Knowledge Distillation : A Survey

Reading By, Harsh Srivastava
B.Tech in Computer Science
IIT Patna
Patna, India
harshsrivasta243@gmail.com

Guided By, Saurabh Sharma
Ph.D, Computer Science
IIT Patna
Patna, India

Supervised By, Dr. Joydeep Chandra
Professor, Dept.of Computer Science
IIT Patna
Patna, India

***Abstract*—This document is a (summarized) reading of the paper titled "Knowledge Distillation : A Survey" which is originally authored by Jianping Gou[1], Baosheng Yu[1], Stephen J. Maybank[2] and Dacheng Tao[1].**

The Venue of Publication :

1. UBTECH Sydney AI centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia
2. Department of Computer Science and Information Systems, Birkbeck College, University of London, UK.

The link to the original paper is here :

https://github.com/harshsrivasta243/1901CS24_CS299_2021/blob/Papers/Paper1.pdf

I. Motivations

A. Challenges : Overview

1) Deep Learning applications and improvisations: Deep learning has applications in computer vision, reinforcement learning, natural language processing, etc. Deep neural networks are good at learning multiple levels of feature representation with increasing abstraction. This is known as representation learning. For increasing efficiency of deep neural networks, focus is laid on :

- efficient building blocks for deep models, including depth-wise separable convolution
- model compression and acceleration

2) Need for model compression and acceleration: Success of deep learning is due to the ability of encoding large scale data and handling too many parameters at once. Such a model has high computational complexity and high storage requirements, which is not feasible on phones or embedded devices; which we use the most.

B. Challenges : Ideation to overcome

1) Model compression and acceleration: It is a technique to reduce the execution complexity and storage requirement of a model so as it could run on systems with limited processing capability and storage space. This methodology is further categorised into :

- Parameter pruning and sharing: removes inessential parameters from the model without any significant change in performance. Subcategories are model quantization, model binarization, structural matrices and parameter sharing.
- Low-rank factorization: identifies redundant parameters of deep neural networks via matrix/tensor decomposition.
- Transferred compact convolutional filters: removes inessential parameters by transferring/compressing convolutional filters.
- Knowledge distillation(KD): distills knowledge from a larger deep neural network (teacher model/network) into a smaller network (student model/network).

II. Problem(s) Addressed

Large deep models tend to achieve good performance in practice, because the over parametrization improves the generalization performance when new data is considered. In KD, a small student model is supervised by a teacher model in general. The key problem addressed by the paper is theoretizing the methodologies and thus devising efficient techniques to transfer the knowledge from the teacher model to the student model.

III. Architecture Used

The paper focusses on the architecture of the teacher and student models mainly, however not in a generalized form, but mainly specifying the student network relative to the teacher network and discussing methods to reduce degradation due to architecture.

A. The Teacher-Student Architecture

In knowledge distillation, the teacher-student architecture is a generic carrier to form the knowledge transfer. In general, knowledge is needed to be transferred from deeper and wider neural networks to shallower and thinner neural networks. Owing to this fact, the student model choices are made usually from among the following :

- a simplified version of teacher network with fewer layers and fewer channels in each layer
- a quantized version of a teacher network in which the structure of the network is preserved

- a small network with efficient basic operations
- small network with optimized global network structure
- network exactly the same as teacher

The model capacity gap between the large teacher neural network and the small student neural network can degrade knowledge transfer. For effective knowledge transfer and minimal degradation due to architecture, the following methods can be utilized :

- introducing a teacher assistant to mitigate the gap between teacher and student models
- making the assistant structure learn the residual error, ie, residual learning
- minimizing the difference in structures of the teacher and student models
- making the student model a small and quantized version of the teacher model
- transferring the knowledge learned by multiple layers to a single layer
- block-wise knowledge transfer from teacher to student along with preserving the receptive field
- using Depth-wise separable convolution to design efficient networks for mobile/embedded devices
- improving performance by searching for a global structure based on efficient meta-operations or blocks
- removing redundant layers in a data-driven way using reinforcement learning

A joint search of student structure and knowledge transfer under the guidance of the teacher model is an interesting topic of further approach and is to be still explored.

IV. Baseline

A. Knowledge Distillation: Overview

Knowledge Distillation (KD) is a representative type of model compression and acceleration which effectively learns a small student model from a large teacher model. A KD system consists of 3 components: knowledge, distillation algorithm and teacher-student architecture. Successful distillation relies on data geometry, optimization bias of distillation objective and strong monotonicity of the student classifier.

A larger model may not be a better teacher because of model capacity gap. Experiments also show that distillation adversely affects the student learning. KD has also been explored for label smoothing, for assessing the accuracy of the teacher and for obtaining a prior for optimal output layer geometry. The training of a student model depends on knowledge categories, training schemes, teacher-student architecture, distillation algorithms, performance comparison and applications.

