RESPONSE-BASED KNOWLEDGE DISTILLATION

Vibhas Kumar Vats

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the Department of Data Science,
Indiana University
May 2021

Accepted by the Graduate Faculty, Indiana University,
in partial fulfillment of the requirements for the degree of
Master of Science.

Master's Thesis Committee

David J. Crandall, Professor

David B. Leake, Professor

Md. Alimoor Reza, Dr.

Date of Defense: 05/07/2021

To my grand-parents, parents and siblings!

ACKNOWLEDGEMENTS

For the first time in my life, I took upon writing a scientific work. This journey was filled with learning. This work would not have been possible without a great deal of support and assistance from my mentor, friends, and family.

I would first like to thank my supervisor, Professor David J. Crandall, whose expertise was invaluable in formulating this research work. Your insightful feedback helped me connect the dots of various experiments and brought my work to a higher level. I have very thankful to you for listening to me throughout the various discussion and especially when I was not making sense to myself.

I would like to acknowledge my colleagues from the IU Computer Vision Lab for their feedback and discussion during many of our group sessions. I would particularly like to single out Nikhil Thakurdesai and Md. Alimoor Reza for critical discussions to shape this work.

I would also like to thank my course tutor, Dr. Minje Kim, for his valuable guidance throughout the deep learning systems coursework. You provided me with the tools that I needed to choose the right direction and complete this work.

In addition, I would like to thank my parents and siblings for believing in me to take up this journey. You are always there for me. Finally, I could not have completed this work without thinking the research ideas out loud with my close friend, Pauravi Wagh, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

Vibhas Kumar Vats

RESPONSE-BASED KNOWLEDGE DISTILLATION

The response-based knowledge distillation process behaves differently for different capacities of teacher and student model pairs. A teacher model with a larger capacity to learn the data distribution does poor distillation as compared to a smaller capacity teacher model. Without much analysis of this unintuitive outcome, the degradation in distillation performance has largely been attributed to the gap in learning capacities of teacher and student models. Our analysis finds that the quality of soft labels of the teacher model, presence and absence of similarity information in soft labels, plays a very significant role in governing the distillation performance. We show that a well-trained large learning capacity teacher model learns very fine discriminative properties for classification tasks and loses similarity information between classes, leading to the degradation in distillation performance. We argue that the presence of similarity information in soft-labels facilitates *one-example-to-many-classes learning*, leading to a faster and better knowledge distillation, whereas the absence of similarity information in soft-labels facilitates *one-examples-to-one-class learning*. We present these results through the soft-label hypothesis. Based on this hypothesis, we also propose some special considerations to train a teacher model for retaining the similarity information in soft labels. We argue that by finding the right balance between batch size and the number of epochs of pre-training of a teacher model, a much better distillation performance can be achieved for a given teacher and student model. We demonstrate the generalization of the hypothesis on three different datasets, MNIST, Fashion-MNIST, and CIFAR-10.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The emergence of deep learning (DL) techniques has brought about many successes in recent years. Wider and deeper DL models can address some of the unattainable problems of the recent past by learning billions of parameters through back-propagation. At the same time, these billion parameter networks have brought along a new set of problems, like high computational expenses and large memory requirements [19] that need attention. Large and cumbersome DL models with billions of parameters are computationally very expensive, they also require large memory for storage and execution. These requirements have become one of the biggest hurdles in its deployment on edge devices [9] like mobile phones and embedded devices.

Many methods have been developed in recent years to compress the cumbersome DL models and make them more viable on edge devices. Like every other problem in DL, researchers have taken to different approaches to address this problem. One such approach focuses on parameter pruning and sharing by removing inessential parameters without significantly affecting the model performance, like model quantization [23], model binarization [2], parameter sharing [7, 22]. Another approach focuses on low-rank factorization [3, 25] by using matrix decomposition to remove redundant parameters. Yet another approach focuses on convolution filters and removes inessential parameters by transferring or compressing filters from different convolutional layers [27]. In this work, we focus on the intricacies of knowledge distillation (KD) approach that aims to distill the knowledge from a cumbersome DL model (teacher model) to a more efficient DL model (student model).

The KD process involves two basic steps. First, knowledge extraction from teacher(s)[1] model, and second, the distillation of the extracted knowledge to a student model. A careful

[1] There can be more than one teacher, but we only talk about a single teacher in the rest of the work.

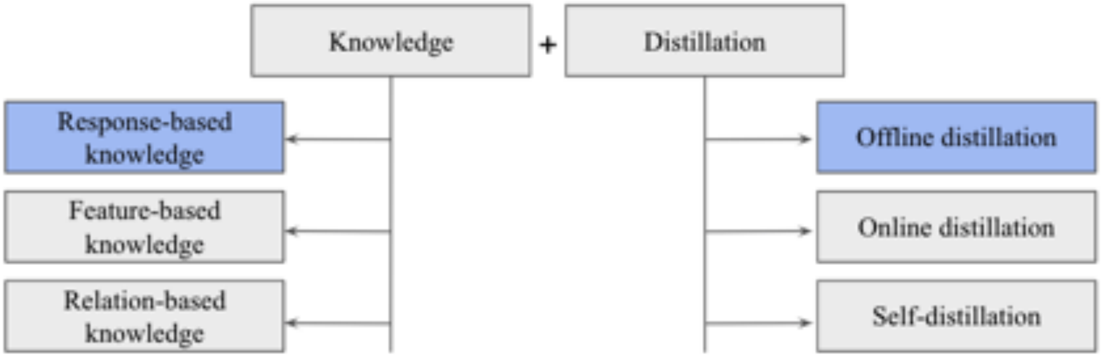


Figure 1.1: KD process categorization [6], in this work, we focus on response-based offline knowledge distillation (the combination in blue box).

stratification of these basic steps has been discussed by Gou et al. in [6]. They categorize the KD process based on different methods used for knowledge extraction and knowledge distillation. Knowledge extraction from the teacher model can be response-based, features-based, or relation-based and distillation to a student model can be offline, online, or self-distillation as shown in figure 1.1. In this work, we focus on one specific KD process that uses the response from a teacher model to acquire the knowledge and distill this knowledge to a student model in an offline manner. This method, proposed by Buciluă et al. [1] and popularized by Hinton et al. [8], is called a response-based offline knowledge distillation process, response-based knowledge distillation (RBKD) [2]. *We focus on understanding the intricacies of the RBKD process by analyzing the quality of response from the teacher model and its impact on the distillation process. We also discuss special considerations for pre-training the teacher model for best distillation performance.*

Both steps, knowledge extraction and knowledge distillation, of RBKD process, are simultaneously performed as shown in figure 1.2. But the teacher model is pre-trained on the given data distribution to be able to extract the knowledge from its response. In the RBKD process, first, the soft labels are generated from the output layer (softmax layer) of the teacher and then, these soft labels, along with original one-hot labels, are used to

[2] We carefully omit the distinction between online and offline distillation steps as we believe that our work applies to both the online and offline distillation methods. We leave it as an open question for further exploration.

train the student network [8, 1]. The soft labels are the response from the teacher network that contains rich similarity information between the classes that helps the student model to mimic the behavior of the teacher model [8]. Similarity information is a probabilistic values of the correct class being similar to each incorrect class as understood by the teacher network. This method is very effective in distilling the knowledge of a cumbersome teacher model into a more efficient student model.

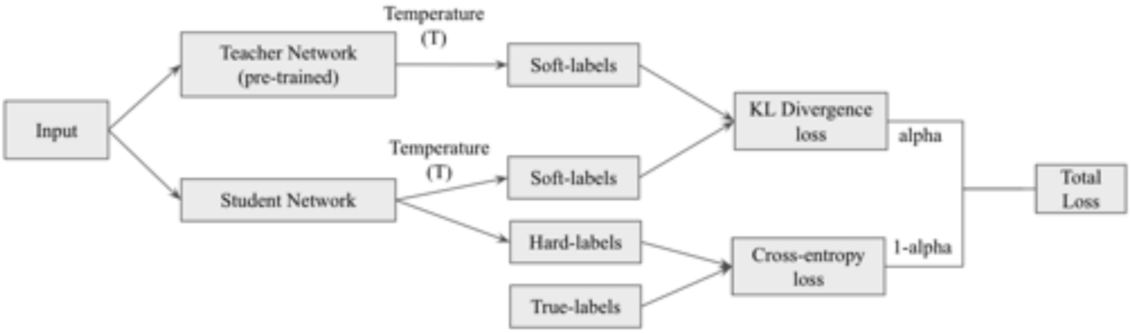


Figure 1.2: The figure shows the network architecture for the response-based knowledge distillation process.

Recent advances [26, 15, 5] suggest that the performance of the RBKD process is affected by the gap in learning capacities of the teacher and student models. As the gap between the learning capacities of the student and teacher model increases, the performance of the student model degrades on distillation [26, 15, 5]. Many methods have been proposed to address this problem, like improving distillation via teacher assistant [15], residual knowledge distillation [5], teacher-free knowledge distillation [26]. These works [26, 15, 5] highlights the problem and then skip to finding a solution without analyzing the root cause of the problem. In this work, we focus on analyzing the root cause(s) of the performance degradation.

The degradation in distillation performance has largely been attributed to the gap in learning capacities of teacher and student models in literature [15]. But our experiments suggest that this is not the only factor. We argue that for a fixed student network, as the teacher model becomes more and more powerful (wider and deeper in size) it loses the rich similarity information present in the soft-labels and behaves like a label smoothing

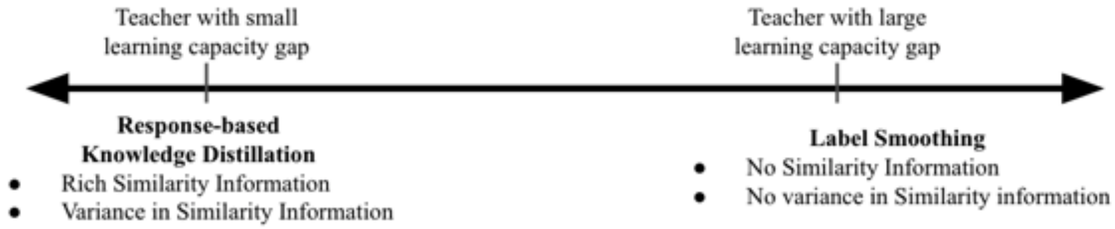

Teacher with small learning capacity gap
Teacher with large learning capacity gap
Response-based Knowledge Distillation
• Rich Similarity Information
• Variance in Similarity Information
Label Smoothing
• No Similarity Information
• No variance in Similarity information

Figure 1.3: The soft-label hypothesis: for a fixed capacity of student model, as the teacher model becomes more and more powerful (capacity of teacher increases along x-axis) the process of RBKD shifts from similarity information rich soft-labels based KD to no similarity containing less confident form of smooth label process of label smoothing (LS).

(LS). A label smoothing[3] process [16, 21] uses a less-confident form of one-hot vector as per equation 3.1. The absence of this rich similarity information (dark knowledge [8]) contributes to the degradation of the student model performance on distillation. Apart from analyzing the quality of soft-labels, we also propose some special consideration for the pre-training step of a teacher model to retain the rich similarity information in its response. Good quality of response from teacher model significantly boosts the performance of a student model on distillation.

Our contributions are as follows:

1. We show that as the teacher model becomes more powerful to learn the data distribution more precisely, it becomes more and more confident and loses the rich similarity information in its response. We demonstrate that this loss of rich similarity information is a contributing factor in the poor distillation performance of a large capacity teacher model.

2. We demonstrate that the rich similarity information in soft-labels facilitate **one-example-to-many-classes learning** [4] and significantly reduces the number of training examples needed per class to distill the knowledge. We also show that the absence of rich similarity information in response leads to **one-example-to-one-class learning** and requires more examples per class to achieve similar performance.

[3] Explained in section 3.3.1

[4] One input example carries relative information of more than one class for training. It is explained in chapter 3 figure 3.1

3. We propose **the soft-label hypothesis** (Figure 1.3) for RBKD process and experimentally show that as the teacher model becomes more capable of learning the data distribution, the process of knowledge distillation moves from being a response-based knowledge distillation (RBKD), with rich similarity information, towards a label smoothing (LS) [16, 21] process, with no similarity information (Figure 1.3).

4. We also propose some special consideration for the pre-training steps of a teacher to retain the rich similarity information in its response. We argue that any teacher model can be trained to retain the rich similarity information in its response by finding the right balance among these three factors:

    - Batch size with which teacher model is pre-trained
    - The number of epochs the teacher model is pre-trained
    - Student-Teacher learning capacity gap

vkvats: Organization of rest of the document

CHAPTER 2

RELATED WORK

With the beginning of the deep learning era, wider and deeper DL models has helped us build more and more sophisticated systems across a broad range of areas [14]. But these cumbersome DL models that can produce state-of-the-art performance have huge computation and memory requirements. This lead to a new line of research to compress the large cumbersome DL models into computationally efficient models with comparable performance [1, 8, 7, 22, 23, 3, 25, 2, 27]. Knowledge distillation, more precisely response-based knowledge distillation (RBKD), is one of the many interesting lines of research for compressing the cumbersome models without significantly affecting the performance [1, 8, 15, 5, 16, 26, 21].

Before the deep learning era, Buciluă et al. [1] devised a method to compress the knowledge of an ensemble of networks into a single model. They used the input to the last layer, generally a softmax layer, of an ensemble of networks to train a more efficient model. In 2015, Hinton et al. [8] popularized this idea of response-based knowledge distillation by modifying the softmax layer. They use a temperature-raised response of the output layer to distill the knowledge to a computationally efficient model. The output layer of a neural network typically produces class probabilities (q_i) by passing the logits (z_i) through a softmax activation function raised at temperature (T) to control the softness of response, shown in equation 2.1. Changing the value of T in equation 2.1 directly affects the response of the teacher model. With the increase in temperature, the correct class probability value decreases, indicating a decrease in confidence of the model, and the incorrect class probabilities increase, indicating an increase in similarity information in response. The response from the teacher model becomes softer and more information-rich at higher values of temperature T [8].

$$q_i = \frac{exp(\frac{z_i}{T})}{\sum_j exp(\frac{z_j}{T})} \tag{2.1}$$

This method of compressing a cumbersome model is effectively used in all major areas of application, like Computer Vision, Natural Language Processing (NLP), and Speech Recognition. Researchers explain the effectiveness of this method in many different ways, like Kim et al. [11] and Müller et al. [16] ascribe the effectiveness of RBKD being similar to the effectiveness of label smoothing (LS). Ding et al. [4] explain RBKD as the regularization effect brought about by the soft-labels of the teacher model. Yuan et al. [26] stress that the soft-labels from a teacher is more of a regularization term than the similarity information between categories. However, the RBKD process relies completely on the output layer response of the teacher model and it fails to properly explain the hidden-layer supervision from the teacher model [6]. There is no clear consensus about the effects of RBKD process. In this work, we explain the intricacies of RBKD process and tie the loose ends in the literature.

Recent studies have shown that the teacher model with different learning capacity[1] performs differently on distillation [15, 26, 5]. A large learning capacity teacher does poor distillation as compared to a small learning capacity teacher. The distillation performance by a teacher model degrades with the increase in its learning capacity for a given student model[15, 26, 5]. This is a counterintuitive result. A common understanding is that as the teacher model becomes more powerful to understand the data distribution, it should do better RBKD to a student. Researchers explain this outcome very differently. Mirzadeh et al. [15] attribute the gap in learning capacities between a teacher and a student as the only reason for the degradation in performance on distillation, and propose a teacher assistant model to address this problem. A teacher assistant model has a learning capacity more than

[1] the term large learning capacity is loosely used for a wider and deeper network with a large number of parameters. It is used in comparison with student learning capacity as reference.

the student but less than the teacher. The knowledge from the teacher is first distilled to the teacher assistant model, which is then used as a pseudo teacher to distill the knowledge to the student model. The original teacher is not used directly to distill the knowledge. This process can have more than one teacher assistant model and the knowledge from the teacher is distilled to a student via all these teacher assistants in a sequential manner. Though this method provides a way to overcome the problem the whole process becomes computationally more expensive with one or more teacher assistants which are orthogonal to the original problem, model compression, it is trying to solve. Mirzadeh et al. [15] propose that reducing the gap in learning capacity improves the distillation performance but they neither explain how reducing the learning capacity gap improves the model nor talks about the changes the teacher assistant model brings to boost the distillation performance. We answer both these questions in this work.

Yuan et al. [26] propose a teacher-free knowledge distillation (TFKD) process to address the gap in learning capacities between a student and a teacher model. They also fail to explain the reason for the performance degradation and propose a new method to reduce the learning capacity gap between a teacher and a student. The TFKD process distills its knowledge to itself during training (self-distillation) instead of using soft labels from a pre-trained teacher model. Theoretically, TFKD reduces the gap in learning capacity between a teacher and a student model to zero, but its knowledge is limited by its learning capacity and amount of training. It does not have much knowledge in itself to distill. Though it can be put forward as a solution, it deviates from the original problem of compressing the knowledge of a cumbersome DL model into a more efficient model.

Gao et al. [5] propose residual knowledge distillation (RKD) to distill the knowledge by introducing an assistant model. During distillation, the student model learns the feature map of the teacher model and the assistant model learns the residual error between them. While this method improves the performance on distillation, but it does not look into the details of the problem causing this degradation. They formulate their solution completely on the explanations of Mirzadeh et al. [15] and Yuan et al. [26]. They also introduce an additional assistant model in this process which adds to the complexity of the whole process.

We focus on finding an explanation of the performance degradation by looking into the changes in the response from the teacher. We find that the degradation in the performance of the student model on distillation is not only caused by the gap in learning capacities but also by the quality of response from the teacher. We argue that it is caused by two main factors; the absence of similarity information in soft labels and the gap in learning capacity between a teacher and a student model. We keep our focus on providing a careful explanation of the role played by rich similarity information in soft labels and its effects on the RBKD process. We find a connection between the RBKD and LS process and propose **the soft-label hypothesis** (see figure 1.3 and 3.2) to join the missing links in literature. We argue that with the increase in learning capacity of a teacher model, the distillation process starts to move away from RBKD process and towards label smoothing (LS) process. This is observed through the loss of similarity information in the response from the teacher. This hypothesis also explains the role of the teacher assistant model [15] and the teacher-free knowledge distillation [26] in improving the distillation performance. To enable better distillation performance by large capacity teacher models, we also propose a method to retain the similarity information in soft-labels and reverse the performance degradation.

