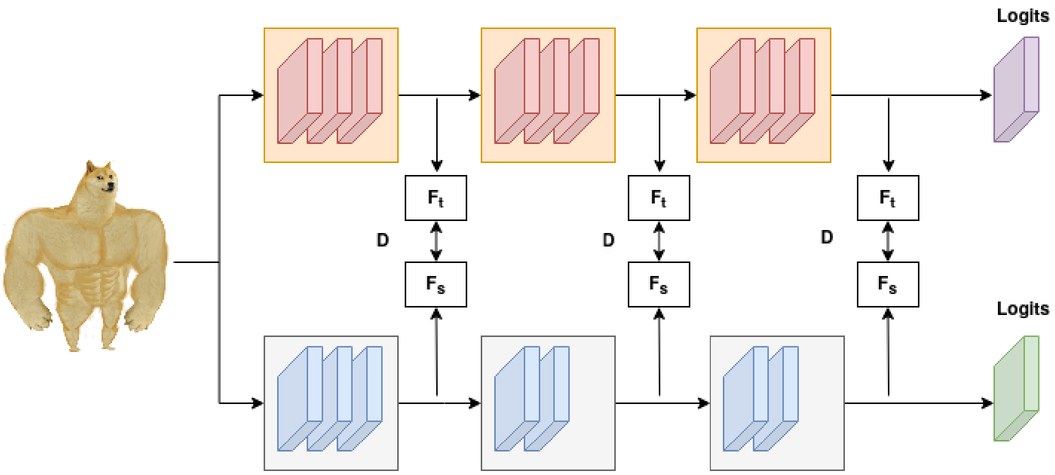

Figure 2. The typical architecture of feature-based knowledge distillation framework.

ers of a deep neural network (DNN) in order to explain KD [6]. What separates this paper from other work is that the author tried to interpret KD from a different perspective of information theory by quantifying, analyzing and comparing encoded information in the intermediate layers between DNNs learned by KD and normal DNNs, which not only investigate the success of KD but also what DNNs have learned during training in order to perform effective inference.

3.1. Setup

Given the teacher network is a pretrained classification model then distill to the student one. The main focus is to compare and analyse the difference betwen the student network with the *baseline network* (DNN learned from raw data) (both three models have the same architecture for fair comparison). Let $x \in R^n$ be the input image, $f_t(x), f_s(x) \in R^L$ be the intermediate-layer features and $y_t = g_t(f_t(x))$, $y_s = g_s(f_s(x))$ be the classification results of teacher and student, respectively. As the effect of KD, $f_s(x)$ is forced to be approximately equivanlent to $f_t(x)$

The author used the method proposed by [21] in order to quantify the discarded information of the input (consider as the conditional entropy$H(X')$) given the intermediate-layer feature $f^* = f(x)$ as follows:

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

where X' denotes the set of images having a specific object instance (represented by a small range of feature $||f(x') - f^*||^2 \leq \tau$). With a proper condition of x' ($x' \sim \mathcal{N}(x, \Sigma = diag(\sigma_1^2, ..., \sigma_n^2))$, where n is the number of images), the entropy $H(X')$ of the entire image can be formed by the entropies of every pixel $\{H_i\}$:

$$H(X') = \sum_{i=1}^{n} H_i$$

where $H_i = \log\sigma_i + \frac{1}{2}\log(2\pi e)$ measures the discarded information of the i-th pixel.

3.2. Proposed hypotheses

Hypothesis 1: Knowledge distillation makes the DNN learn more reliable visual concepts then learning from raw data.