The former extensions of KD are Teacher-Student learning, Mutual learning, Assistant teaching, Lifelong learning, and Self learning. KD and its extensions focus on compressing deep neural networks. The resulting lightweight student networks are then applied in visual recognition, speech recognition, natural language processing, etc. The transfer of knowledge from one model to another can be extended to data augmentation, data privacy, data security, adversarial attacks and others.

Dataset distillation is a technique inspired by KD which compresses the training data, ie, transfers knowledge from a large dataset to a small dataset to reduce training loads of deep models. KD distills the knowledge from a large teacher model to a compressed, lightweight student model without a significant drop in accuracy and performance. The main idea is that the student model mimics the teacher model in order to obtain a competitive or even a superior performance.

B. Vanilla Knowledge Distillation Method

A vanilla KD framework contains one or more large pre-trained teacher models and a small student model. The supervision signal from a teacher model, called "knowledge" learned by the teacher model, helps the student model to mimic the behaviour of the teacher model, with comparable accuracy. The logits, ie, the output of last layer in a deep neural network, are used as the carriers of knowledge viz-a-viz the teacher and student models, which, in turn, aren't explicitly provided by the training data examples. This knowledge learned by the teacher model, is called dark knowledge.

C. Transfer of dark knowledge in a vanilla KD framework

Given a vector of logits z as the output of the last fully connected layer of a deep model, such that z_i is the logit for ith class, then the probability that the input belongs to the ith class is given by the softmax function,

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

The reasons for using the softmax function for the purpose is,

- exponentiation(raising e to power z_i) is done in order to ensure that we have only positive values to work with.
- normalization(dividing by $\sum_j e^{z_j}$) is done to make sure all probabilities sum to 1, ie, a probability distribution is created.

Probability distributions(soft targets) in place of one-hot labels(hard targets) are preferred so that ground truth probability distribution could be compared with predicted probability distribution effectively and efficiently. These predicted soft targets by the teacher model comprise of dark knowledge and can be used to transfer the same from the teacher model to the student model.

D. The Temperature(T) factor

The temperature factor is induced in the softmax function in order to control the importances of the soft targets and to obtain a smoother probability distribution. Smoother here means that unlike regular softmax there is no big spike corresponding to one entry. This modifies the softmax function,

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

Where higher temperature provides smoother distribution.Typically, when T tends to infinity all classes share same probability and when T tends to zero soft targets become one-hot labels.

E. Distillation Loss

The distillation loss is defined to match the logits of the teacher model and the student model. The logits of the teacher model are matched by the cross-entropy gradient wrt the logits of the student model. The gradient of the distillation loss function wrt the ith student model logit(z_{si}), for assumed high temperature computed via Taylor Series can be given by

$$\frac{1}{T}\left(\frac{1+\frac{z_{si}}{T}}{N+\sum_j z_{sj}/T}-\frac{1+\frac{z_{ti}}{T}}{N+\sum_j z_{tj}/T}\right)$$

wherein if we further assume that the logits of each transfer training set example are zero-mean, ie, $\sum_j z_{sj}/T = \sum_j z_{tj} = 0$, then the same gradient will become,

$$\frac{1}{NT^2}(z_{si}-z_{ti})$$

From which we can conclude that the distillation loss is equal to matching the logits between the teacher model and the student model under the conditions of high temperature and zero-mean logits , ie, minimizing ($z_{si}-z_{ti}$).Hence, distillation through matching logits with high temperature can convey very much useful knowledge information learned by the teacher model to train the student model.

F. Student Loss and combining the losses

The student loss is defined as the cross-entropy between the ground truth label and the soft logits of the student model. In distillation and student losses, both use the same logits of the student model but different temperatures. The temperature is T=1 in student loss and T=t (t is greater than 1) in the distillation loss. Finally, the benchmark model of a vanilla KD is the joint of the distillation and student losses.

$$L(x,W) = aL_D + (1-a)L_S;$$

where L_D stands for distillation loss, L_S for student loss, and, x is a training input on the transfer set, W is a student model parameter and 'a' is a regulation parameter.
On a general basis, the teacher model is always pre-trained and the student model is trained using the knowledge only from soft targets of the pre-trained teacher model. It will be learned later that this may also be called as offline knowledge distillation with response-based knowledge.