CHAPTER 3
THEORY OF RESPONSE-BASED DISTILLATION

3.1 Overview

In this chapter, we discuss the theory of response-based knowledge distillation (RBKD) process, label smoothing (LS) process and their relation. We start with describing the LS and the RBKD processes and mathematical analysis of their loss functions in section 3.3.1 and 3.3.2, respectively. Then, we draw some high level parallels between RBKD and LS processes and discuss their theoretical understanding in section 3.3.3. In section 3.4, we discuss a detailed analysis of soft-labels and its essential components for good distillation. Towards the end of this chapter, we conceptualise the ideas used in experiments and propose *the soft-label hypothesis* to tie the loose ends of the RBKD and LS methods.

3.2 Important Terminologies

1. Similarity information 2. Soft-labels 3. Response based knowledge distillation process 4. Label smoothing process. 5. Dark Knowledge 6. Confidence labels 7. Similarity labels 8. Variance in soft-labels

3.3 Label Smoothing and Response-Based Knowledge Distillation

3.3.1 Label Smoothing

The LS process produces less confident form of one-hot label. It has two important components U and V. U is the true labels, V is a given distribution, and α_{LS} as a smoothing factor. The generalized equation for LS process is shown in 3.1.

$$q_i^{'} = (1 - \alpha_{LS}) * U_i + \alpha_{LS} * V_i \tag{3.1}$$

The LS process produces softened labels as per equation 3.1, where the distribution U is the one-hot vector and V is a uniform distribution generated by $1/c$, c is the number of classes in the dataset as per equation 3.2. The generated labels are shown in table 3.1.

$$y_c^{LS} = (1 - \alpha_{LS})\ y_c + \alpha_{LS}\ \frac{1}{C} \tag{3.2}$$

We denote the true soft-label distribution $q(c|x)$ (x is the input) as $q^{'}$ for LS process, q for RBKD process and use p to denote distribution $p(c|x)$ generated by model during training. As shown by Yuan et al. in [26], the cross-entropy calculated over these smooth labels are given as:

$$\begin{aligned} H(q^{'}, p) &= -\sum_{c=1}^{c} q^{'}\ log(p) \\ &= (1 - \alpha_{LS})\ H(q,p) + \alpha_{LS}\ H(u,p) \\ &= (1 - \alpha_{LS})\ H(q,p) + \alpha_{LS}\ (D_{KLdivergence}(u,p) + H(u)) \end{aligned} \tag{3.3}$$

where D_{KL} is the Kullback-Leibler divergence (KL divergence) and $H(u)$ is a fixed entropy value for uniform distribution. Thus,we can write the loss function of LS process as:

$$\mathcal{L}_{LS} = (1 - \alpha_{LS})\ H(q,p) + \alpha_{LS}\ D_{KLdivergence}(u,p) \tag{3.4}$$

Table 3.1: Soft-labels for LS process generated with $\alpha_{LS} = 0.6$ as per equation 3.1 (where U is one-hot vector and V is uniform distribution) for 10 classes of MNIST

Input class	Soft-labels for label smoothing process									
	0	1	2	3	4	5	6	7	8	9
	0.06	0.06	0.06	0.06	0.06	0.06	**0.46**	0.06	0.06	0.06
	0.06	0.06	0.06	0.06	0.06	0.06	**0.46**	0.06	0.06	0.06
Digit 6	0.06	0.06	0.06	0.06	0.06	0.06	**0.46**	0.06	0.06	0.06
	0.06	0.06	0.06	0.06	0.06	0.06	**0.46**	0.06	0.06	0.06
	0.06	0.06	0.06	0.06	0.06	0.06	**0.46**	0.06	0.06	0.06

A closer look at the soft-labels generated by the LS process, in table 3.1, reveals that this process makes the model less confident about the correct class and provides no similarity information about other classes [16, 21]. The absence of rich similarity information in soft-labels promotes **one-example-to-one-class learning** training paradigm as shown in figure 3.1(b). In this learning paradigm, each example provides useful information only about its class and treats all other classes equally.

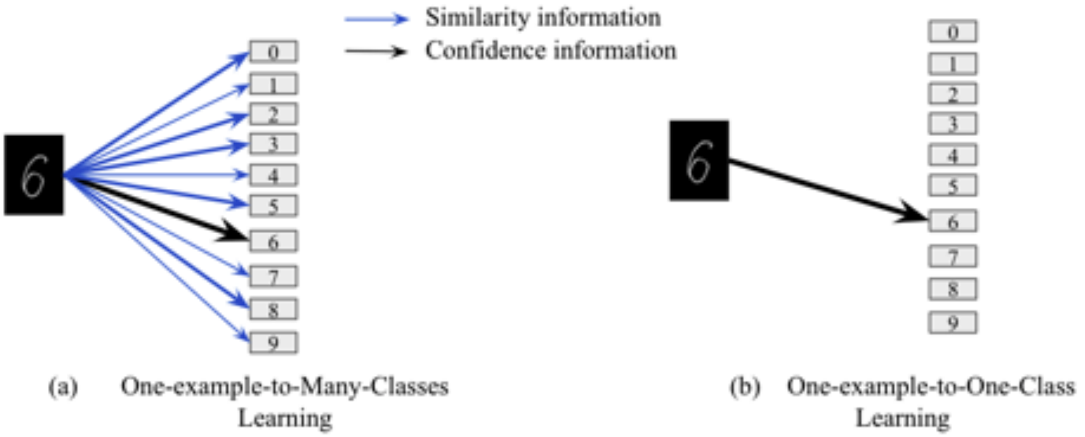

Similarity information
Confidence information
6
0
1
2
3
4
5
6
7
8
9
(a) One-example-to-Many-Classes Learning
6
0
1
2
3
4
5
6
7
8
9
(b) One-example-to-One-Class Learning

Figure 3.1: There are two basic paradigms of learning, first shown in (a) *one-examples-to-many-classes learning*, in which one example from a class provides similarity-based information about all other classes as well as information about its class. The thickness of the arrow shows a high value as per the first row of table 3.2. Second is shown in (b) *one-examples-to-one-class learning*, in which each example provides information only about its class and no similarity-based information is shared for other classes.

3.3.2 Response-Based Knowledge Distillation

As discussed earlier, the RBKD process modifies the output of the last layer, the softmax layer, of a neural network as per equation 2.1 to produce probabilistic distribution, soft-labels, for each class at a temperature T. These soft-labels contain rich similarity-based information between classes to support **one-examples-to-many-classes learning** paradigm [8]. The pictorial depiction of one-examples-to-many-classes learning is shown in 3.1(a). These information-rich soft-labels are used to improve the performance of the student model by minimizing the weighted sum of Kullback-Leibler divergence (KL divergence) and cross-entropy losses as per equation 3.5.

$$\mathcal{L}_{KD} = (1 - \alpha_{KD})\ H(p, q) + \alpha_{KD}\ D_{KLdivergence}(p^t_T, p_T) \tag{3.5}$$

where $H(p, q)$, p, p_T, and p^t_T denotes the cross-entropy loss between student prediction and true labels (q), output prediction of student model, output prediction of student model softened at temperature T and output prediction of teacher model softened at temperature T, respectively. α_{KD} decides the contribution of each loss in the overall loss function $\mathcal{L}_{KD}$.

From equation 3.4 and equation 3.5 , we learn that the loss functions of RBKD and LS processes have similar formulation but they use different methods for generating soft-labels. In these two equations, the only difference is the p_T^t in $D_{KLdivergence}(p_T^t, p_T)$, which is generated by a pre-trained teacher model and u in $D_{KLdivergence}(u, p)$, which is a uniform distribution [26]. From this comparison, we can conclude that LS is a special case of RBKD process in which soft-labels are generated as per some pre-defined distribution as prior knowledge instead of a pre-trained teacher model as a source of prior knowledge.

We have understood the theoretical similarity and differences between the RBKD and LS processes. This theoretical understanding is corroborated by the mathematical resemblance of the loss function of these processes. But there is a significant difference between the soft labels generated by these two processes as discussed earlier. While the soft-labels generated in the LS process are easy to understand because of the known behavior of the pre-defined distribution, the soft-labels of the RBKD process changes with the change in any hyper-parameters, batch size, number of epochs during pre-training, of a teacher model. The quality of these soft labels is also dependent upon the learning capacity of the teacher model. We discuss these in the next section.

3.3.3 Parallels and Differences

We can draw parallels between RBKD and LS processes at a high level of abstraction. Keeping our attention on the effects of these processes and not focusing on its values or method of its generation, both RBKD and LS convert high-confidence labels, one-hot vector, into less-confident smooth labels, smooth form of one-hot labels. Both these processes bring about the regularization effect and improve the generalization of the model [16, 4, 26]. But if we focus on its numerical values and method of its generation, both these processes become very different. The LS process smooths the high-confidence one-hot labels according to a given distribution, like Gaussian distribution, as per the equation 3.1. The soft labels of this process are not much different from each other, it has only one high-value label of the correct class and all other labels are small in magnitude as per the distribution V.

As shown in equation 3.1, the LS labels are a mixture of one-hot vector U and uniform distribution V with α_{LS} as a smoothing factor. But this is not the case for soft-labels generated in the RBKD process. This process uses a pre-trained teacher network to generate less-confident smooth labels. The values generated by it are quite different from one another. Even for the examples belonging to the same class, the soft-label generated with the RBKD process is quite different as shown in table 3.2. Their values are based on the similarity information captured by the teacher model during its pre-training. In this process, the confidence removed from the most confident class, 1 of the one-hot vector, is distributed to other classes as per the similarity relation learned by the teacher model. This is very different from the labels generated with the LS process. In this process, the confidence removed from the most confident class, 1 of the one-hot vector, is distributed to other classes as per distribution V in equation 3.1.

3.4 Understanding Soft-labels: Confidence, Similarity and Variance in Similarity labels

Majority of contribution in the loss function of RBKD is by $D_{KLdivergence}(p_T^t, p_T)$ as the α_{KD} value is kept as high as 0.99 (from equation 3.5). This contributes soft-labels (p_T^t) by the teacher model very significant in the whole process. The student model tries to mimic the probabilistic distribution, soft-labels, of the teacher model by minimizing the KL divergence of its softened response with the response of the teacher at the same temperature T. This makes it important to have the best quality of soft labels suitable for distillation in the RBKD process. With a clear understanding of the methods used for generating the soft-labels in both RBKD and LS processes, now, it is important to understand how these methods affect the desirable quality of soft-labels for good knowledge distillation and how the absence of these desirable qualities make the RBKD process behave like a LS process.

Table 3.2: Soft-labels by small learning capacity teacher model with 3 hidden layers on MNIST data

Input class	Soft-labels by small learning capacity teacher									
	0	1	2	3	4	5	6	7	8	9
	0.087	0.048	0.095	0.093	0.077	0.144	**0.206**	0.051	0.122	0.078
	0.087	0.048	0.089	0.1	0.090	0.119	**0.177**	0.056	0.114	0.095
Digit 6	0.090	0.078	0.089	0.097	0.115	0.091	**0.179**	0.071	0.108	0.082
	0.107	0.068	0.089	0.076	0.104	0.1	**0.229**	0.070	0.086	0.071
	0.118	0.079	0.095	0.075	0.101	0.081	**0.210**	0.069	0.098	0.073

In the soft-label generation process, each example from a class generates soft labels for every other class including its class. The soft-labels generated by an example for its class can be termed as *same class soft-labels* and the soft-labels generated by the same example for other classes can be called *different class soft-labels*. Same class soft-labels represent the confidence of the teacher model incorrectly classifying the examples at a given temperature T and can be termed **confidence label**. The *different class soft-labels* represents the similarity information between the correct class and the remaining incorrect classes based on the understanding of the teacher model, at a temperature T, and can be termed as **similarity labels**. Table 3.1, 3.2, and 3.3 show the confidence label in bold text and similarity labels in plain text. The hyper-parameter $\alpha_{LS} = 0.6$ for LS process in table 3.1, temperature $T = 9$ for RBKD process in table 3.2, and 3.3.

Table 3.3: Soft-labels generated by a large learning capacity teacher model with 6 hidden layers on MNIST data

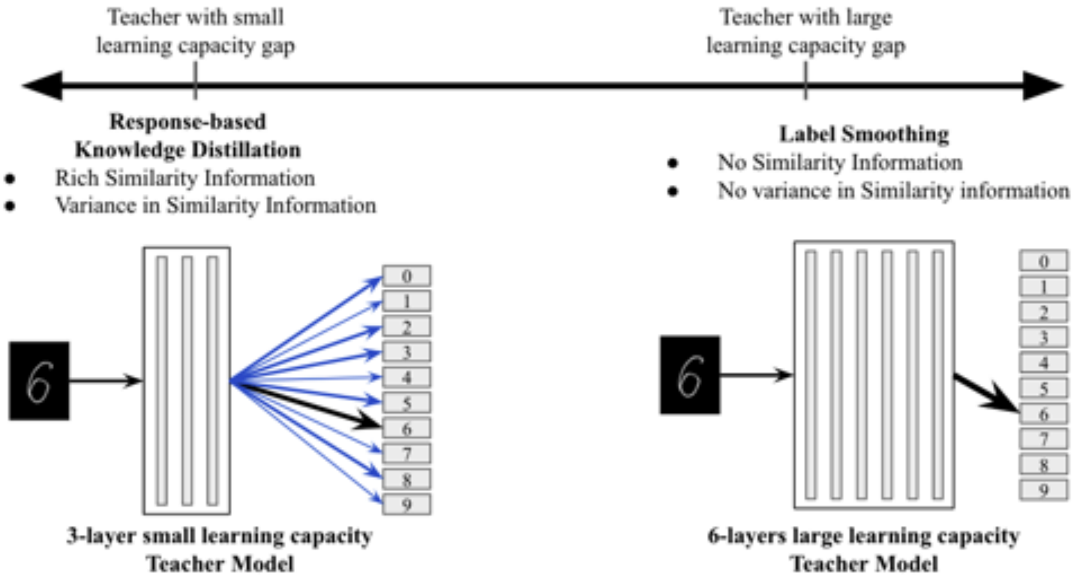
Input class	Soft-labels by large learning capacity teacher									
	0	1	2	3	4	5	6	7	8	9
	0.078	0.069	0.063	0.059	0.070	0.135	**0.356**	0.039	0.078	0.053
	0.083	0.077	0.073	0.056	0.078	0.107	**0.339**	0.042	0.090	0.053
Digit 6	0.077	0.067	0.068	0.047	0.073	0.086	**0.425**	0.034	0.079	0.044
	0.076	0.062	0.057	0.039	0.059	0.090	**0.485**	0.027	0.065	0.039
	0.077	0.067	0.061	0.043	0.068	0.089	**0.450**	0.031	0.075	0.041

For the LS process, the value of confidence labels and similarity labels changes with the smoothing factor α_{LS}. With the increase in smoothing factor, the value of the confidence label increases and the value of similarity labels decreases. The Sum of all soft-labels values is always equal to one. It is important to note that the similarity labels in LS process is a uniform distribution for our experiments and the same value is assigned to all *different class soft-labels*. In the RBKD process, the value of confidence labels and similarity labels vary with the temperature T. As the temperature, T is increased the value of the confidence label decreases, and the value of similarity labels increases. It is also important to note that, unlike the LS process, the increase or decrease in similarity labels are not uniform in the RBKD process, rather it is dependent on the similarity relation learned by the teacher model during its pre-training.

At a given temperature T, a teacher model with a large learning capacity, high confidence about the data distribution, assigns relatively higher value to *confidence labels* and smaller values to *similarity labels*. But at the same temperature T, another teacher model with smaller learning capacity, less confident about data distribution, assigns relatively smaller value to *confidence labels* and higher value to *similarity labels*. It should be noted that different examples belonging to the same class get different confidence and similarity label values. The confidence label and similarity labels generated by each example are dependent only on the knowledge of the teacher model. With better knowledge of the data distribution, a teacher model produces less confusing soft-labels by assigning a high value to confidence labels, but a less knowledgeable teacher model produces more confusing soft-labels by assigning higher values to similarity labels for the same example. Each example gets a *confidence label* based on the confidence of the teacher model to correctly classify it and *similarity labels* as per similarity perceived by the teacher model between the correct class and other classes. This knowledge dependent variation in similarity labels for different classes of the same example, row-wise in table 3.2, and for the same class of different examples, column-wise in table 3.2, can be termed as **variance in similarity labels**. The variance in similarity labels can be observed in table 3.2 and table 3.3 on MNIST data set. Both the tables were generated by two different learning capacity teacher models, large and small, using the same input images of digit 6 from the MNIST dataset at $T = 9$.

With a clear understanding of confidence labels, similarity labels, variance in similarity labels, different learning capacities of teacher models, and their effects, we present our hypothesis, called **the soft-label hypothesis**, in the next section. This hypothesis connects all these variables for a better understanding of both the RBKD and LS processes.

3.5 The Soft-label Hypothesis



Teacher with small learning capacity gap
Teacher with large learning capacity gap
Response-based Knowledge Distillation
Rich Similarity Information
Variance in Similarity Information
Label Smoothing
No Similarity Information
No variance in Similarity information
6
0
1
2
3
4
5
6
7
8
9
6
0
1
2
3
4
5
6
7
8
9
3-layer small learning capacity Teacher Model
6-layers large learning capacity Teacher Model

Figure 3.2: The soft-label hypothesis: for a fixed learning capacity of student model, as the learning capacity of teacher model is increased (learning capacity increases along positive x-axis) the process of response-based knowledge distillation (RBKD) shifts from a similarity-information rich soft-labels based process of knowledge distillation (KD) to a no-similarity containing less confident form of one-hot label based process of label smoothing (LS). The size of model is shown relative to MNIST dataset.