The paper aims to compare the number of visual concepts that are encoded in the baseline network and those in the student network. They assumed that since those pixels having high H_i would not contribute much to the prediction result of the model, they would be considered as task-irrelevant feature, while those having smaller value are called *visual concepts*: image region whose information is significantly less discarded and mainly used by the DNN. Using the same concept of information theory, this is how the authors quantify the knowledge:

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

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

$$\lambda = \mathbb{E}_{x \in \mathbf{I}} \left[\frac{N^{fg}_{concept}(x)}{N^{fg}_{concept}(x) + N^{bg}_{concept}(x)} \right]$$

where $N^{bg}_{concept}(x), N^{fg}_{concept}(x)$ are the number of visual concepts encoded on the background and foreground respectively, $\Lambda_{bg}, \Lambda_{fg}$ are the sets of pixel of background and foreground w.r.t. image x, $\overline{H} = E_{i \in \Lambda_{bg}}[H_i]$ is the average entropy value of the background pixels which can be used and a baseline entropy. The pixels that having smaller entropy value than this baseline by an small amount b are considered task-relevant visual concepts. Finally, λ is proposed as the metric used to measure how effective the model's feature extraction process. The reason why λ can do such thing is argued by the author that statistically the foreground contains more informative features than the background, so a well-trained DNN model should focus on the visual concepts in the foreground.

Hypothesis 2: Knowledge distillation ensures that the DNN is prone to learning various concepts simultaneously, in constrast of DNN learning from raw data which learns these sequentially.

The author proposed another metrics based on the number of learned foreground visual concepts along each training epochs and take measurement with respect to the epoch that the model learn those the most: $\hat{m} = \text{argmax}_k N_k^{fg}(I)$ and "weight distance" function to estimate the learning effect $\sum_{k=1}^{\hat{m}} \frac{||w_k - w_{k-1}||}{||w_0||}$ (where w_i is the weight of the model at epoch i-th):

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

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

where D_{mean} is the average weight distance where the DNN usually extracts the mosts task-relevant visual concepts, the smaller the value the faster the DNN learns the visual concepts during training. While D_{var} was considered inversely proportional to how simultaneously the model learns different visual concepts during training.

Hypothesis 3: Comparing to learning from raw data, knowledge distillation produces more consistent optimization directions.

The final metrics is computed by comparing the amount of learned and chosen foreground visual concepts in the final model $||S_M(I)||$ with the union of those learned in each epoch $|| \bigcup_{j=1}^{M} S_j(I)||$. The proper form of the metric is defined as follow:

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

The higher the value ρ indicates the less detours and more stable training process.

3.3. Discussion

The paper offers a novel and understandable interpretation of knowledge distillation from the perspective of information theory, to be more precise, measuring the information encoded in a DNN's intermediate layers. For each hypothesis, the author proposed metrics that allow us to partly theoretical explain and empirically prove that corresponding theory. But some confusing details might need further investigation: 1) this explaining framework only focuses on classification problem, which is one of the simplest forms, to apply this framework to another application might need extra information 2) The mathematical concept of presenting object and assumption about the distribution of the image set: given a diverse training dataset, whether these metrics function similarly? 3) The estimated epoch $\hat{m}$ used for the second metrics is not a precise estimation.

4. Regularizing Knowledge distillation

As one of the most beneficial characteristic of KD, any student can learn from any teacher disregarding the structural difference. But it is emperically proved that well-trained large DNN doesn't often make good teachers due to the mismatched capacity, which makes the student unable to mimic it [7].

To tackle this problem, multiple work shares the same idea of regularizing the teacher model [7, 20]. In [7], Cho et al. proposed a new process ESKD (Early-stopped knowledge distillation) after emprically prove sequential knowledge distillation is also not that efficient since it can only outperform one model training from scratch but not ensemble of those. They argued that the found solution space of the teacher is not accessible from the student, which means to find a teacher whose discovered solution should be discoverable by the student. Based on another works, the author assumed early stopping allows large model to behave as a small network while still having better search space than smaller ones. Different from Cho et al., Lukasik et al. [20] investigates the denoising effect of label smoothing on noisy data then applies it on the distillation process in order to test its effectiveness. The ending result is that applying label smoothing on the teacher significantly enhances over vanilla distilaltion, while applying the same on the student has mixed results.