G. Knowledge

The knowledge that we can use for distillation can come from different portions of the teacher model. Hence we classify knowledge into 3 different categories, where each category has its own properties and ways of handling. This categorisation is depicted below:

1) Response-Based Knowledge: It refers to the logits' vector of the teacher model, ie, the last layer neural response of the teacher network. The distillation loss for response-based knowledge is the Kullback-Liebler(KL) divergence loss. Recently, this knowledge, which is also called "dark knowledge" has been further explored to address the information of ground-truth label as the conditional targets. However, since this focuses on the last layer only and not the information in the intermediate layers, there are many areas it leaves untouched. As the soft logits are the class probability distribution, this knowledge is limited to supervised learning.

2) Feature-Based Knowledge: Both the output of the last layer and the output of intermediate layers, ie, feature maps, is referred to as the feature-based knowledge. The main idea is to directly match the feature activations of the teacher and the student. A lot of research work went under this topic, a few major ones are listed below:

- derivation of attention maps to express knowledge
- generalization of attention maps using neuron selectivity transfer
- transfer of knowledge by matching the probability distribution in feature space
- introduction of "factors" for easy and feasible transfer of teacher knowledge
- developing route constrained hint learning for performance matching
- using activation boundary of hidden neurons for knowledge transfer

The parameter sharing of intermediate layers of the teacher model together with response-based knowledge is also used as the teacher knowledge. Though feature-based knowledge transfer provides favorable information for the learning of student model, the following areas about it are still to be explored and investigated :

- how to effectively choose the hint layers from the teacher model and the guided layers from the student model ?
- due to the significant differences between sizes of hint and guided layers, how to properly match feature representations of teacher and student ?

3) Relation-Based Knowledge: Response-based knowledge refers to the knowledge of relationships between different layers or data samples. A Flow of Solution(FSP) process was proposed for interpreting relationships between different feature maps, ie, relation-based knowledge. The properties of the same are listed below :

- FSP is defined by the gram matrix between two layers
- The FSP matrix summarizes the relations between pairs of feature maps
- It is calculated using the inner products between features from two layers

The relation-based knowledge is a very vital category of knowledge for KD. Its varying utilities and methods pertaining to a smooth distillation are :

- Using relation-based knowledge as the distilled knowledge, KD via singular value decomposition is capable of extracting key information from feature maps.
- To use the knowledge from multiple teachers, 2 graphs could be formed by respectively using the logits and features of each teacher model as the nodes.
- The importance and relationships of the different teachers are modeled by the logits and representation graphs before the knowledge transfer.
- Multi-head graph-based knowledge distillation was also proposed.
- The graph knowledge is the intra-data relations between any two feature maps via multi-head attention network.
- For delving into the pairwise hint information, the student model also mimics the mutual information flow from pairs of hint layers of the teacher model.

V. Broad Techniques

A. Knowledge transfer and distillation methods

After analysing the knowledges which can be used to train a student model from a teacher model for knowledge distillation, many knowledge transfer techniques were hence proposed, a few are listed here :

- Individual knowledge distillation : the individual soft targets of the teacher are directly distilled into the student.
- Instance relationship graph : the transferred knowledge contains instance features, instance relationships and the feature space transformation cross layers.
- Relational knowledge distillation : transfers the knowledge from instance relations. Based on manifold learning, the student network is learned by future embedding, preserving the feature similarities of samples in the intermediate layers of the teacher networks. The relations between data samples are modelled as probabilistic distribution using feature representations of data, and the probabilistic distributions of teacher and student are matched by knowledge transfer.
- Similarity-preserving knowledge distillation : similarity-preserving knowledge, which arises from the similar activations of input pairs in the teacher networks, is transferred into the student network, with pairwise similarities preserved.
- Correlation congruence knowledge distillation : the distilled knowledge contains both the instance-level information and the correlations between instances. The student network can learn the correlation between instances by this method.

Further, distilled knowledge can be classified from different perspectives such as structured knowledge of data or privileged information about input features. However, how to model the relation information from feature maps or data samples as knowledge still needs to be explored.

VI. Techniques Proposed

A. Distillation Schemes

The learning schemes of KD(Distillation schemes) can be divided into three main categories :

1) Offline Distillation: The teacher model is pre-trained before distillation and the knowledge from the teacher model, in form of logits or intermediate outputs of feature maps, is then distilled and used for the training of the student model. Offline distillation usually employs one-way knowledge transfer and two-phase training procedure, is easier to implement, and the training of the student model is efficient this way.

2) Online Distillation: Both the teacher and student models are updated simultaneously, and the whole KD framework is end-to-end trainable. The advancements in online distillation techniques have been eminent frequently and the major ones are listed below :

- introduction of Deep Mutual Learning(DML) for online distillation
- extension of DML by using ensemble of soft logits for improving generalizing ability
- introduction of auxiliary peers and group leader in DML for forming a set of peer models
- proposition of a multi-branch architecture in DML for reducing computational cost
- introduction of feature fusion module for constructing the teacher classifier
- replacement of convolution layer with cheap convolution operations for forming student model
- employment of online distillation for training large-scale distributed neural network
- introduction of codistillation for training multiple models with same architecture in parallel
- proposition of an online adversarial KD method for training multiple networks by the discriminators

Online distillation is a one-phase end-to-end training scheme with efficient parallel computing. However, how to address the high-capacity teacher in online settings is still to be explored.