The soft-label hypothesis is shown in figure 3.2, it can also be called a RBKD-LS hypothesis as it establishes a relation between the RBKD and LS methods. The hypothesis states that *for a given student model, as the learning capacity of the teacher model is increased, learning capacity increases along the positive x-axis in figure 3.2, the process of response-based knowledge distillation (RBKD) starts to move away from a similarity-information rich soft-labels based distillation process to a no-similarity containing a less* *confident form of the one-hot label based label smoothing process*. In other words, as the learning capacity of the teacher model is increased, the distillation process shifts away from *one-example-to-many-class learning* paradigm to *one-example-to-one-class learning* paradigm.

A closer look at the soft-labels generated by large learning capacity teacher, table 3.2, and by small learning capacity teacher, table 3.3, reveal that the soft-labels in table 3.2 has more row-wise and column-wise variance in similarity labels as compared to the soft-labels in table 3.3. With a high confidence label, table 3.3 has less confusing soft-labels as compared to table 3.2 that has more confusing labels. It should be carefully noted that the soft-labels assigned to the same class by the same teacher, see columns in table 3.2, and the soft-label assigned to different classes for same examples, see rows in table 3.2, are significantly different from each other. This variation within the same class, column-wise, and for different classes, row-wise, is very crucial information for better distillation of knowledge. This variation in soft-labels, column-wise and row-wise, diminishes with the increase in the learning capacity of teacher models and directly affects its distillation performance.

In chapter 5, we present more evidence to support the soft-label hypothesis. We also show that the best result on distillation is achieved by using a teacher model with a moderate understanding of the data distribution that produces moderately confused soft labels. As the quality of distillation depends on the quality of soft labels, we also propose special considerations during teacher model pre-training to retain the quality of soft labels for any capacity of teacher model in chapter 6. But first, we discuss the details of the experimental setup in the next chapter.

CHAPTER 4
THE EXPERIMENTAL SETUP

4.1 Overview

In this chapter, we discuss various experiments used to provide evidence for the soft-label hypothesis. We start with the details of the data sets used for this work in section 4.2. Then, we discuss the rationale behind selecting different teacher and student models in section 4.3. Building on these critical details, we proceed to discuss various experiments to provide evidence in support of *the soft-label hypothesis* in the latter part of this chapter. Some of these experiments has been previously used in similar context, like similarity information in soft-labels [8] (section: 4.4), relation between similarity information and entropy [8] (section: 4.4.1), entropy-based transfer set [1] selection [8] (section: 4.6) and penultimate layer representations [16] (section: 4.7). We also track change in entropy of soft-labels with change in temperature in section 4.5 and externally induce variance in soft-labels by adding noises in section 4.8. We try to explain the rationale behind these experiments in this chapter and present their results in the next chapter.

4.2 The Datasets

We test the soft-label hypothesis on different datasets with varying level of difficulty, Modified National Institute of Standards and Technology (MNIST) [13], Fashion - MNIST (F-MNIST) [24] and Canadian Institute For Advanced Research (CIFAR-10) [12]. We describe these data set in this section.

[1] A transfer set is the set of examples that is used to distill the knowledge from teacher(s) to a student model. It can be a subset of training set, test set or a combination of training and test sets

4.2.1 MNIST

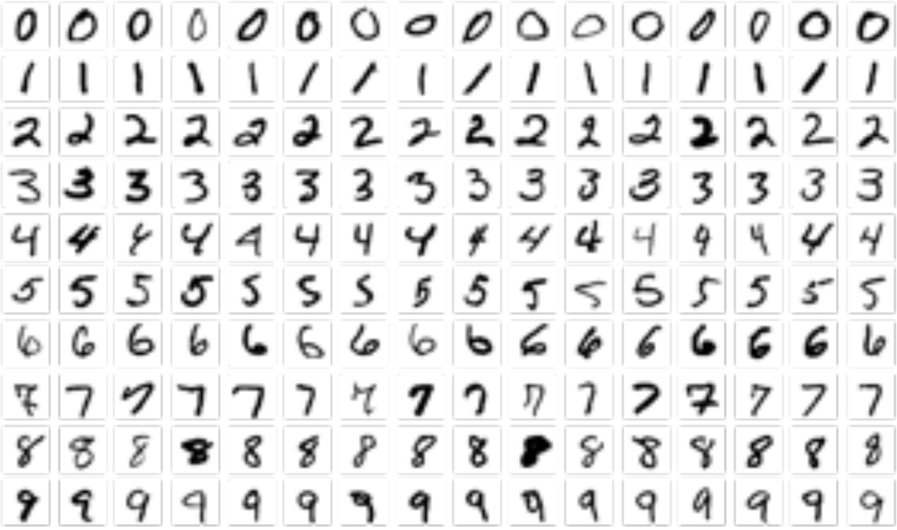


Figure 4.1: The figure shows examples from the MNIST dataset. Each row corresponds to one examples from one class. It shows the natural variation in writing same digits. Capturing this variation is important for our experiments

MNIST is a large dataset of handwritten digits. It has a training set of 60,000 examples and a test set of 10,000 examples associated with 10 different classes. The 10 different classes represent each digit from 0 to 9 as shown in figure 4.1. The digits in this dataset have been size-normalized and centered in a 28x28 size image. All the examples are grayscale. The training and test set examples are taken from a different special database and contain examples from approximately 250 writers, this brings a natural variation in handwriting into digit representation for this dataset [13]. This natural variation is important for our experiments and our hypothesis. This variation in the writing style of 250 writers forces the network to learn different ways of writing each digit and find differences and similarities between its classes. It is a natural variation [2] and it brings variance in similarity labels (refer section: 3.4) as understood by the pre-trained teacher model, this plays a very significant role in the RBKD process.

[2] This variation can be found in all image dataset as a result of many factors that naturally vary while snapping photos.

4.2.2 Fashion-MNIST

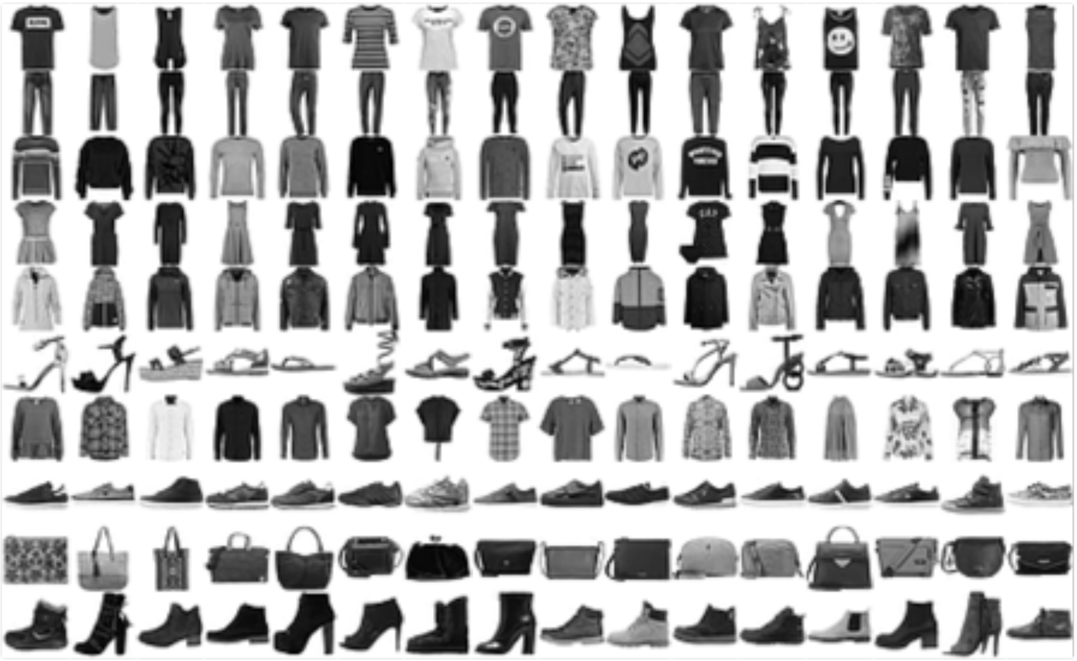


Figure 4.2: The figure shows the examples of the F-MNIST dataset. Each row shows examples from same class. It can be seen that there is a variation in examples by design and size. This variation plays a very important role in the distillation process.

F-MNIST is a dataset of Zalando's article images. It was designed based on MNIST and shares the same image size and structure of training and testing splits. It is considered slightly difficult than the MNIST dataset. It has 60,000 training and 10,000 test examples associated with 10 class labels. The images are a grayscale of size 28x28. The 10 different classes represent T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot as shown in figure 4.2. All these classes are mutually exclusive. The curators of the F-MNIST data set intend to serve it as a direct drop-in replacement for the original MNIST dataset [13] for benchmarking machine learning algorithms [24]. Just like MNIST, this dataset has also the natural variation within the class because of the variation in designs and size of items.

4.2.3 CIFAR-10

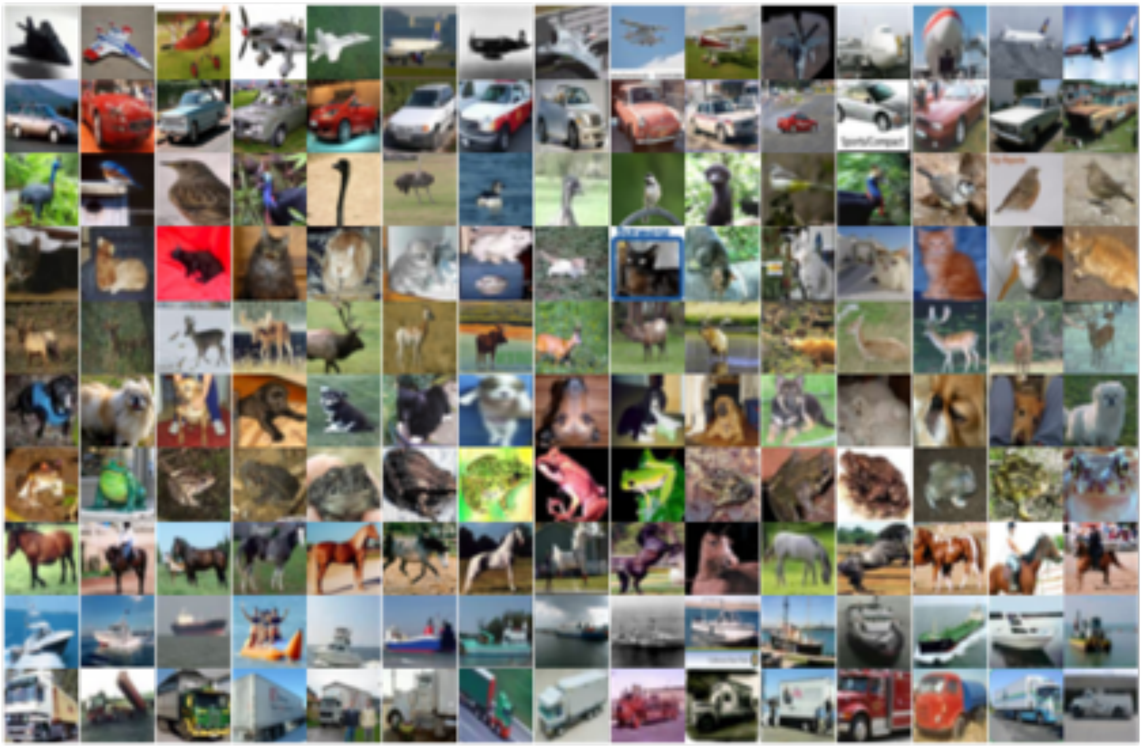


Figure 4.3: The figure shows some random examples of the CIFAR-10 dataset [12]. Each row represents examples from same class. This dataset also has a natural variation like any other dataset of natural images.

We also wanted to test our hypothesis on a dataset containing natural color images. CIFAR-10 is one of the most widely used color image datasets. It is the most difficult dataset in our experiments. It has 10 different classes and comprise of 60,000, 32x32x3 shape images, with 6000 images per class [12]. It has 50,000 training images and 10,000 test images. The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks as shown in figure 4.3. Like any other data of natural images, this data also has natural variation. All these classes are completely mutually exclusive and there is no overlap between automobiles and trucks. The automobile class includes sedans, SUVs, and similar things. Truck class includes only big trucks [12].

4.3 Teacher and Student Models

The teacher models are cumbersome models or an ensemble of many models with a large learning capacity to understand the data distribution. A large capacity well-trained teacher model can classify any given example with very high confidence. But the large computation and memory requirements lead to many problems. The knowledge from this model is distilled to a smaller computationally more efficient model through the RBKD process. We vary the capacity of teacher models to test our hypothesis. In our experiments, we vary the learning capacity of the teacher model in two ways, first, we use two different capacity teacher models with a significant difference in learning capacities within each dataset, and second, we vary the learning capacities of a teacher with the complexity of data distribution. MNIST, being the least difficult dataset, use the least powerful teacher model as compared to teacher models of Fashion-MNIST and CIFAR-10 datasets. CIFAR-10, being the most difficult dataset, uses the most powerful teacher model.

The soft-label hypothesis requires us to vary the learning capacities of teacher models within a dataset. We vary the learning capacity of teachers by changing the number of hidden layers, the number of filters in convolutional layers, and the number of units in fully connected dense layers. These variations effectively change the number of trainable parameters in the model. We also use concepts like dropout [20], and batch normalization [10] to make model generalize better and train faster. We assume that a deeper model, with more number of trainable parameters, is more capable of learning as compared to a shallow model. We also keep the student model less powerful than any variations of teacher used in our experiment under the same criteria. We discuss the details of all teacher models used in our experiments for each dataset in appendix A.

The student models are the more efficient models and are relatively less complex in all criteria as compared with any teacher model. Ideally, the student models should have a much smaller learning capacity than any teacher. We use the same criteria for measuring the learning capacity of a student model as described above for teacher models. We assume that a model which is shallower and has a smaller number of trainable parameters has lesser learning capacity. Throughout the experiments, we choose a student model with a much smaller learning capacity as compared to any of the teacher models. We provide specific details of each student model used for each dataset in appendix B.

4.4 Similarity Information in Soft-labels

Similarity information in soft-labels controls the quality of response from the teacher model. In the soft-label hypothesis, we argue that the presence and absence of similarity information in soft labels decide the behavior of the distillation process. In the presence of similarity information, the distillation process follows a one-example-to-many-classes learning paradigm, and in the absence of similarity information, it follows a one-examples-to-one-class learning paradigm.

We check the presence of similarity information based on a method developed by Hinton et al. [8]. In this method, we remove all examples of one or more classes from the transfer set and test the accuracy of the student model only on the examples of the removed class. For example, in the MNIST dataset, if we remove digit 6 from the transfer set then we will test the performance of the student model only on the examples of digit 6. So from the perspective of a student model, 6 is a mythical digit it has never seen. The only source of information about digit 6 is through the soft labels provided by the teacher model during the distillation process. If there is sufficient similarity information in the soft labels, then the student model will learn to correctly classify digit 6 even though it has never seen it during training. We repeat this process one by one for each class in our dataset and observe the accuracy of the model for each class. High accuracy on the removed class test set indicates the presence of similarity information in the soft labels and smaller accuracy points to near-absence of similarity information in the soft labels. We present these results on all three datasets in the next chapter.

4.4.1 Similarity Information and Entropy

The Shannon entropy [18] is a good indicator of presence of similarity information in soft-labels [8]. It is calculated using the equation 4.1

$$E_{Soft-labels} = -\sum_{i=1}^{C} p_i \; log \; (p_i) \tag{4.1}$$

where C is the number of classes in the dataset, p_i is the probabilistic value from teacher model, soft labels. The entropy of soft-labels is proportional to the presence of rich similarity information, a high value of entropy indicates the presence of similarity information and facilitates one-example-to-many-classes learning during the distillation process, whereas a small value of entropy indicates the absence of similarity information and facilitates one-example-to-one-class learning during the distillation process. Instead of using the removed class experiment proposed by Hinton et al. [8], we use entropy as an indicator of similarity information in the rest of the work.

4.5 Change in Entropy with Temperature

The entropy of the soft labels is an indicator of rich similarity information [8]. But the softness of the response from the teacher is controlled by temperature T in equation 2.1. The entropy of soft-labels increases with the temperature, thick black and blue lines in figure 4.4 indicates average entropy of soft-labels, for any model, but it does not always help the distillation process. As the temperature increases the value of *confidence labels* decreases and the values of *similarity labels* increases. This change can be beneficial at smaller values of temperature, but at a very high value of temperature, typically beyond 50, the soft-labels start to become more like a uniform distribution and are not useful for the RBKD process. We do not go deeper into this behavior of soft labels at very high temperatures and leave it for future work.

The figure 4.4 is a typical representation of the behavior of entropy with different val-

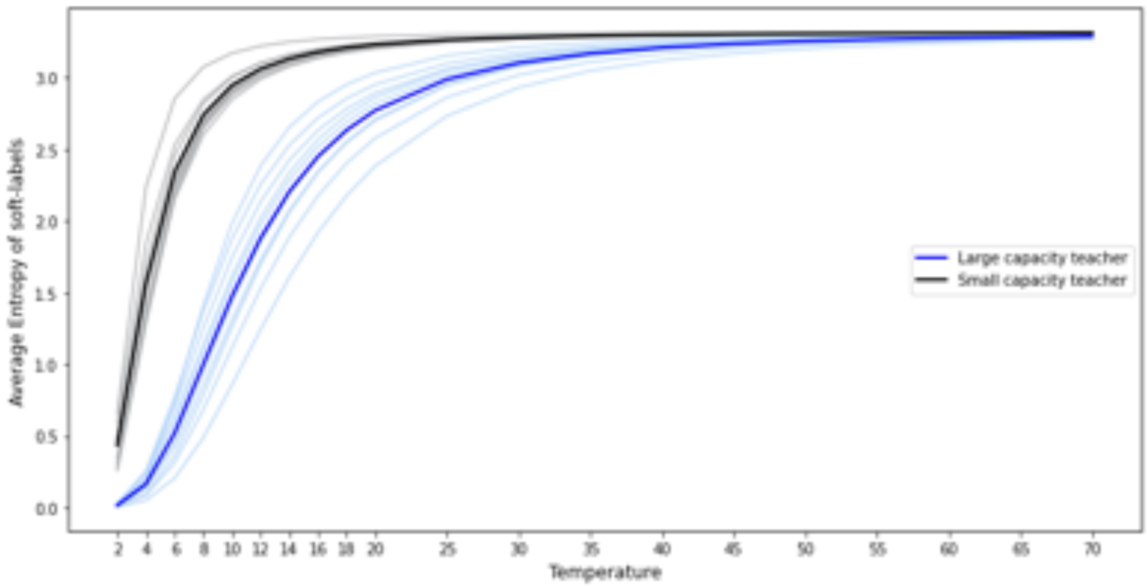


Figure 4.4: The figure shows line plot of average entropy of the soft-labels generated by two different teachers at different temperatures. The bold lines (black and blue) shows the mean line of the average entropy values.

ues of temperature. The faded black and blue lines represent entropy of the soft-labels generated by small and a large learning capacity teacher models, respectively. Each of these faded lines represents one class of the MNIST dataset. The thick black and blue lines represent the average of all the faded black and blue lines, respectively. For better visual presentation, we only show the thick lines in the rest of our work instead of one line for each class as shown in 4.4.

