Knowledge Distillation

Aditya Kumavat   
 *MIT School of Computing*  
*MIT Art, Design, and Technology University*Pune, 412201, India

+91 7020653917  
adityakumavat03126@gmail.com

Raviraj Verma  
 *MIT School of Computing*  
*MIT Art, Design, and Technology University*Pune, 412201, India

+91 9970562525  
raviraj72013@gmail.com Abhishek Das  
 *MIT School of Computing*  
*MIT Art, Design, and Technology University*Pune, 412201, India

+91 9874920971  
abhishek.das@mituniversity.edu.in

Yash Bora   
 *MIT School of Computing*  
*MIT Art, Design, and Technology University*Pune, 412201, India

+91 9922911206  
yashbora1210@gmail.com

Aman Jha   
 *MIT School of Computing*  
*MIT Art, Design, and Technology University*Pune, 412201, India

+91 7249619459  
amankj4930@gmail.com

**ABSTRACT**

Knowledge distillation is simply converting large model into smaller model for particular requirement. Large model also known as teacher and smaller as student. Student model learn from teacher model by using some loss function which combines cross entropy with ground-verify markers and make teacher logits soft. It can reduce computational cost in real-world scripts and retains their delicacy and performance. This is helpful for business to put their big models in real-world constraints. Usually, it is performed on neural network which contains complex infrastructures with several layers and model parameters. Also, the researches done in last decade helped in success of fields like speech recognition, natural language process and image recognition. This way knowledge distillation gained elevation in real-world operations.

**KEYWORDS**

Deep learning, Tensorflow, Distillation algorithm, teacher-student Model.

# INTRODUCTION

In this paper research it uses deep literacy models to produce knowledge distillation result. Which allows business to efficiently put their big models in real-world applications. Recent researches on Artificial Intelligence (AI) and Machine Learning (ML) have enabled development of knowledge distillation operations. It can be used by associations to make another model with accuracy of main model but reduce the size of parent model, and maintain same delicacy. When associations developing a model, their high concern is to achieve maximum delicacy for that they make use of billions of parameters, which leads to big and cumbrous model. To put this or similar big and cumbrous models, it’s almost impossible to make it compatible for mobile and small devices.

# EASE OF USE

# The Knowledge Distillation paper make it easy for associations to efficiently put their model in mobile devices and embedded bias. This paper explains the use of deep learning models to make distillation algorithms as per current persisting trends.

# KNOWLEDGE

## In knowledge distillation, distillation strategies, knowledge types, and teacher-student infrastructures play major role in student learning. This section focuses on colorful feathers of knowledge for knowledge distillation in this part.

## Response–Based Knowledge

## Response-based knowledge often refers to the neural response of teacher model’s last affair sub-caste. The main goal is to directly mimic the final vaticinator of the teacher model. The response-based knowledge distillation is simple but effective for model contraction and has been highly used in different tasks and operations.

## 2.2 Feature–Based Knowledge

## Deep neural networks are good at learning multiple situations of point representation by adding abstraction. This is known as representation literacy. Thus, both the affair of intermediate layers and affair of the last sub-caste, i.e. point charts, can be used as the knowledge to supervise the training of the student model. Specifically, point-grounded knowledge from the intermediate layers is a good extension of response-grounded knowledge, especially for the training of thinner and deeper networks. The main idea is to directly match the point activations of the teacher and the student.

## Relation–Based Knowledge

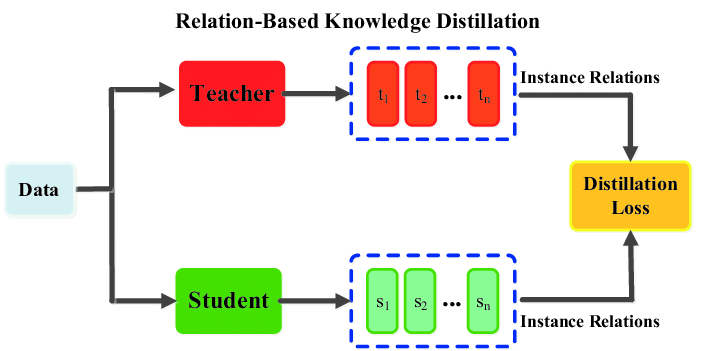
There are two main types: response-based knowledge and feature-based knowledge. Both of these types of knowledge utilize specific layers in the teacher model. However, there is also a third type of knowledge called relation-based knowledge, which explores the connections between different layers or data samples. One interesting approach to exploring these connections is the Flow of Solution Procedure (FSP), which was proposed by Y et al. (2017). FSP is defined by the Gram matrix between two layers, and it summarizes the relations between dyads of point charts. The correlations between point charts can then be used as distilled knowledge. To further prioritize crucial information in the point maps, knowledge distillation via singular value corruption has been proposed.