3) Self Distillation: The same networks are used for the teacher and the student models. It is a special case of online distillation. The chief idea is to distill the knowledge from the deeper sections of the network to the shallower sections. Extended methods, namely, teacher-free KD method, novel self-KD method, class wise self-KD method, etc. have also been proposed, and it is also adopted to optimize deep models.

B. Distillation Algorithms

The different algorithms that the paper suggests for distilling knowledge from teacher model to student model are summarized as below :

- Adversarial Distillation : A discriminator and a generator of GAN(generative adversarial network) generate synthetic data which can be utilized in many ways for efficient knowledge distillation.
- Multi-Teacher Distillation : The average response from all teachers can be used as the supervision signal, and

with slight modifications and extensions, it can be used in other applications apart from KD.
- Cross-Modal Distillation : Given a teacher model pre-trained on one modality, knowledge transfer can be done from teacher model to student model with a new unlabelled input modality.
- Graph-based Distillation : The main idea of this distillation method is to use the graph as the carrier of teacher knowledge and to use the graph to control the message passing of the teacher knowledge.
- Attention-based Distillation : The core of attention transfer is to define the attention maps for feature embedding in the layers of a neural network. It is used to boost the performance of student network.
- Data-Free Distillation : There is no training data available. The transfer data is synthetically generated by a GAN. It is reconstructed by using the layer activations or spectral activations of the teacher network.
- Quantized Distillation : It reduces the computation complexity of neural networks by converting high precision teacher networks into low precision student networks.
- Lifelong Distillation : It accumulates the previously learned knowledge and also transfers the learned knowledge into future learning. Its extensions can address the problem of catastrophic forgetting.
- NAS-based Distillation : NAS(Neural Architecture Search) is an autoML technique and aims to reduce the capacity gap between teacher and student by finding appropriate student architecture for KD.

VII. Performance Aspects

The performances of the teacher and student models, whilst the student model is trained via knowledge distillation, have been compared for several cases. The results could be summarized as follows :

- KD can be simply realized on different deep models and model compression of different deep models can be easily achieved.
- The online distillation through collaborative learning can significantly improve the performance of deep models.
- The self-distillation as well can improve the performance of the deep models
- The performance of lightweight student deep models can be easily improved by the knowledge transfer from the high capacity teacher models using knowledge distillation.

VIII. Results and Conclusions

A. Applications

KD is applied in varietic fields of technology, especially in the Artificial Intelligence(AI) domain viz. visual recognition, speech recognition, natural language processing(NLP) and recommendation systems. It can also be used for purposes like data privacy and defense against adversarial attacks. The following points illustrate further :

- KD provides efficient and effective teacher-student learning for a variety of visual recognition tasks.
- KD can make full use of different kinds of knowledge in complex data sources.
- KD provides efficient and effective lightweight language deep models.
- The teacher-student knowledge transfer can easily and effectively solve many multilingual tasks.
- In deep language models, the sequence knowledge can be easily transferred from large to small models.
- The lightweight student model can give real-time responses and have high recognition accuracy.
- RNN teacher models can transfer temporal knowledge from real acoustic data to a student model.
- Sequence-level KD can be well applied to sequence models with good performance.
- KD can also be applied in problems of multi-accent and multilingual speech recognition.

B. Further Directions of Research and Investigation

The challenges that are still persistent and which need to be explored and investigated are :

- This paper introduces us to different forms of knowledge, viz., response-based, feature-based and relation-based. However, how to model these different forms of knowledge in a unified and complementary framework is still a challenge.
- To improve the efficacy of knowledge transfer, the relationships between the model complexity and existing distillation schemes or other novel distillation schemes could be further investigated.
- The design of an effective model or construction of a proper teacher and student model are still challenging problems in knowledge distillation.
- A deep understanding of generalizability of knowledge or the quality of the teacher-student architecture, is still very difficult to attain.

Apart from solving the above issues, the directions which could be further investigated to extend knowledge distillation and its application areas in future are vast and wide. Here we try to enlist a few of those :

- Devising a technique to decide the order for applying different compressing methods will be an interesting topic for future study.
- Knowledge distillation has been proven to have a potential in fields other than model compression, which are seldom investigated further. It would be nice to extend KD in other domains.
- Knowledge distillation could be integrated with other learning schemes like adversarial learning, auto machine learning, lifelong learning and reinforcement learning for practical challenges in the future.