4.6 Entropy-Based Transfer-set Selection

The relation between similarity information in soft-labels and entropy can be exploited to select highly suitable examples for the transfer set. Hinton et al. [8] postulate that examples with high entropy in soft labels provide much more information per training case as compared to those examples with low entropy in soft labels. More information per training cases facilitates *one-example-to-many-classes* learning paradigm during distillation and requires lesser number of examples per class for the RBKD process. This idea also aligns with the intuitive our intuitive understanding. If one example provides information about itself and other classes, then we should require a fewer number of examples per class to achieve a good performance. Based on the entropy of soft-labels, we select a different number of examples, varying from 5 to 2000 examples per class, per class to exploit *one-example-to-many-classes* learning process to our advantage. We evaluate the quality of soft-labels and the performance of the distillation process based on the minimum number of examples required per class to achieve similar performance by both types of teacher models on a test set [3]. The performance on the test set directly relates to the RBKD process, lower the number of examples required for distillation, higher is the similarity information in soft-labels and better is the distillation of knowledge by the teacher model. We present the results of this experiment on all three datasets in the next chapter.

4.7 Penultimate Layer Representations

Müller [16] developed a visualization method to visually understand the effects of LS technique. This method helps in understanding the change in representations of the penultimate layer activations of a model by visualizing its linear projections. In the LS process, smooth labels are generated as per equation 3.2. These labels, see table 3.1, is a less-confident form of one-hot vectors generated as per some distribution V. Training with these labels encourages the differences between the logits of the correct and incorrect class to be dependent on α_{LS} [16]. A mathematical description of this can be given by writing the prediction of a neural network as a function of penultimate layer activations as shown in equation 4.2

$$p_c = \frac{e^{x^T \ w_c}}{\sum_{c=1}^{C} \ e^{x^T \ w_l}} \tag{4.2}$$

where, p_c is the likelihood of C^{th} class, w_k represents the weights and biases of the last layer and x is the penultimate layer activations. The logit $x^T \ \ w_c$ of the C^{th} class can be seen as the measure of the squared Euclidean distance between x and w_c. Mathematically,

$$||x - w_c||^2 = x^T \ x \ - \ 2x^T w_c \ + \ w_c^T \ w_c \tag{4.3}$$

[3] We use the predefined test set to record the performance of distillation

Each class has a template w_c, $x^T\ x$ is factored out when calculating the softmax output and $w_c^T\ w_c$ is usually constant across classes as shown in equation 4.3.

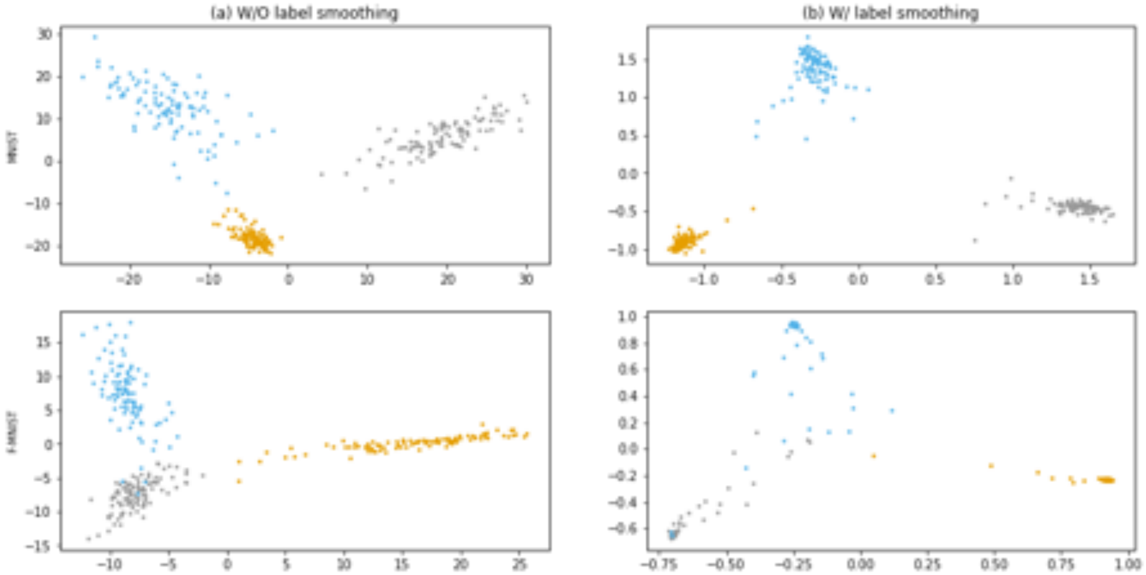


Figure 4.5: The figure shows the effect of training with LS process on MNIST and Fashion-MNIST. Column (a) shows penultimate layer activation clusters (spread clusters) for three different classes when trained without LS, and column (b) shows much tighter clusters for same examples when trained with LS process. For each dataset, the corresponding model were trained with same parameters and for same number of epochs, the only difference is once the model is trained without LS and then it is trained with LS

Therefore, it can be concluded that LS encourages the activations of the penultimate layer to be close to the template of the correct class and equally distant to the templates of the incorrect classes [16] as shown in figure 4.5. Using this visualization technique, the behavior of the RBKD process with different learning capacities of the teacher model can be identified by checking the compactness of the penultimate layer activation representation. We show this representation for smaller and large learning capacity teachers for each dataset in the next chapter.

4.8 Adding noise in soft-labels

We stressed on *variance in similarity labels* while describing the soft-labels in chapter 3. We argue that there should be some variance in similarity labels, row-wise and column-wise, and these variances should be a result of an understanding of the data distribution of a teacher model. We call this **natural variance** as it is completely dependent on the natural variance in the dataset understood by the teacher model. To further stress the importance of this natural variance, we artificially induce the variance by combining different type of noises with the soft-labels of the large capacity teacher models as well as for the LS method as shown in equation 4.4 below

$$\begin{aligned} y'_{soft-label} &= y^{Teacher}_{soft-label} + U \\ y'_{soft-label} &= y^{LS}_{soft-label} + U \end{aligned} \tag{4.4}$$

Where $y^{Teacher}_{soft-label}$ is soft-label generated by a teacher model, $y^{LS}_{soft-label}$ is soft-labels generated by LS process and U is a distribution, a uniform or normal distribution of different standard deviation. The soft labels are normalized after adding the noise U. This method of generating soft labels induces some induced variance but it does not bring any conducive effect for better distillation. Instead of helping in distillation, this way of inducing variance is soft-label brings more irregularity in the target distribution. Due to the random sampling and addition of values from distribution U, at each iteration, a different value is added to the same soft-label generated by the same example. In this way, the same example gets different target distribution at each epoch, which KL divergence is trying to minimize. This irregularity is repeated at each iteration and brings inconsistency in the target distribution during the optimization process. This does not help in providing the similarity information needed for distillation. This small experiment further supports our argument that the variance in soft-labels should be generated by a teacher model which has some understanding of the data distribution to achieve better distillation.

CHAPTER 5

THE EXPERIMENTAL RESULTS

We discuss the results of the experiments described in chapter 4 on three different datasets, MNIST [13], Fashion-MNIST [24] and CIFAR-10 [12]. The results provide empirical evidence for the soft-label hypothesis. We club the results across different dataset under each experiment to stress on the consistency and generalization of the hypothesis.

5.1 Similarity Information in soft-labels

We perform this experiment as described in section 4.4 on all three datasets. Each teacher and student model is different in its learning capacity for different datasets. The learning capacities of these models are based on the general understanding of the complexity of data distribution. The details of teacher and student models are discussed in appendix A and B, respectively.

Table 5.1: Student accuracy on missing class (MNIST) when distilled with small learning capacity teacher

Missing class	Student accuracy(%) on missing classes at temperature(T)						
	T=3	T=6	T=9	T=12	T=15	T=18	T=20
Digit 0	99.18	99.28	99.48	99.18	98.67	98.87	98.57
Digit 1	98.67	98.67	98.67	98.59	98.50	98.32	98.23
Digit 2	95.73	96.80	96.12	95.83	94.86	94.76	94.76
Digit 3	98.11	98.41	98.31	98.11	98.01	97.62	97.92
Digit 4	98.47	98.67	98.57	98.67	98.37	97.96	96.94
Digit 5	96.86	97.53	97.30	97.19	96.41	96.86	97.86
Digit 6	97.39	97.49	97.28	97.18	97.18	96.76	96.65
Digit 7	96.10	96.30	96.40	95.91	94.84	94.84	95.03
Digit 8	97.12	96.71	97.22	96.61	96.40	95.58	96.09
Digit 9	96.13	96.63	96.23	95.83	95.63	95.63	94.25

The results of the small learning capacity teacher model on MNIST dataset are shown in table 5.1, and of the large learning capacity teacher model in table 5.2. Each row of these

Table 5.2: Student accuracy on missing class (MNIST) when distilled with large learning capacity teacher

Missing class	Student accuracy(%) on missing classes at temperature(T)						
	T=3	T=6	T=9	T=12	T=15	T=18	T=20
Digit 0	2.8	0.20	0.61	0.0	0.0	0.0	0.0
Digit 1	0.0	0.26	0.0	0.26	0.08	0.0	0.0
Digit 2	2.5	0.38	0.58	0.48	1.25	0.48	0.19
Digit 3	0.0	0.79	1.38	6.23	2.47	1.38	2.27
Digit 4	0.10	0.10	0.0	0.0	0.0	0.10	0.0
Digit 5	0.67	0.56	1.35	2.80	2.46	3.36	4.26
Digit 6	17.32	2.08	12.83	7.09	18.58	11.37	7.41
Digit 7	0.29	0.29	3.50	0.58	0.77	0.58	1.84
Digit 8	0.0	0.10	0.0	0.0	0.0	0.0	0.0
Digit 9	0.0	0.69	0.10	1.48	1.28	2.08	2.18

tables represents the accuracy on the missing digit, removed class, at different temperatures. For example, row 'digit 0' shows the accuracy of the student model to correctly classify 'digit 0' when it has not seen 'digit 0' during its training. In this way, the student model is forced to learn about missing class solely from the similarity information in soft-labels provided by other examples, like digit 1 - digit 9. The accuracy on missing class achieved by the same student model on distillation from a large capacity teacher is very different. The student model achieves very high accuracy upon distillation by a small learning capacity teacher (table 5.1), but it performs very poorly upon distillation by a large learning capacity teacher model (table 5.2).

It should be noted that the only difference in generating table 5.1 and table 5.2 is the learning capacities of teacher models. We use two different teacher models and keep all other parameters the same in this process. For a student model, the only source of information to learn about the missing class is through rich similarity information present in the soft labels of the teacher model. The drop in accuracy indicates the absence of this crucial information from the soft labels. Consequently, the distillation performance of the RBKD process degrades with it. It can be concluded that a well-trained large learning capacity teacher models fail to retain similarity information between classes and distill very little or no similarity knowledge during the distillation process.

Table 5.3: Student accuracy on missing class (Fashion-MNIST) when distilled with small learning capacity teacher

Missing class	Student accuracy(%) on missing classes at temperature(T)						
	T=3	T=6	T=9	T=12	T=15	T=18	T=20
T-shirt/top	84.70	77.20	62.8	46.70	32.69	17.70	23.70
Trouser	94.90	94.19	92.90	91.39	88.09	88.59	89.99
Pullover	64.89	53.79	44.20	28.99	17.49	10.49	15.00
Dress	84.50	82.09	76.70	69.99	60.50	53.20	46.90
Coat	79.19	71.49	59.79	44.49	28.40	18.99	16.20
Sandal	91.50	92.00	90.49	87.30	84.89	82.09	82.30
Shirt	55.80	42.30	30.30	17.49	6.40	1.70	2.70
Sneaker	91.79	87.69	83.30	69.99	63.89	50.19	57.09
Bag	95.20	95.20	94.19	91.60	88.80	86.79	86.19
Ankle boot	91.29	89.99	87.99	83.39	79.69	75.80	74.00

Table 5.4: Student accuracy on missing class (Fashion-MNIST) when distilled with large learning capacity teacher

Missing class	Student accuracy(%) on missing classes at temperature(T)						
	T=3	T=6	T=9	T=12	T=15	T=18	T=20
T-shirt/top	0.6	6	7.6	7	8.4	8	9
Trouser	0.0	4.2	15.4	19.2	23.2	21.5	16.8
Pullover	0.0	0.0	0.1	0.1	0.1	0.1	0.1
Dress	0.8	0.7	1.8	1.0	1.7	1.8	1.4
Coat	0.3	0.4	0.4	1.4	1.0	2.0	1.6
Sandal	0.2	0.4	0.4	0.4	0.4	0.5	0.5
Shirt	0.2	0.7	1.1	1.7	2.1	1.5	1.9
Sneaker	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Bag	0.0	0.5	0.4	0.4	0.9	0.5	0.4
Ankle boot	1.1	2.6	4.4	3.6	3.7	4.1	3.7

We observe a similar trend on Fashion-MNIST dataset (table 5.3 and 5.4), and on CIFAR-10 dataset (table 5.5 and 5.6). All these results point to the same conclusion that distillation with a large learning capacity teacher model leads to poor performance by the student model on missing class, but under similar circumstances, the same student model performs much better upon distillation by a small learning capacity teacher model. We already established that accuracy on removed class is solely dependent upon the similarity information between classes. Therefore, it is safe to conclude that a well-trained teacher model with large learning capabilities to understand the data distribution fails to provide this crucial similarity information in the RBKD process.

Table 5.5: Student accuracy on missing class (CIFAR-10) when distilled with small learning capacity teacher

Missing class	Student accuracy(%) on missing classes at temperature(T)						
	T=3	T=6	T=9	T=12	T=15	T=18	T=20
Airplane	70.99	66.6	52.1	42.6	39.7	34	31.5
Automobile	80.8	77.5	70.6	63.8	55.8	46.8	45
Bird	48.1	41.8	35	26.9	25.7	16	14
Cat	50.8	42.1	33	22.6	8.7	6.2	4.6
Deer	61.59	53.2	55.7	43.1	26.2	20.6	16.2
Dog	67.4	55.6	29.1	20.2	16	10	10.8
Frog	79.5	76.3	65.9	60.5	47.1	40.6	37.7
Horse	67.69	63.8	63.2	54.2	52.7	45.2	40
Ship	73.79	70.3	67.5	59.7	54	44.4	46.3
Truck	79.4	75.4	73.3	66.8	55.1	51.8	43.3

Table 5.6: Student accuracy on missing class (CIFAR-10) when distilled with large learning capacity teacher

Missing class	Student accuracy(%) on missing classes at temperature(T)						
	T=3	T=6	T=9	T=12	T=15	T=18	T=20
Airplane	1.79	0.89	0.49	0.70	0.80	0.60	0.60
Automobile	2.09	0.70	0.30	0.20	0.30	0.30	0.10
Bird	0.89	0.60	0.30	0.40	0.30	0.20	0.20
Cat	0.80	0.70	0.49	0.30	0.30	0.70	0.10
Deer	0.99	1.49	0.89	0.60	0.80	1.09	0.40
Dog	0.70	0.70	0.40	0.20	0.30	0.10	0.20
Frog	2.30	1.89	1.60	1.99	1.4	2.4	2.09
Horse	0.80	0.20	0.40	0.40	0.20	0.30	0.49
Ship	3.99	2.09	1.49	1.09	1.09	0.60	0.49
Truck	3.09	2.99	1.09	0.80	0.09	0.60	0.80

We observe a similar pattern on three different datasets, but we are still to understand the role played by the presence or absence of similarity information in soft labels. The RBKD process uses only the response from the teacher model, in the form of rich similarity

Table 5.7: Student accuracy on missing class (MNIST, F-MNIST and CIFAR-10) when soft-labels are generated by LS process.

Missing class	Student accuracy(%) on missing classes at different α_{LS}						
	$\alpha_{LS} = 0.2$	$\alpha_{LS} = 0.3$	$\alpha_{LS} = 0.5$	$\alpha_{LS} = 0.6$	$\alpha_{LS} = 0.7$	$\alpha_{LS} = 0.8$	$\alpha_{LS} = 0.9$
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

information in soft-labels, to distill its knowledge to a student model. We argue that the rich similarity information in soft-labels carries information about more than one class and helps the student model to learn about many classes from a single example through *one-example-to-many-classes learning* paradigm discussed earlier. Because of this hidden information in soft labels, the student model can learn about the class which it never sees during training. But in the absence or scarcity of this similarity information, it fails to learn about more than one class from a single example, *one-examples-to-one-class learning*, leading to poor accuracy on the missing classes during testing.