Having the same idea, Yuan et al. [36] investigates and compares KD with label smoothing regularizer and later on proposed a novel Teacher-free Knowledge distillation (TfKD) framework. It started with the observation that using either poorly trained teacher to distil student or student to teach teacher model still can improve and enhance the guided models, which suggests considering KD as a regularization term (strongly related to label smoothing regularizer). To put it another way, we may think of KD as an adaptive variant of label smoothing, implying that it can inherit much of label smoothing's regularization advantages, such as model regularization and improved calibration, without being overconfident. Having that in mind, the paper consider replacing the class distribution predictions of teacher model with a simppler one, implemented in the TfKD framework. This framework is particularly useful in circumstances where a more efficient teacher model is inaccessible or where only minimal computing resources are available. There are two methods of distilling in TfKD framework: 1) self-training distillation and 2) Combine KD with Label Smoothing Regularizer to create a 100% accuracy teacher. Both of the methods are very simple yet effective and also emperically proved to be performant.

5. Conclusion

The main technical information and applications of knowledge distillation have been covered in this survey. We also briefly review the taxonomy methods for current KD approaches and include description of the problem. We then investigate how KD is perceived and explained in the current past work. Finally some methods of regularizing KD while applying in real-world application are also discussed.

References

[1] Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations, 2020.

[2] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Noise as a resource for learning in knowledge distillation, 2020.

[3] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression, 2018.

[4] Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535541, New York, NY, USA, 2006. Association for Computing Machinery.

[5] Wei-Chun Chen, Chia-Che Chang, Chien-Yu Lu, and Che-Rung Lee. Knowledge distillation with feature maps for image classification, 2018.

[6] Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge, 2020.

[7] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation, 2019.

[8] Joseph DiPalma, Arief A. Suriawinata, Laura J. Tafe, Lorenzo Torresani, and Saeed Hassanpour. Resolution-based distillation for efficient histology image classification, 2021.

[9] Mengya Gao, Yujun Shen, Quanquan Li, and Chen Change Loy. Residual knowledge distillation, 2020.

[10] Mengya Gao, Yujun Shen, Quanquan Li, Junjie Yan, Liang Wan, Dahua Lin, Chen Change Loy, and Xiaoou Tang. An embarrassingly simple approach for knowledge distillation, 2019.

[11] Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4):20512062, Apr 2019.

[12] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation, 2019.

[13] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation, 2019.

[14] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons, 2018.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[16] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation, 2019.

[17] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer, 2017.

[18] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding, 2020.

[19] Zhizhong Li and Derek Hoiem. Learning without forgetting, 2017.

[20] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise?, 2020.

[21] Haotian Ma, Yinqing Zhang, Fan Zhou, and Quanshi Zhang. Quantifying layerwise information discarding of neural networks, 2019.

[22] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation, 2019.

[23] Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong. Conditional teacher-student learning. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[24] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019.

[25] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation, 2019.

[26] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5142–5151. PMLR, 09–15 Jun 2019.

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015.

[28] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning, 2019.

[29] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[30] David Svitov and Sergey Alyamkin. Margindistillation: distillation for margin-based softmax, 2020.

[31] Tiancheng Wen, Shenqi Lai, and Xueming Qian. Preparing lessons: Improve knowledge distillation with better supervision, 2020.

[32] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020.

[33] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan Yuille. Knowledge distillation in generations: More tolerant teachers educate better students, 2018.

[34] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L. Yuille. Snapshot distillation: Teacher-student optimization in one generation, 2018.

[35] J. Yim, D. Joo, J. Bae, and J. Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7130–7138, 2017.

[36] Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization, 2021.

[37] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2017.

[38] Chenrui Zhang and Yuxin Peng. Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification, 2018.

[39] Yang Zhao, Yifan Liu, Chunhua Shen, Yongsheng Gao, and Shengwu Xiong. Mobilefan: Transferring deep hidden representation for face alignment. *Pattern Recognition*, 100:107114, 2020.