Fig: The general case relation-based knowledge distillation

# DISTILLATION SCHEMES

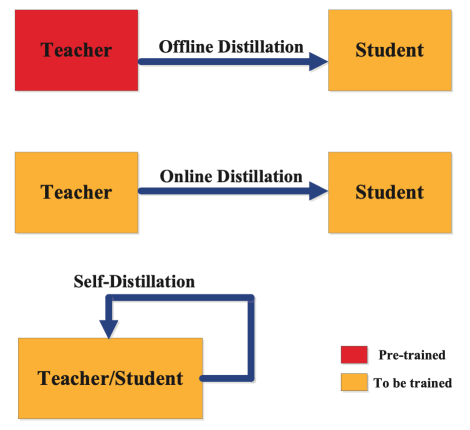
In this section, I'll be discussing the distillation schemes, also known as training schemes, for both teacher and student models. There are three main categories of knowledge distillation based on whether the teacher model is streamlined contemporaneously with the student model or not: offline distillation, online distillation, and tone-distillation. 

Fig: Different distillations. The red color for “pre-trained” means networks are learned before distillation and the unheroic color for “to be trained” means networks are learned during distillation.

## Offline Distillation

## Most of the knowledge distillation methods are designed to work offline. In the case of characterless knowledge distillation (H et al. 2015), the knowledge is transferred from a pre-trained teacher model to a student model. The training process includes two stages. Firstly, the large teacher model is trained on a set of training samples before the actual distillation process. Secondly, the teacher model is used to extract the knowledge in the form of logits or intermediate features, which are further utilized to guide the training of the student model during the distillation process.

## Online Distillation

## Although offline knowledge distillation approaches have been effective, they have also received significant attention from researchers who have identified some limitations. To address these limitations, online distillation has been proposed as an alternative method to further enhance the performance of student models. Online distillation is especially useful when a high-performance teacher model with a large capacity is not available. In online distillation, both the teacher and student models are trained concurrently, and the entire knowledge distillation framework is end-to-end trainable.

## 3.3 Self-Distillation

# In the context of knowledge distillation, self-distillation involves using the same neural network architecture for both the teacher and student models. Essentially, self-distillation is a type of online distillation where the teacher and student models are the same network. Moreover, offline, online, and self-distillation can also be interpreted from the perspective of human teacher-student learning. In offline distillation, the teacher imparts knowledge to the student. In online distillation, the teacher and student learn together. In self-distillation, the student learns by oneself. These three types of distillation can be combined to complement each other, taking advantage of their respective strengths. For instance, the multi-knowledge transfer framework integrates both self-distillation and online distillation.

# DISCUSSIONS

## Are bigger models more preceptors? The fundamental concept of KD is that the soft targets (probabilities) generated by a pre-trained teacher model convey more information about the data distribution than the hard targets (actual labels) (1). Moreover, it is expected that as the teacher model becomes more robust, the soft targets it produces become more reliable and capture the true class distribution. This implies that a stronger teacher model can provide more informative guidance to the student model, thereby improving its performance. However, contrary to this intuition, empirical evidence suggests that using a larger and more powerful teacher model does not always result in better student performance. In fact, as the capacity of the teacher model increases, the student model's accuracy tends to improve only up to a certain point, after which it starts to deteriorate.

**Is a pre-trained teacher important?** Focus of majority workshops is on training smaller student model based on a pre-trained teacher model. However, this approach is not guaranteeing effectiveness, especially when capacity difference between the teacher and the student models is high. This can make it hard for the student model to follow the teacher model, which leading to optimization difficulties. So, the question is whether pre-trained teacher model is essential for training a compact student model with high performance. One alternative approach is to learn from peer students who have the similar or same model architecture. This approach has advantage of being effective because pre-training a high-capacity teacher model is not required. Instead of being taught, the student peers learn to collaborate and achieve optimal learning results.