We present our argument through the soft-label hypothesis shown in figure 3.2. It states that the shift from *one-example-to-many-classes learning* to *one-example-to-one-class learning* is due to the absence of similarity information in the soft labels and this change in similarity information is caused by the change in learning capacity of the teacher model. At one end of the distillation process, where the distillation happens through one-example-to-many-classes learning, the distillation process behaves like a RBKD process, but at the other end, the distillation happens through one-example-to-one-class learning behaving like a LS process. We provide further evidence for the soft-label hypothesis by performing the same experiment on the LS training process as shown in table 5.7. The soft-labels are generated using the equation 3.2. The student model fails to correctly classify even a single missing class example for all three datasets and any value of smoothing factor, α_{LS}.

5.2 Change in Entropy with Temperature

In section 4.4.1, we discuss the relation between entropy and the rich similarity information present in the soft labels. We use entropy as the indicator of similarity information. In section 4.5, we discuss the relation between temperature T and entropy. Here, we present the change in entropy with temperature for RBKD process in figure 5.2 and with smoothing factor α_{LS} for LS process in figure 5.1. Observing these plots together [1] provides further evidence in support of the soft-label hypothesis.

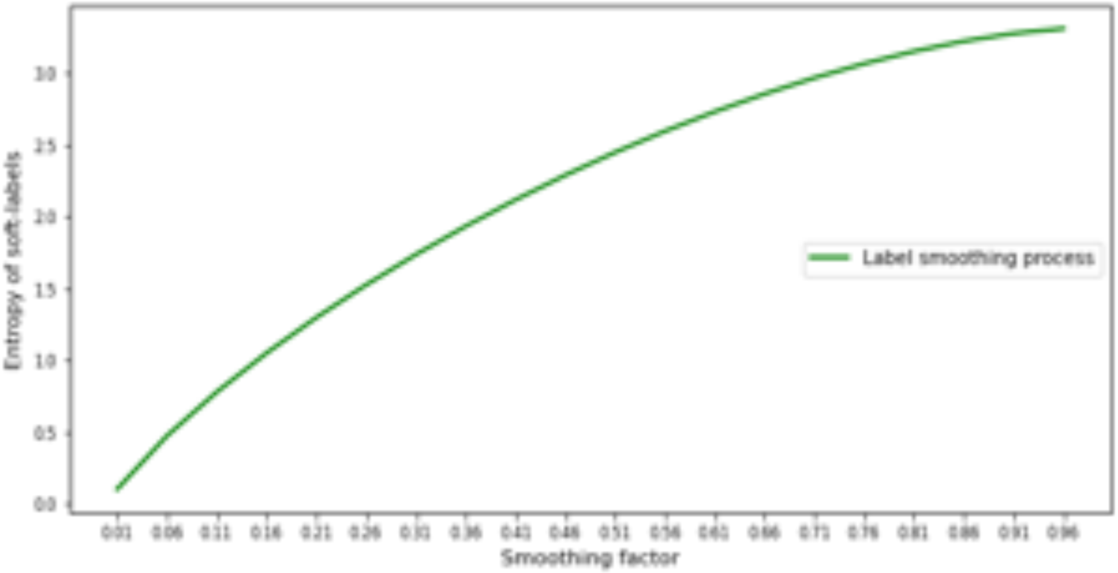


Figure 5.1: Change in entropy with change in smoothing factor (α_{LS}) for Label Smoothing process.

The results of the similarity information presence test for the LS process, shown in table 5.7, show that the student model fails to correctly classify even a single example of the missing class. This indicates that the soft-labels of the LS process does not have any similarity information in its soft-labels. Based on these observations, we argued that the LS process is a special case of the RBKD process, where no similarity information exists in

[1] We do not plot fig. 5.1 and 5.2 on the same x-axis because it is hard to find a one-to-one correspondence between temperature and smoothing factor

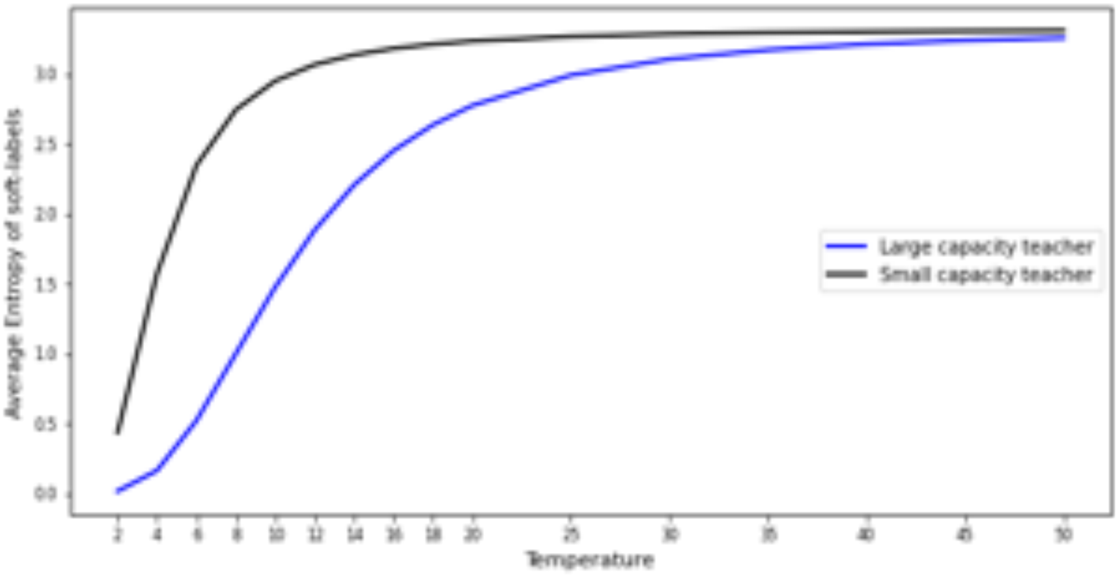


Figure 5.2: Change in entropy with change in temperature for small and large learning capacity teacher models on MNIST dataset.

the soft-labels. The figure 5.1 shows the change in entropy of the soft-labels generated by the LS process with a change in the smoothing factor α_{LS}. The line in this plot shows the entropy line which corresponds to the complete absence of similarity information in soft labels. In the soft-label hypothesis, this corresponds to the rightmost end, in figure 3.2, of the distillation process.

In figure 5.2, we show the change in average entropy of the soft-labels generated by small and the large learning capacity teacher models at different temperatures. There is a significant difference in the average entropy values of soft-labels generated by two different capacity teacher models. At all temperatures, the average entropy line traced by the small learning capacity teacher model remains higher as compared to the line traced by the large learning capacity teacher. The entropy values of the large capacity teacher model are comparable to the entropy of the LS process. A visual comparison of these two plots points to the same conclusion. This is exactly the idea proposed in the soft-label hypothesis. As the soft-labels lose rich similarity information, their entropy decreases and the RBKD process starts to behave like the LS process.

We observe a similar pattern on Fashion-MNIST, shown in figure 5.3, and on CIFAR-

10, shown in figure 5.4, datasets. For both these datasets, their corresponding small capacity teacher model produces soft-labels with higher entropy as compared to the large learning capacity teacher models. All this points to a similar conclusion and adds to the evidence supporting the generalization of *the soft-label hypothesis*.

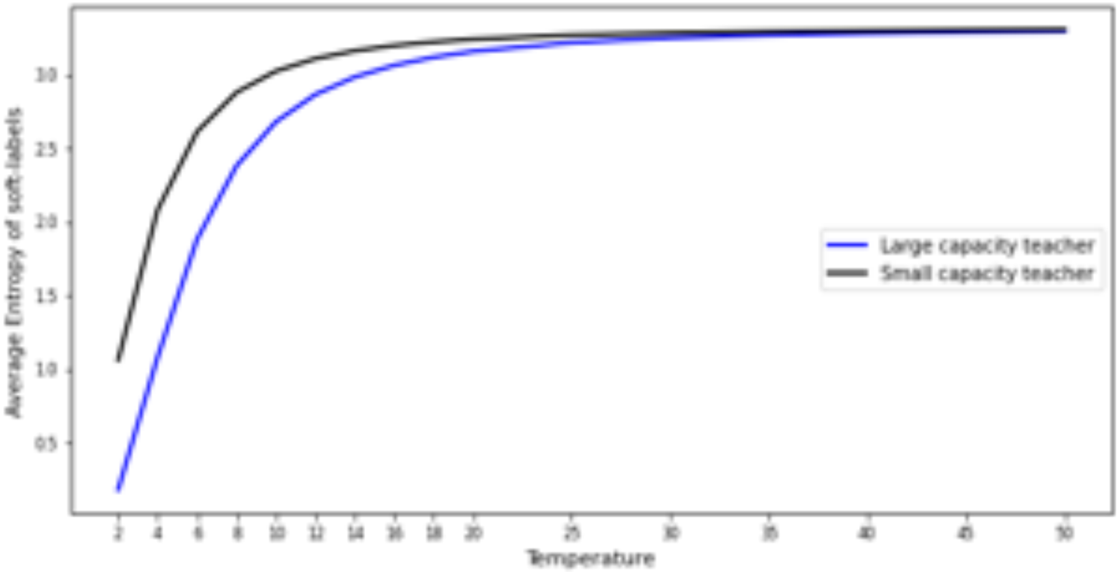


Figure 5.3: Change in entropy with change in temperature for small and large learning capacity teacher models on Fashion MNIST dataset.

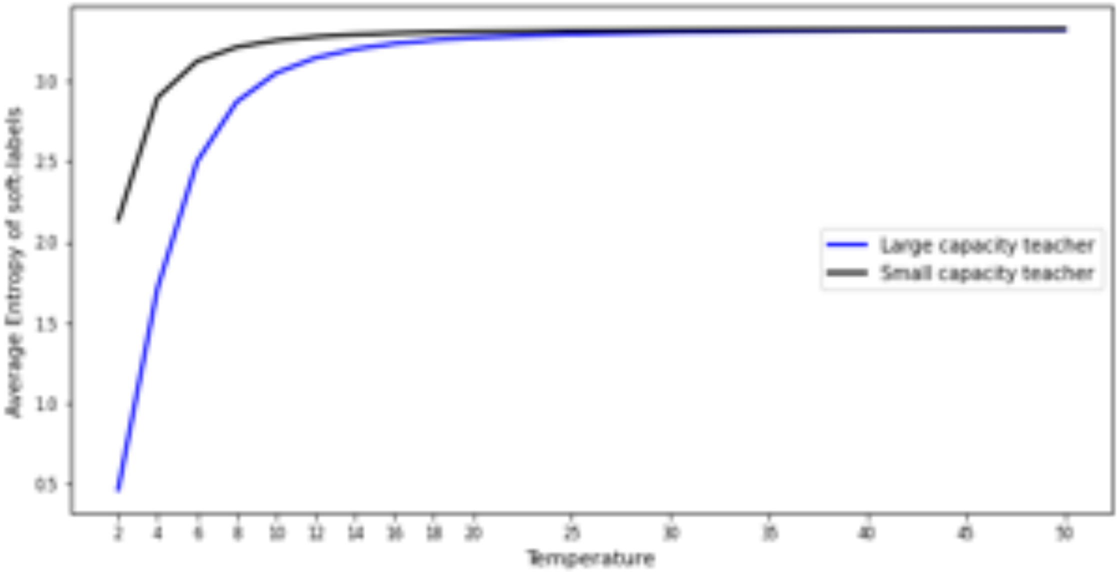


Figure 5.4: Change in entropy with change in temperature for small and large learning capacity teacher models on CIFAR-10 dataset.

There are two important regions in these plots that need a little more description, the low-temperature entropy region, and the high-temperature entropy region. In section 3.4, we talk about confidence labels, similarity labels, and variance in similarity labels. We argue that for effective distillation, there should be a variance in similarity labels along with rich similarity information. This variance should not be as per any pre-defined distribution (see section 4.8 for detail), rather this should be as per the understanding of the data distribution of the teacher model. From figure 5.2, 5.3, and 5.4, one might conclude that the soft-labels generated by a large learning capacity teacher model can be enriched with similarity information by raising the temperature, but our experimental results do not provide any such evidence. As we argued in section 4.5, at high temperature, soft-labels become more like a uniform distribution by assigning similar value, up to many significant decimal places, to each class and losses the rich similarity information as well as the variance in similarity labels from the soft-labels. This is also the reason to restrict temperature value in range $[2, 20]$ as argued in [8]. With soft labels roughly in the same decimal range, the entropy increases but the *variance in similarity labels* are lost.

We have already discussed the importance of variance in similarity information for the RBKD process in section 3.4. We argue that as long as there is enough variance in similarity labels, the distillation process happens. At a small temperature, confidence labels are high and similarity labels have low values, but even in those small values, the small capacity teacher model manages to provide enough variance in similarity labels as per its understanding of the data distribution. This facilitates one-example-to-many-classes learning, even at a small temperature. The accuracy on missing classes can be seen in table 5.1, 5.3 and 5.5 which support this argument.

In this section, we provided additional empirical evidence to support the soft-label hypothesis using entropy as a proxy for similarity information in soft labels. The soft-label hypothesis can also be presented in terms of the entropy of the soft labels. This is shown in figure 5.15. With the help of the entropy version of the soft-label hypothesis, in chapter 6, we show that the decrease in entropy caused by the large learning capacity teacher model can be reversed with some special consideration during its pre-training.

5.3 Entropy based transfer-set selection

We discussed in section 4.6 that one-example-to-many-classes learning should require a lesser number of examples per class as compared to one-examples-to-one-class learning. To test this assertion, we use the entropy of the soft-labels generated by each type of teacher model to select top N examples, where $N \in [5, 2000]$. We compare the number of examples required by both the teacher models to achieve similar performance on distillation. It provides empirical evidence to show that the rich similarity information encourages *one-example-to-many-classes learning* in the RBKD process.

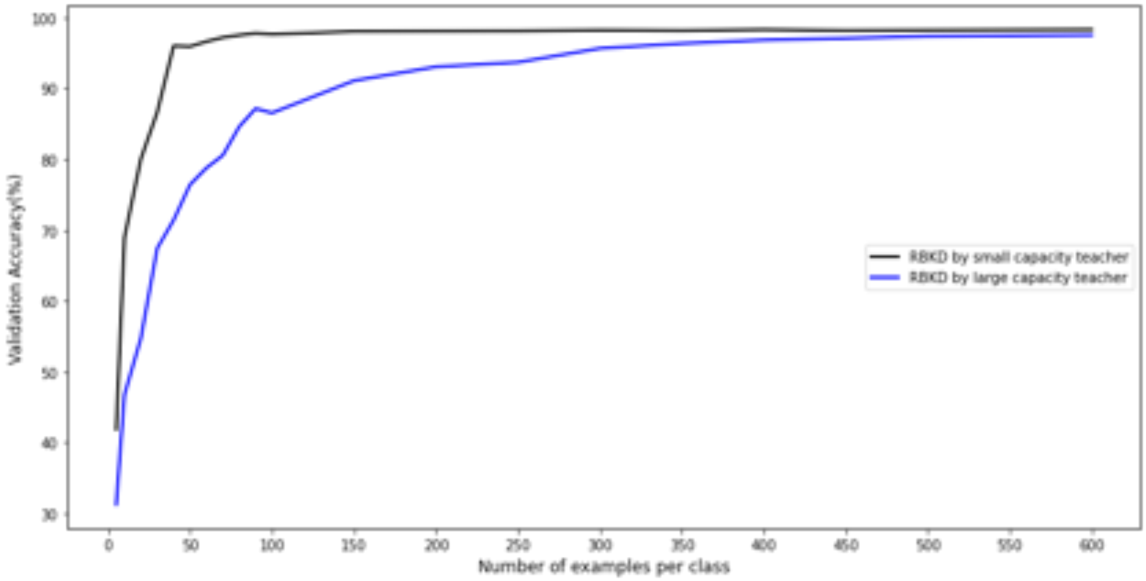


Figure 5.5: The performance of a student model on the test set of MNIST on distillation by small and large learning capacity teacher with the number of examples per class. The smaller the number of examples per class required to achieve good accuracy, the better is the RBKD process. In this plot, the small capacity teacher model always outperforms large capacity teachers. It needs 60 examples per class to achieve more than 95 % validation accuracy while a large capacity model requires around 500 examples per class for that.

The figure 5.5 [2] shows the validation accuracy of a student model with the number of high-entropy examples per class. The two lines represent distillation by a small, black line, and large, blue line, learning capacity teachers. The performance of the student model

[2] For MNIST we plot N up to 600 as it is enough to understand this experiment on this data, for other two datasets, F-MNIST and CIFAR-10, we show N till 2000.

distilled with a small capacity teacher is always better than the performance of the same student model on distillation by a large capacity teacher. The distillation by small capacity teacher takes as small as 60 examples per class to achieve more than 95% validation accuracy, whereas the large capacity teacher can reach a similar performance around 500 examples per class on the MNIST dataset. The fact that the distillation by a small capacity teacher requires a very small number of examples per class suggests that during distillation the student model learns through a one-example-to-many-classes learning paradigm. This is facilitated by the presence of rich similarity information in soft labels. This process enables the student model to quickly learn about the data distribution from a very limited number of examples per class. At the same time distillation through a large teacher does not support one-examples-to-many-classes learning and requires around 500 examples per class to achieve similar performance.

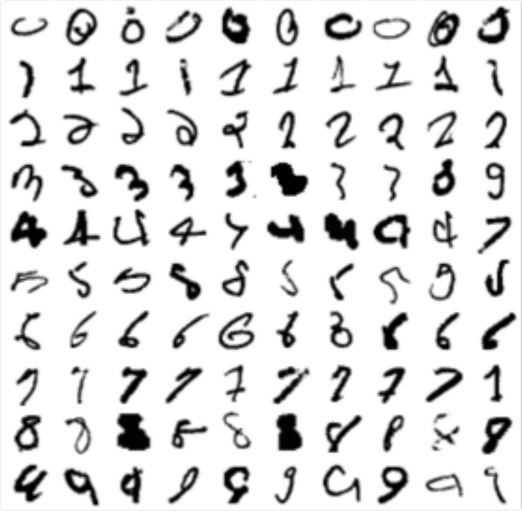


Figure 5.6: MNIST: top 10 high entropy examples from each class selected by small learning capacity teacher at temperature 9.

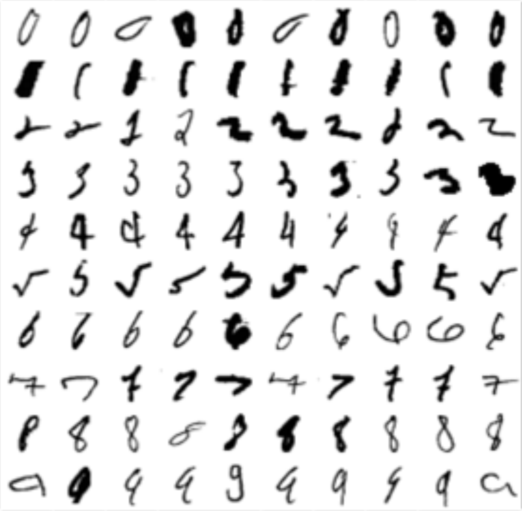


Figure 5.7: MNIST: top 10 high entropy examples from each class selected by small learning capacity teacher at temperature 9.

This phenomenon can be further understood by looking at the high entropy examples selected by each of the teacher models based on their entropy calculation at a given temperature. For all three datasets, MNIST, Fashion MNIST, and CIFAR-10, the temperature T is set at 9. These examples, shown in figure 5.6 and 5.7, provide an insight about the most suitable examples for the RBKD process as seen by respective teacher models. But it is interesting to see that the high entropy examples selected by both these teachers are quite different from one another. The examples selected by large learning capacity models are much more difficult to learn by a machine learning model or in some cases, even by humans. Most of these examples are highly distorted and miss-labeled, like 9 labeled as 3, 1 labeled as 7, 9 labeled as 7. It is also very hard to find the visual similarity among different classes in these examples, like 4 looks nothing like 9 or 3 looks nothing like 8. But the examples selected by small learning capacity teachers are comparatively less difficult to learn by the model. The examples are less distorted and we can also find some visual similarities among different classes. These slight differences in example selection by the two teacher models make a great impact on the RBKD process. The effect on distillation is so large that a well-trained large-capacity teacher model needs around 9 times more examples per class to reach a similar performance.

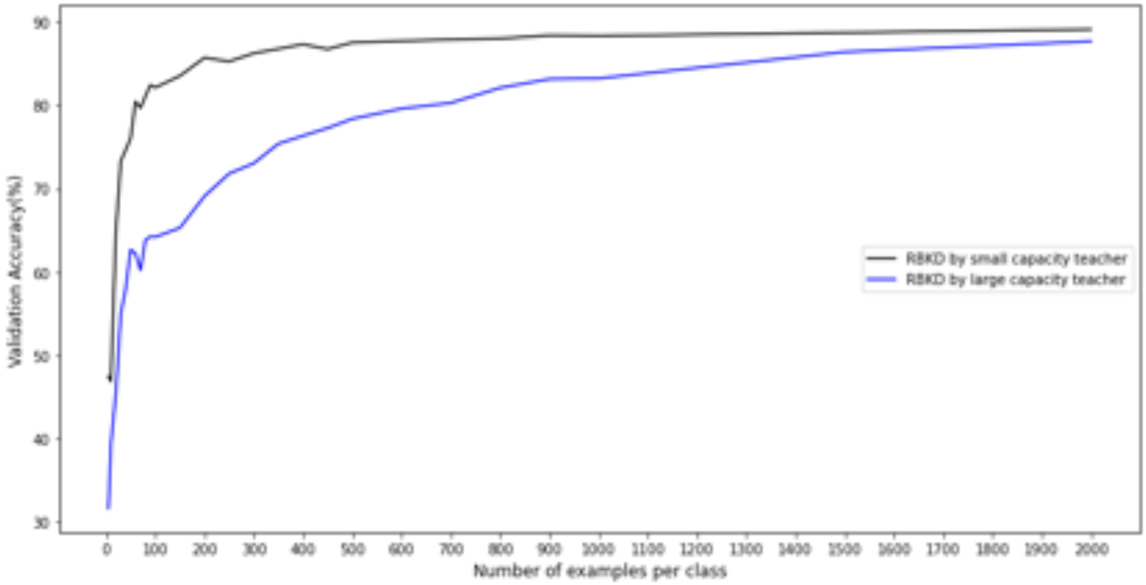


Figure 5.8: The performance of a student model on the test set of F-MNIST on distillation by small and large learning capacity teacher with change in the number of examples per class. For this dataset, the performance gap has further widened as compared to MNIST. The small capacity teacher model always outperforms large capacity teachers.

The MNIST dataset is considered easier than the Fashion-MNIST and CIFAR-10 datasets. Because of this reason, a large learning capacity teacher model requires only 9 times more examples per class. But on Fashion-MNIST, as shown in figure 5.8, the gap in performance of student model on distillation further widens. The large capacity teacher mode takes around 1900 examples per class to reach the validation accuracy achieved by the student model on distillation with the small capacity teacher model at 100 examples per class. With more difficult data distribution the gap further widens. Looking at the high entropy examples selected by these teacher models, shown in figure 5.9 and 5.10, reveals a similar pattern as observed on MNIST.

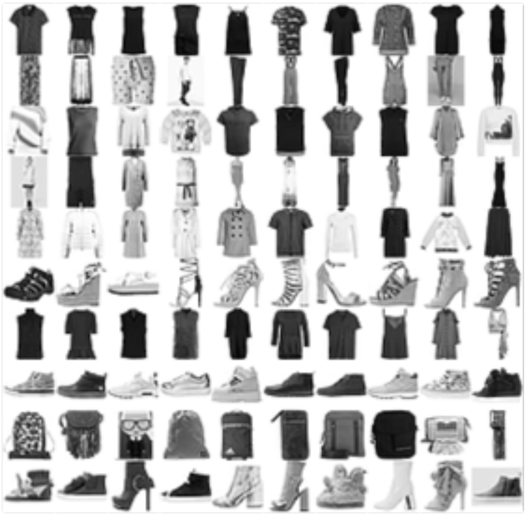


Figure 5.9: F-MNIST: top 10 high entropy examples from each class selected by small learning capacity teacher at temperature 9.

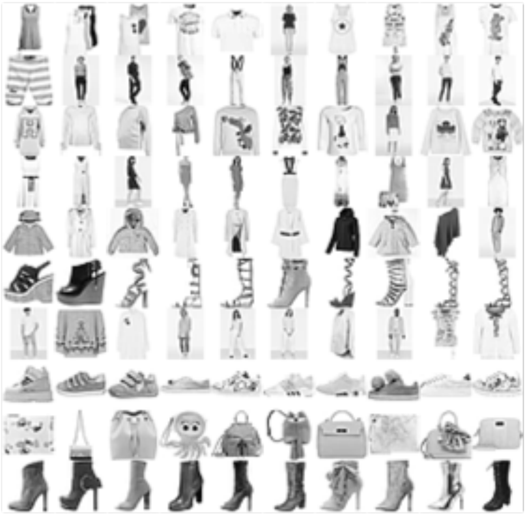


Figure 5.10: F-MNIST: top 10 high entropy examples from each class selected by small learning capacity teacher at temperature 9.

A row-wise comparison between the above two figures shows the stark difference between the selected examples by the teacher models. Most of the examples chosen within a class by a large learning capacity teacher model look similar and do not have distinguishing properties. For example, in row 1, 2, 7 and 9 most of the examples are black objects which looks kind of similar to each other but it is hard to find any visual similarity with examples from other classes. The examples selected by the small learning capacity teacher model look more diverse and have different visual properties. It selects a variety of examples that can represent properties of more than one class, like row 2 of figure 5.10 it selects examples that seem like a crossover between two classes. This variety in examples can represent the properties of more than one class to facilitate one-example-to-many-classes learning.

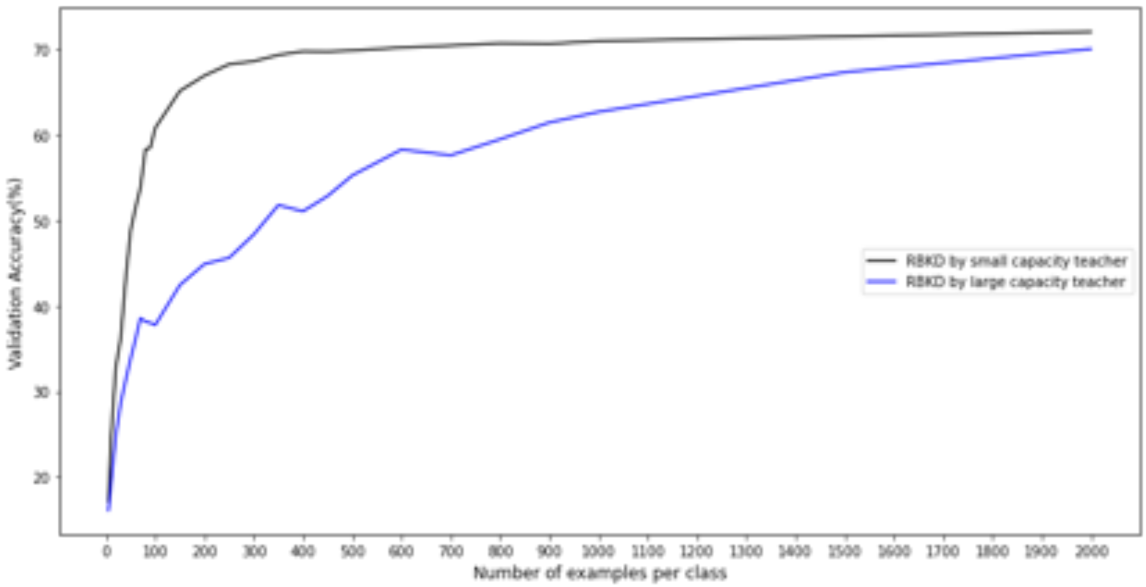


Figure 5.11: The performance of a student model on test set of CIFAR-10 on distillation by small and large learning capacity teacher with change in number of examples per class. The gap in performance has further widened. In this case as well, the small capacity teacher model always outperforms large capacity teacher.

We observe a similar pattern on the CIFAR-10 dataset but with more gaps in performance for this experiment. As shown in figure 5.11, the gap in performance of the student on distillation has further widened for teacher models with small and large learning capacity. A closer look at the high entropy examples selected by these two models, 5.12 and 5.13, can further explain the performance gap. For example, row 3, bird class, of both the figures are very different. It is filled with images of the head of the birds when selected using a large capacity teacher model, but it has complete images of birds when selected using a small capacity teacher. These differences at the smallest level of example selection make a big difference in the RBKD process.

With the significant difference in the required number of examples per class by the large and small capacity teacher models, it can be concluded that soft labels from small capacity teacher models provide rich similarity information that facilitates one-example-to-many-classes learning. This further supports the soft-label hypothesis presented in section 3.5.

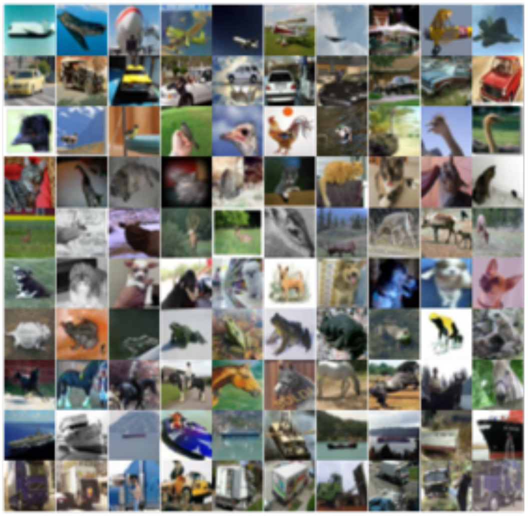


Figure 5.12: CIFAR-10: top 10 high entropy examples from each class selected by small learning capacity teacher at temperature 9.

Figure 5.13: CIFAR-10: top 10 high entropy examples from each class selected by small learning capacity teacher at temperature 9.

5.4 Penultimate layer Representation

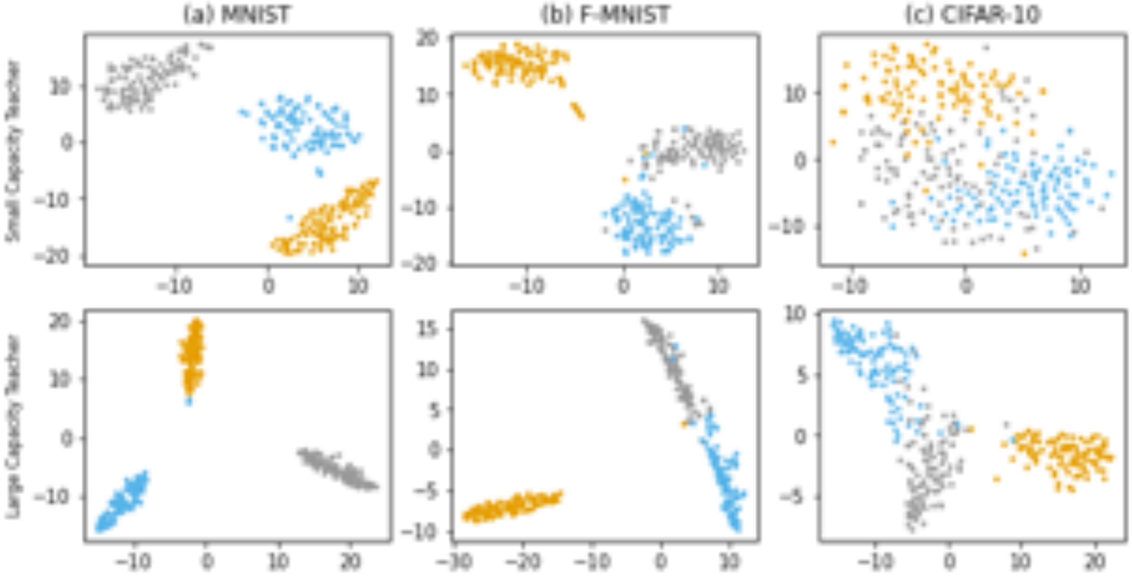


Figure 5.14: The linear projection of activation of penultimate layer for small and large learning capacity teacher models. The plot is constructed for 3 randomly chosen classes having 100 examples each. The first row shows the representation of the small learning capacity teacher model and the second row shows the representation of the large learning capacity teacher model. Column (a), (b) and (c) shows representations for MNIST, Fashion-MNIST, and CIFAR-10, respectively.

This method of visualizing penultimate layer representations of a model is used to show that a tighter cluster formation leads to loss of similarity information between different classes [16]. We use this analogy to support the soft-label hypothesis. In figure 5.14, we show clusters of penultimate layer activations of image classifier teacher networks trained on MNIST, Fashion-MNIST, and CIFAR-10 datasets. The first and second rows of the figure show representation of small and large learning capacity teacher model, respectively. The columns represent different datasets. The plot is constructed for three randomly selected classes, each cluster corresponds to a class, having 100 examples each. In the first row plots, we observe that the representations are spread into defined but broad clusters. But in the second-row plots, the clusters are much tighter. The broad clusters indicate that the presence of similarity information between classes that transcends to soft-labels generated by respective teachers, whereas the tighter cluster indicates the absence of similarity information between classes.

This behavior is shown by penultimate layer representation when a model is trained with LS process [16]. We have already shown that the soft-labels of the LS process do not carry similarity information (see table 5.7). In the distillation process, this behavior is shown by a well-trained large learning capacity teacher model. This supports our hypothesis that as the teacher model becomes more powerful, it loses similarity information in its response. This aligns with the soft-label hypothesis which states that as the learning capacity of the teacher model increases, the RBKD process moves towards LS process. We observe this behavior in all three datasets, which shows the generalization capability of the soft-label hypothesis across different datasets.

5.5 Conclusion

The degradation in the performance of the student network has been largely attributed to the gap in the learning capacity of a teacher and a student model. But our experiments show that similarity information in soft labels plays a very important role as well. We test the presence of similarity information by testing the student model on missing classes, classes removed from the transfer set. It shows a stark difference in accuracy on distillation from small and large capacity teachers. We also show that the similarity information can be tracked with the entropy of soft-labels and show that soft-labels from a large capacity teacher model have much less entropy as compared to soft-labels from a small capacity teacher at the same temperature. It clearly explains the role of similarity information in distillation. To show the direct relation between the one-example-to-many-classes learning and the RBKD process, we show that distillation happens faster and with fewer examples per class in presence of rich similarity information. To concretize this idea, we show the penultimate layer activation projection for different teacher models on different datasets. All of them show a similar pattern. The tighter clusters formed by penultimate layers of a large capacity teacher model indicate a lack of similarity information in its response leading to degraded performance on distillation.

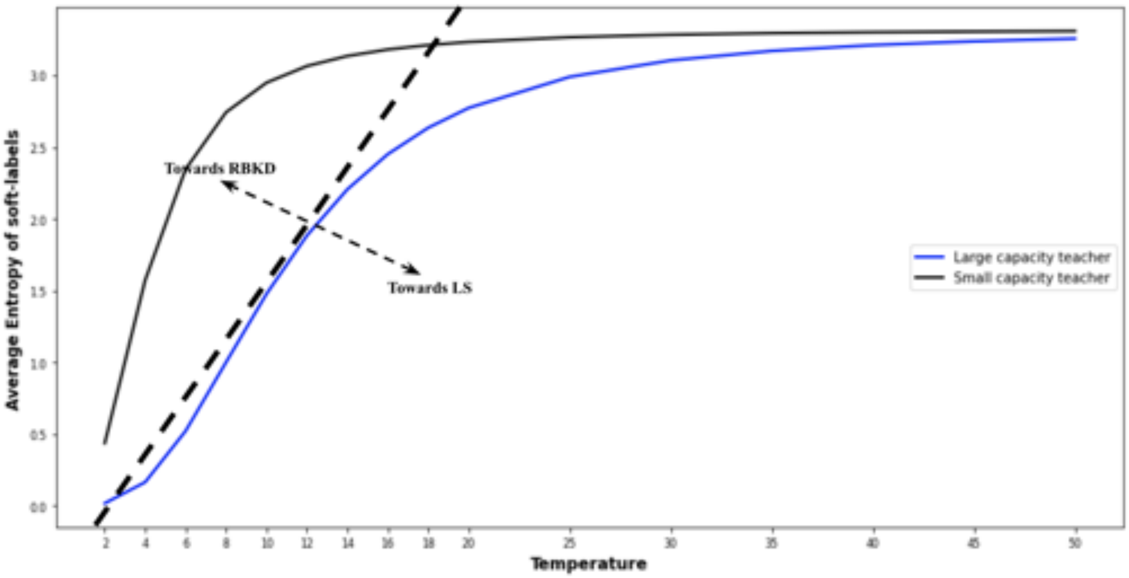


Figure 5.15: The soft-label hypothesis (figure 3.2) can also be presented in terms of entropy. The figure shows the average entropy of soft-labels generated by small (black line) and large (blue line) capacity teachers. The dashed lines roughly separate the LS and RBKD nature of the distillation process. If the line moves away from the dashed line downwards, the distillation process will get close to LS process and if the entropy line moves away from dashed line upwards, the distillation process will get close to RBKD process.