**Logits vs features**. When it comes to knowledge distillation (KD) styles, the knowledge can be defined from three aspects: logits, feature maps (intermediate layers), or both. However, it is still unclear which one of these represents better knowledge. Some workshops have focused on the interpretation of feature maps and claimed that they contain richer information, while others have argued that logits can represent each sample by a class distribution and that a student can learn intra-class variations easily. Nevertheless, it is evident that KD via logits has some significant downsides. First, its effectiveness is limited to the SoftMax loss function and relies on the number of classes, making it unsuitable for low-resolution vision tasks. Second, when the teacher and student models have a large capacity difference, it can be difficult for the student to follow the teacher's class probabilities. Therefore, it is safe to say that feature maps provide richer knowledge than logits and generalize better to problems without class probabilities.

**CONCLUSION:**

In today's world, deep learning operations are dependent on neural networks which has large capacity, memory footprints, and slow conclusion time. However, applying such models to real-life products is big challenge. To solve issues like this, knowledge distillation provides an better way to train a smaller, faster, and cheaper student model derived from a large, complex teacher model. Knowledge distillation is complex process which involves various types of knowledge like training schemes, infrastructures, and algorithms. It has proven to be highly successful in many fields such as computer vision, natural language processing, and speech, among others.

**ACKNOWLEDGMENTS**

We're extremely happy to present our project “Knowledge Distillation” with you. There are many individuals who directly or indirectly contributed to this project and make it this better, and we would like to express our gratitude to them all.

We're especially grateful to our mentor Abhishek Das for supporting and guiding us throughout the project. We are also thankful to him for his valuable insights and cooperation during the project. Last but not least, we're grateful to our team members for their support and motivation which helped us to complete the project successfully. We hope that this solution will serve the purpose for which it's developed thereby highlighting the success of the process.

# REFERENCES

1. Iulia Turc, Ming- Wei Chang, Kenton Lee, Kristina Toutanova (2019). WELL-READ STUDENTS LEARN BETTER: ON THE SIGNIFICANCE OF PRE-TRAINING COMPACT MODELS. <https://arxiv.org/pdf/1908.08962.pdf>
2. Samuel Stanton, Pavel Izmailov, Polina Kirichenko (2021). DOES KNOWLEDGE DISTILLATION WORK? <https://arxiv.org/pdf/2106.05945.pdf>
3. LucasBeyer,XiaohuaZhai,AmelieRoyer,LarisaMarkeeva,Rohan,Anil,Alexander Kolesnikov(2022)

KNOWLEDGE DISTILLATION: A GOOD TEACHER IS PATIENT AND CONSISTENT.

<https://arxiv.org/pdf/2106.05237.pdf>

1. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam (2017)

MOBILENETS: EFFICIENT CONVOLUTIONAL NEURAL NETWORKS FOR MOBILE VISION APPLICATIONS

<https://arxiv.org/pdf/1704.04861.pdf>

1. Lukas Bossard, Matthieu Guillaumin, Luc Van Gool (2014)

FOOD-101 – MINING DISCRIMINATIVE COMPONENTS WITH RANDOM FORESTS.

<https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/static/bossard_eccv14_food-101.pdf>

1. Jianping gou, baosheng yu. Stephen j. maybank & dacheng tao (2021)

KNOWLEDGE DISTILLATION: A SURVEY

<https://link.springer.com/article/10.1007/s11263-021-01453-z>

1. W Park, D Kim, Yan Lu, M Cho (2019)

RELATIONAL KNOWLEDGE DISTILLATION

<https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html>

1. M Phuong, C Lampert (2019)

TOWARDS UNDERSTANDING KNOWLEDGE DISTILLATION

<https://proceedings.mlr.press/v97/phuong19a.html>

1. L Wang, kuk-jin yoon (2021)

KNOWLEDGE DISTILLATION AND STUDENT-

TEACHER LEARNING FOR VISUAL INTELLI-

GENCE: A REVIEW AND NEW OUTLOOKS

<https://ieeexplore.ieee.org/abstract/document/9340578?casa_token=B9vZEkR8XNMAAAAA:RGZAHUwEMRk7ZdD28-B_nlddmJxKl8JFi4hIi0VfadBc0U5xgpUfwLGi4gtLX4hbgiFGP5dqSvGe-ss>

1. Y Liu, K Chen, C Liu, Z Qin, Z Luo, J Wang

STRUCTURED KNOWLEDGE DISTILLATION

FOR SEMANTICS SEGMENTATION

<https://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Structured_Knowledge_Distillation_for_Semantic_Segmentation_CVPR_2019_paper.html>

kno

KN