To better explain these findings, we propose the soft-label hypothesis. It says that for a fixed learning capacity of the student model, as the learning capacity of the teacher model has increased the process of response-based knowledge distillation (RBKD) shifts away from a similarity-information rich soft-labels based distillation process towards a no-similarity containing a less confident form of the one-hot label based label smoothing process. All experimental results support the soft-label hypothesis. With the intricate relation between the similarity information and entropy, the soft-label hypothesis can also be presented in terms of entropy of the soft labels. It is shown in figure 5.15. The figure shows that based on the entropy of soft labels, we can differentiate between the different behaviors of the distillation process. With the decrease in soft-label entropy, caused by the increase in teacher learning capacity, the distillation process starts to behave more like a label smoothing process. But at high entropy of the soft-labels, it behaves like a RBKD process. We can draw an imaginary line, the dashed line in figure 5.15, to roughly define the boundary between these two behaviors of the distillation process. Both the forms of the soft-label hypothesis, the general form (figure 3.2) or the entropy form (figure 5.15), present the same idea.

We see that the large capacity teacher models do poor distillation, but for large and complicated datasets, we have to train DL models with billions of parameters. Does this mean that the knowledge from a well-trained large teacher model can not be distilled? We discuss this problem in the next chapter and propose a few special considerations for the pre-training step of the teacher models for better knowledge distillation.

CHAPTER 6

RETAINING THE SIMILARITY INFORMATION IN A MODEL

6.1 Overview

The soft-label hypothesis highlights the important role played by the similarity information in the distillation process. It argues that a well-trained large capacity teacher model loses the critical similarity information in its response, leading to poor performance on distillation. Does this mean that the knowledge of a well-trained large learning capacity teacher model can not be distilled to a more efficient model? In this chapter, we discuss this problem and propose a few special considerations for pre-training the large cumbersome teacher model to retain the critical similarity information in its response. First, we discuss the theory of retaining the similarity information in any teacher model in section 6.2, then, we discuss the experimental setup and show the results in section 6.3. In section 6.4, we discuss the entropy form of the soft-label hypothesis in more detail and present our conclusion.

6.2 Theory of retaining similarity information

The soft-label hypothesis argues that the teacher model starts to lose similarity information in its response with an increase in its learning capacity for given data distribution. The absence or loss of similarity information directly affects the distillation process by pushing the *one-example-to-many-classes* learning process to *one-example-to-one-class learning* process. This loss of similarity information in response can be reverted by taking a few special considerations during the pre-training of the teacher model. We argue that for a given student model, any teacher model can be trained to retain the rich similarity information in its response by finding the right balance between two important factors, batch size of teacher pre-training, number of epochs of teacher pre-training, and the gap in learning capacity between teacher and the student. Figure 6.1 represents this idea through the Venn diagram.

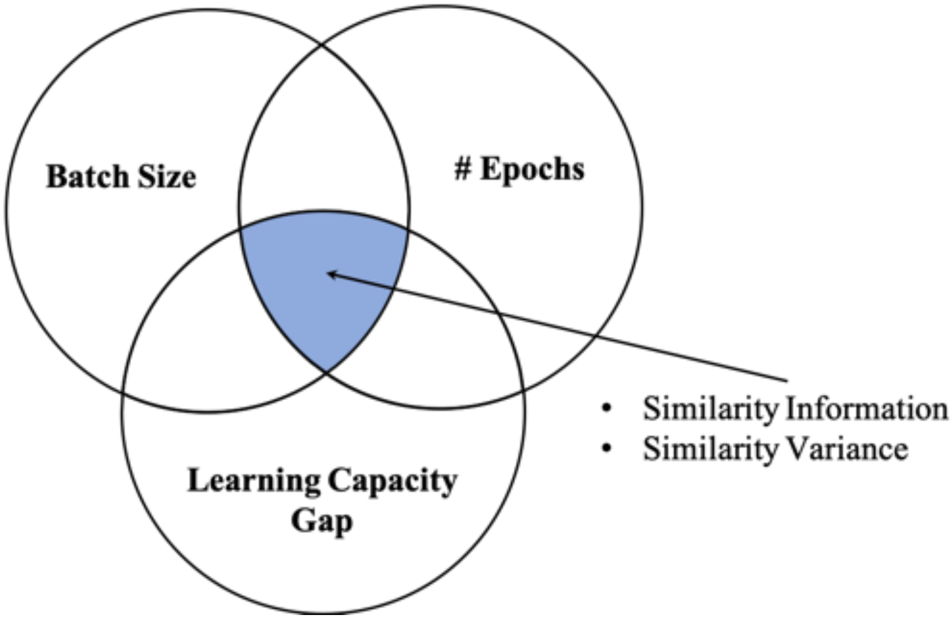


Figure 6.1: The lost similarity information in soft-labels can be regained by finding the right balance in three factors, batch size of teacher pre-training, number of epochs of teacher pre-training, and learning capacity gap between teacher and student. The blue-filled region symbolizes the area of balance that generates soft labels with similarity information and variance in similarity information. The Venn diagram is symbolic.

To develop a theoretical understanding of this problem, it is important to understand the classification problem setting with DL framework and the importance of batch size and number of training epochs. For multi-class classification problems in computer vision, the DL models are trained with a cross-entropy loss function. Training with cross-entropy loss function promotes accuracy with high confidence. A highly confident model tends to learn the most precise discriminative properties between classes and ignores the similarities between those classes. While it is an ideal situation for the classification problem, it is not the ideal situation for distillation as it thrives on the similarity information between classes. The ideal situation for distillation is to keep the teacher model moderately confused among different classes so that it retains some of the similarity information between classes in probabilistic form. But the model should still be able to achieve high accuracy during training so that it provides correct and relevant information in its soft labels. The high accuracy of the moderately confused model is facilitated by the default nature of softmax function, equation 2.1 with $T = 1$, which assign classes based on the highest probability of the output layer. Therefore, the basic idea behind retaining the similarity information in response to a teacher model is to not let the model perform classification tasks with very high confidence. A moderately confused model with high accuracy is the most suitable model for knowledge distillation.

Every DL model has a learning capacity. Generally, this learning capacity is a function of the number of trainable parameters and the number of hidden layers in the model. The model starts with a random value of parameters and learns through a back-propagation algorithm [17]. Typically, this learning happens in batches of examples. If a model is capable of processing all information from each batch to the best of its understanding, then it can quickly become very confident on the classification task, but if the batch size is larger than its processing capability, then it is not able to process all the information from each of the examples in the batch and does not learn quickly to be very confident about the classification task. But the same model, even with a large batch size can become very confident on the classification task, if it is trained for a long time. By controlling these two factors, batch size and number of epochs, for a given student and teacher model pair, the teacher model can retain similarity information between different classes.

For better knowledge distillation, it very important to find the right balance among these three factors, batch size, number of epochs, and the gap in learning capacity of a teacher and a student model. For a given teacher and a student model, the gap in learning capacity is fixed. The other two parameters, batch size and the number of epochs, can be adjusted to retain maximum similarity information in the response of the teacher model. It is important to note that if the gap in learning capacity between the teacher and the student model changes, these parameters will also change. First, we focus on the batch size. The batch size of the pre-training step should increase with the increase in the learning capacity of a teacher model. It should be large enough to overwhelm the teacher. If a teacher processes a batch size larger than its learning capacity, then it is not able to learn the fine-grained discriminative properties between different classes and is less confident in the classification task. The large batch size forces the teacher model to learn only the high-level distinction between the classes. In other words, it keeps the teacher model mildly confused between different classes and still achieves a good accuracy. This is the best-suited condition for the distillation process as the mild confusion is translated into similarity information in the soft labels. Once this condition is achieved, this batch size becomes the base batch size that can be fine-tuned in tandem with the number of epochs of the pre-training step.

As discussed earlier, even after finding a suitable batch size for the given learning capacity gap between teacher and student model, the similarity information in soft-labels can still be lost by training the model for a large number of epochs. Any DL model with enough learning capacity becomes a bit more confident incorrectly classifying each example at the end of every epoch. With enough learning capacity, it can slowly learn the fine-grained details from the data with more training. This is against the suitable requirement for knowledge distillation. To arrest losing the similarity information due to more training, the number of epochs of pre-training should not be kept higher than the optimal value. This optimal value of epochs varies for a different combination of batch size and learning capacity gap. The entropy of the soft labels at a given temperature can be used as a measure to fine-tune these two factors. We argue that all these three factors should be perfectly balanced for the best distillation performance. We present much empirical evidence to support our argument in this chapter.

6.3 The experimental setup and results

6.3.1 Experimental setup

We perform the same experiments as described earlier in chapter 4. We show the change in entropy with temperature as described in section 4.5, entropy-based transfer set selection as described in section 4.6 and penultimate layer representation of teacher models as described in section 4.7. Apart from this we also use a single-layer neural network as a student model to show the improvement in the performance of this model with change in similarity information in soft labels.

The single-layer acts as a classifier layer or output layer and has dense units equal to the number of classes in the dataset. Since there can not be any student model with a lower capacity than this, we call it a **baseline student model**. Choosing a baseline student model serves a dual purpose. First, for any teacher model, it maximizes the gap in learning capacities of a teacher and a student model. This helps in analyzing the behavior of distillation at the far end of the learning capacity gap. Second, since the RBKD process relies only on the response of the output layer, all the effects associated with the RBKD process are directly attached with the soft labels of the teacher. By reducing the student network to a single layer model, we eliminate any other factor capable of influencing the performance of the student model on distillation. The performance of the student model is affected only by the quality of the soft labels. This setup provides us a better way to critically analyze the role played by similarity information in the RBKD process.

6.3.2 Results

We use entropy as an indicator of similarity information in soft labels. All soft-labels are generated at temperature $T = 9$. We pre-train the large capacity teacher model with different batch sizes and for a different number of epochs to track the changes in the entropy of the soft labels. We present these results in figure 6.2 and 6.3.

In the last section, we explained the effects of increasing the batch size on the entropy of the soft labels during the pre-training step of a teacher model. Figure 6.2 shows the change in entropy of the soft-labels with change in batch size for the large learning capacity teacher model. We vary the batch size in the range $[16, 8192]$ and generate the soft-labels using equation 2.1. The entropy of soft-labels increases with an increase in the batch size of teacher pre-training. We also argue that even with large batch size, training for more

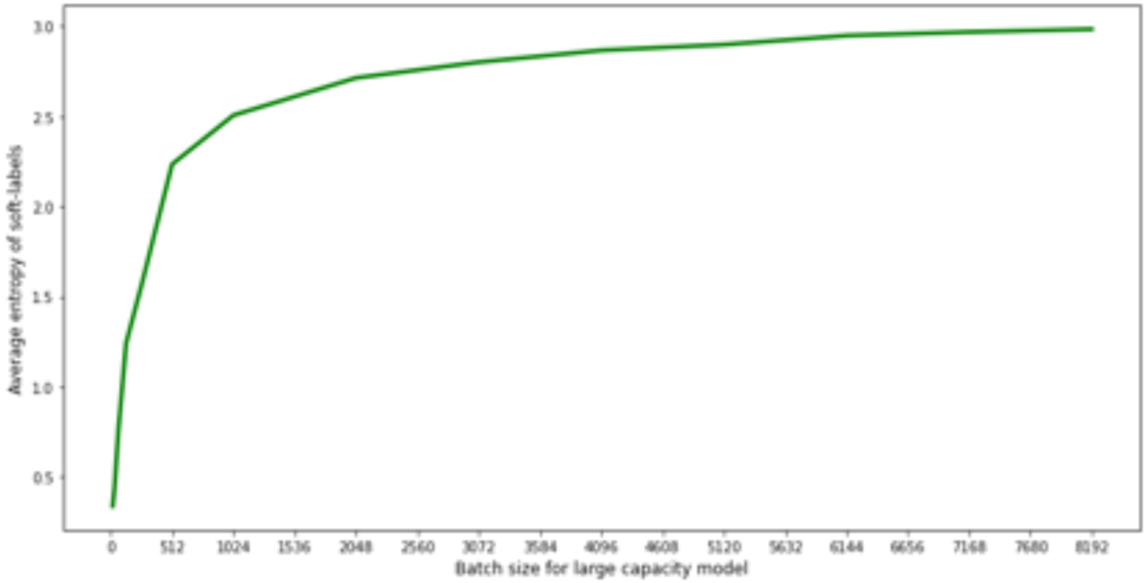


Figure 6.2: The plot shows the change in average entropy of the soft-labels as the large learning capacity teacher model is trained with different batch size for 50 epochs on MNIST. We vary the batch size from 16 to 8192. The entropy of the soft-labels increases with the increase in batch size during pre-training.

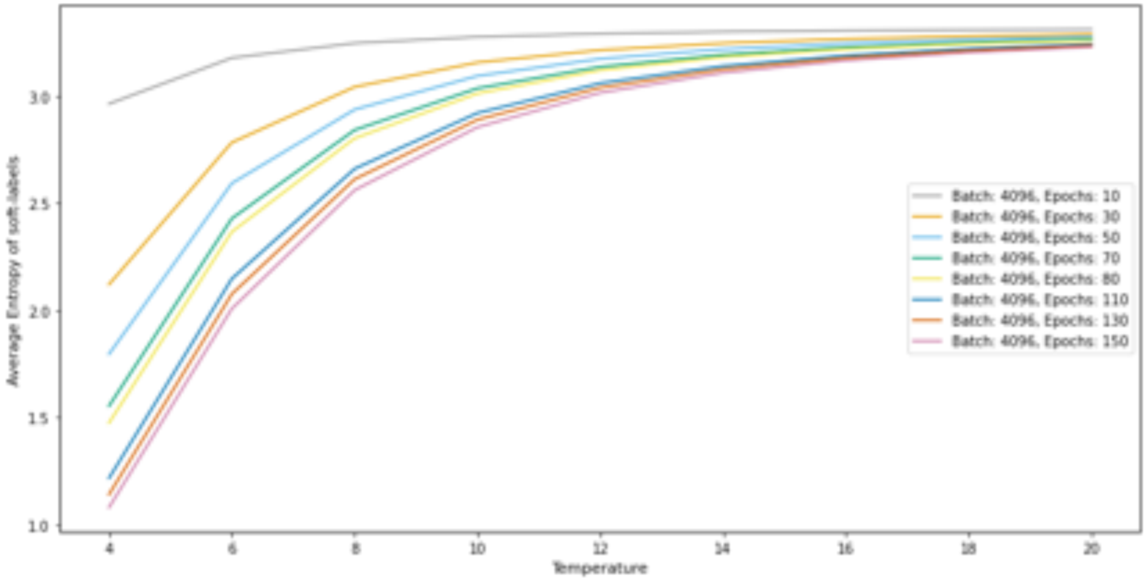


Figure 6.3: The plot shows change in entropy of soft-labels for large learning capacity teacher model trained with batch size of 4096 for various epochs on MNIST. The entropy of the soft-labels decreases as the model is trained for more epochs.

numbers of epochs results in a decrease in entropy of soft labels. Figure 6.3 shows a consistent decrease in entropy of soft-labels of large capacity teacher model as it is trained for more and more epochs. Both these results indicate that it is very important to find a balance between the batch size and number of epochs during pre-training of teacher model for better RBKD process.

In section 5.2, we discussed the change in entropy of soft-labels for small and large learning capacity teacher models on the MNIST dataset (figure 5.2). In figure 6.4, we show the improvement in entropy by increasing the batch size during pre-training of the teacher model. The bold black and blue lines correspond to the lines in figure 5.2. The large-capacity model pre-trained with small batch size, say 16, traces entropy line very close to LS process as shown in 5.1. It indicates a behavior closer to the LS process as argued in the soft-label hypothesis. But when the same teacher is trained with large batch size, say 8192, its behavior comes close to the small capacity teacher model, indicating the presence of similarity information in soft labels.

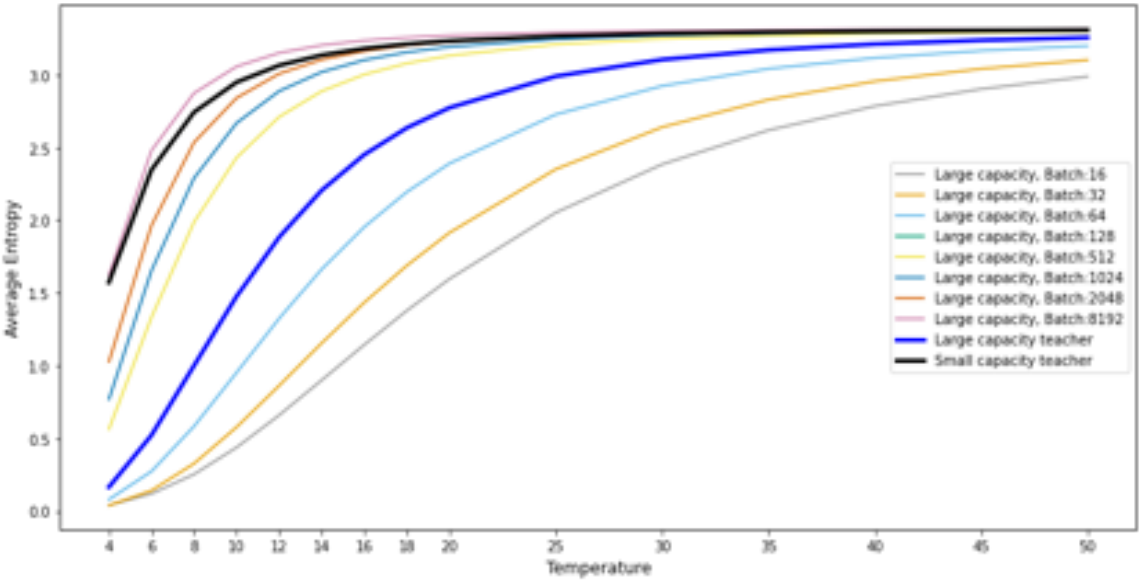


Figure 6.4: The plot shows the average entropy of soft-labels of large capacity teacher model when it is trained with different batch sizes for 50 epochs. The bold black line shows the average entropy of the small capacity teacher model. With a large batch size, the entropy of the large-capacity model can be increase to match the entropy level of the small capacity teacher.

It is important to remember that using this method, we can only gain the performance degraded due to loss of similarity information. The degradation in performance caused by the gap in learning capacity does not improve. To prove this point, we evaluate the performance of the RBKD process on the baseline student model. With a baseline student model, we maximize the gap in learning capacity between teacher and student. It also removes any other factor that could affect the performance of the student model other than the soft labels. Figure 6.5 shows the outcome of this experiment.

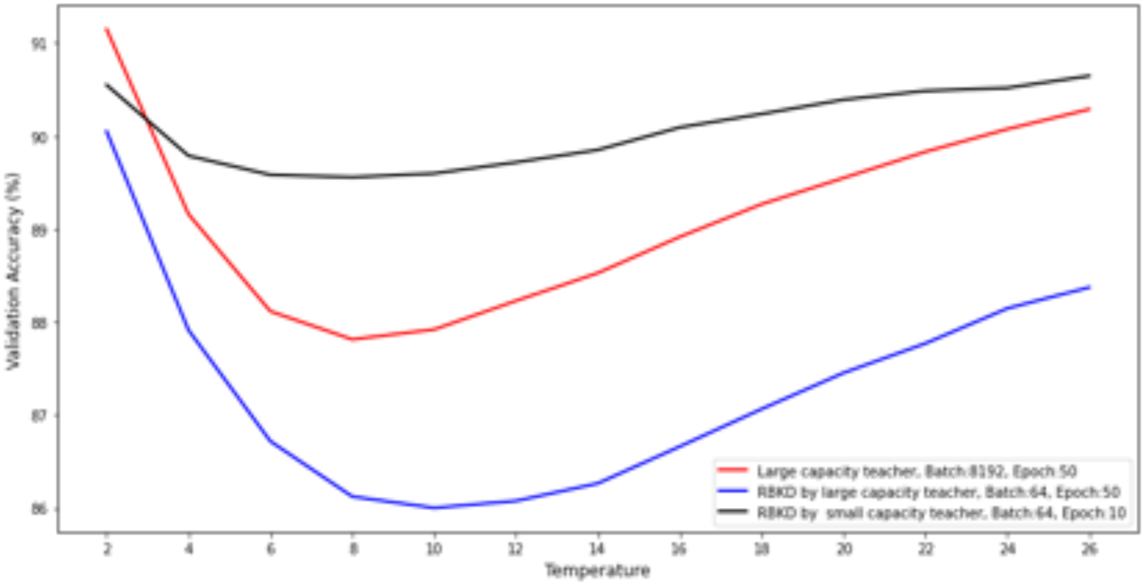


Figure 6.5: The plot shows improvement in the performance of a fixed student model for different capacity teachers, pre-trained with different batch sizes for different epochs on MNIST. A less capacity teacher trained with a batch size of 64 for 10 epochs (blue curve) can do much better distillation as compared to a large capacity teacher trained with 64 batch size for 50 epochs (black curve) at any temperature. It also shows that the same large capacity teacher can perform significantly well when it is forced to retain the rich similarity information (red curve). The degradation in performance attributed to the gap in the capacity of learning can be significantly reduced at all temperatures.

The three lines in figure 6.5 correspond to three different situations of the RBKD process. The black line shows the performance on baseline student model on the validation set of MNIST with small learning capacity teacher model trained with a batch size of 64 for 10 epochs, the blue line shows the performance of the same baseline student model with large learning capacity teacher model trained with a batch size of 64 for 50 epochs, and the red line shows the performance of the baseline student model with large learning capacity teacher model trained with a batch size of 8192 for 50 epochs. We see that the performance of the baseline model improves from the blue line to redline on increasing the batch size, but still, it is not quite close to the performance by the small capacity teacher model, black line. This improvement, from the blue line to the red line, is caused by the presence of similarity information in soft-labels of the large capacity teacher model. But it could not get close enough to a distillation performance with a small capacity teacher model due to the gap in learning capacity of the student and the teacher model. This supports our argument that with these special considerations during pre-training of the teacher model, we can regain the performance of the student model lost due to the absence of similarity information. The degradation of performance caused by the gap in learning capacity of teacher and student is not addressed with this method.

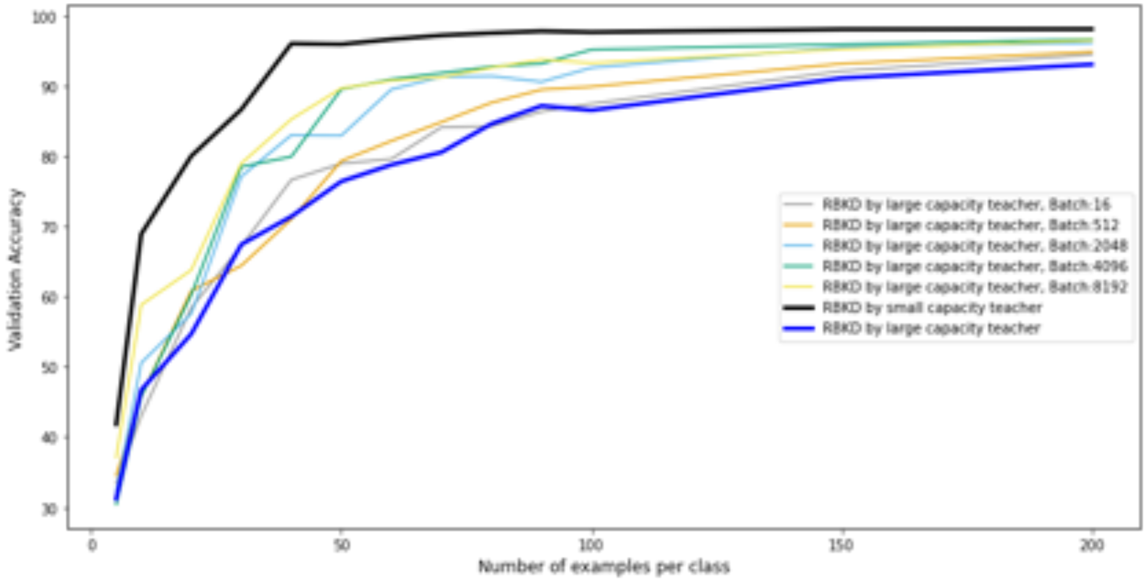


Figure 6.6: The figure shows the improvement in distillation performance of large learning capacity teacher model when trained with larger batch size. The number of examples per class reduces between 150-200 to achieve the same performance as the small capacity learning teacher model.

We argue that the presence of similarity information promotes one-example-to-many-classes learning. The gain in similarity information in soft-labels of a large capacity teacher should reflect in the number of examples it requires for distillation. The figure 6.6 shows this result on MNIST data. The bold black and blue lines correspond to small and large capacity teacher model shown in figure 5.5 for the same experiment. RBKD by a large capacity teacher requires around 500 examples per class to achieve similar performance achieved by a small capacity teacher model with 60 examples per class on distillation. After training with the special considerations, the same large capacity teacher requires close to 100 examples per class to reach similar performance. This shows that the special considerations during pre-training of teachers improve the presence of similarity information in soft-labels resulting in an overall improvement in the performance of the student model on distillation.

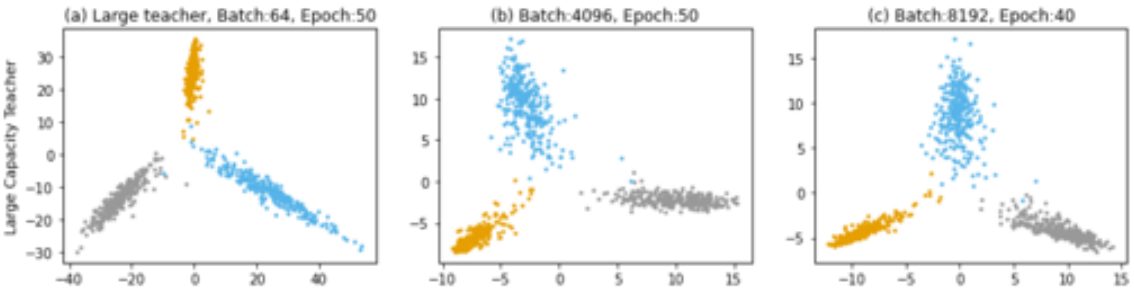


Figure 6.7: The figure shows the improvement in penultimate layer representation on training with the proposed special considerations. Column (b) and (c) shows a significant spread in the representation indicating re-gain of lost similarity information.

We already discussed in section 5.4 the difference in penultimate layer representations of small and large capacity teacher models. In figure 6.7, we show the changes in the penultimate layer representation of the large capacity teacher model when trained with the special considerations. Column (a) of the figure shows the old representation of large capacity teacher model trained with batch size 64 for 50 epochs (figure 5.4), the other two columns, (b) and (c), show the representations of same teacher model trained with batch size 4096 for 50 epochs and batch size 8192 for 40 epochs, respectively. The clusters in column (a) have spread out for large batch size, indicating the gain in similarity information. This result adds further evidence to our argument presented in figure 6.1.

6.4 Conclusion

We argue that finding the right balance between the batch size and the number of pretraining epochs of a teacher model for a given student model can improve the quality of soft-labels better suited for the RBKD process. Using entropy as an indicator of similarity information, we show how a large capacity teacher model can be pre-trained for better

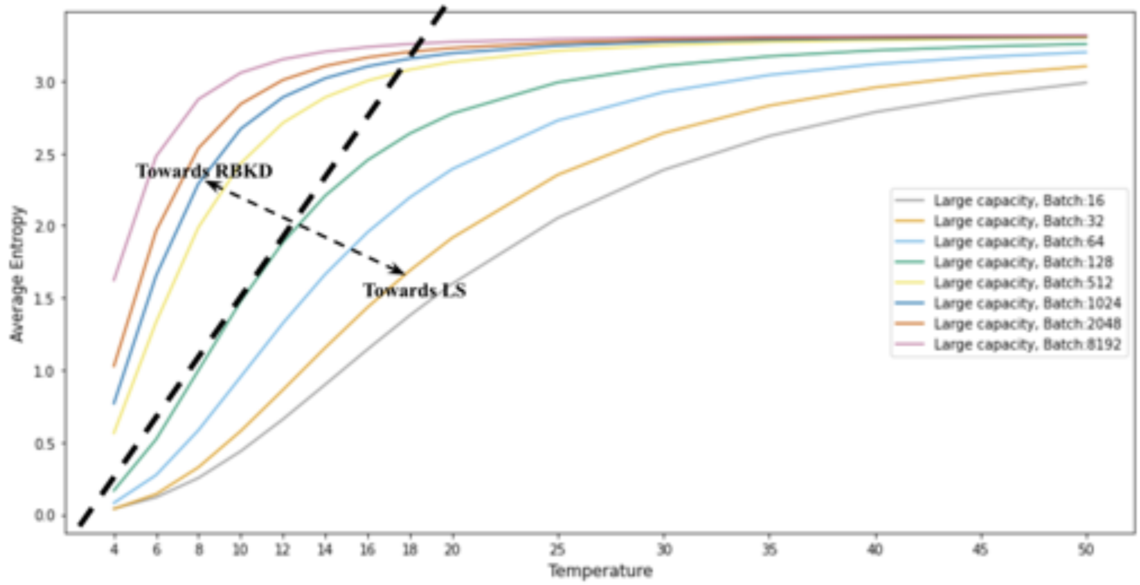


Figure 6.8: The figure shows the entropy version of the soft-label hypothesis. It shows that the same large capacity teacher model can perform different types of distillation based on its pre-training parameters.

response-based distillation. We also argue that these considerations can only help in regaining the performance degradation caused by the loss of similarity information in soft labels. The degradation in performance of the student model caused by the gap in learning capacity is not affected by these considerations. With these results, we present a more precise version of the soft-label hypothesis in terms of entropy as shown in figure 6.8. This is an elaborated version of figure 5.15 presented in chapter 5. The figure shows that the same large capacity teacher model can perform different types of distillation based on its pre-training parameters. A large capacity teacher pre-trained with large batch size can behave like a RBKD process due to the presence of critical similarity information in soft-labels. But when the same teacher model is trained with a smaller batch size, it loses the critical similarity information in soft-labels and behaves like LS process.

CHAPTER 7
CONCLUSIONS

7.1 Summary

We present this work in two parts. In the first part of the work, we argue that the degradation in performance of a student model on distillation is caused by two factors: the absence of similarity information in soft-labels and the gap in learning capacity of teacher and student model. we explain this phenomenon through the soft-label hypothesis. It describes the nature of the distillation process based on the quality of soft-labels generated by the teacher model. We show how the quality of soft labels is dependent upon the learning capacity of the teacher for given data distribution. The hypothesis also explains the different learning paradigms, one-example-to-one-class learning, and one-example-to-many-classes learning, that are facilitated by the presence and absence of similarity information in soft labels. It also talks about the shift from RBKD process to LS process caused by the loss of similarity information in soft-labels. We showed that this hypothesis generalizes across different datasets. The soft-label hypothesis also explains the underlying reason for the improvement in using a teacher assistant model for the distillation process proposed by Mirzadeh et al. [15, 5].

In the second part of this work, we discuss some special considerations for retaining the similarity information in soft-labels of a large capacity teacher model. We show improvement in the quality of soft labels by finding a suitable balance between batch size and the number of epochs of pre-training for a given teacher and student model. The soft-label hypothesis and the special consideration for pre-training a teacher model should generalize to any response-based knowledge distillation process.

7.2 Future Works

In this work, we keep our on the quality of soft labels for better distillation and try to develop a theoretical understanding of the response-based knowledge distillation process. We talk specifically about the offline distillation process, but we believe that the same understanding should be applied to explain online and self-distillation processes. Similar experiments with some modifications can be used to analyze these two methods. We also keep our focus on classification problems in the visual domain, but this theory of response-based knowledge distillation can be explored in Natural Language Processing and Speech Recognition related problems in the future.

Appendices

APPENDIX A
TEACHER MODELS

A deep learning (DL) model with more hidden layers and/or more trainable parameters has a greater capacity to learn a given data distribution. Concepts like Dropout [20], and Batch Normalization [10] further help the learning capacity of a DL model. In this work, we roughly measure the learning capacities of models based on the number of layers, convolutional and fully connected dense layers, and the total number of trainable parameters. Table A.1 shows the different capacity teacher models used for all three datasets. Apart from showing the number of convolutional and fully connected layers, we also indicate the use of dropout and batch normalization techniques for each teacher model.

Table A.1: Details of teacher models for different datasets. Conv, Dense, Dropout and BN denotes number of convolution layer, number of Dense layer, dropout layer and Batch Normalization layer, respectively.

Dataset	Learning Capacity	Conv	Dense	Dropout	BN	Total Parameters
MNIST	Large	3	3	✓	×	2,560,906
	Small	2	1	×	×	1,433,610
F-MNIST	Large	4	4	✓	×	2,339,850
	Small	3	1	×	×	1,558,538
CIFAR-10	Large	8	3	✓	✓	26,902,442
	Small	3	1	×	×	5,674,634

We test the soft-label hypothesis for different teacher models with a very large number of parameters and a small number of parameters. For the MNIST dataset, we reduce the number of parameters of the small capacity teacher model to 55 % of the large capacity teacher. For Fashion-MNIST and CIFAR-10 datasets, we reduce the number of parameters of the small capacity teacher to 66 % and 21 % of the large capacity teacher model, respectively.

APPENDIX B
STUDENT MODELS

We follow a similar rule of thumb to decide the learning capacity of a student model as defined in appendix A. We keep the learning capacity of a student model much smaller than that of the small capacity teacher. For MNIST, we keep the number of parameters of the student model restricted to 0.8 % of the large capacity teacher model and 1.43 % of the small learning capacity teacher models. For Fashion-MNIST, the student model has 2.14 % and 3.22 % number of parameters as compared to the large and small capacity teacher models, respectively. For CIFAR-10, we keep the number of parameters of the student model restricted to 1.99 % and 9.42 % of large and small capacity teacher models, respectively.

Table B.1: Details of teacher models for different datasets. Conv, Dense, Dropout and BN denotes number of convolution layer, number of Dense layer, dropout layer and Batch Normalization layer, respectively.

Dataset	Student Type	Conv	Dense	Dropout	BN	Total Parameters
MNIST	General	2	1	×	×	20,490
	Baseline	×	1	×	×	7,850
F-MNIST	General	2	1	×	×	50,186
	Baseline	×	1	×	×	7,850
CIFAR-10	General	3	1	×	×	534,666
	Baseline	×	1	×	×	30,730

We significantly reduce the number of parameters in a student model by reducing the number of hidden layers, the number of filters for convolutional layers, and the number of units for fully connected layers. The details of the number of layers and number of parameters of each student model are shown in table **??**. We do not use dropout and batch normalization layers for any of the student models.

Table B.2: Details of baseline-students models for different datasets. Conv, Dense, Dropout and BN denotes number of convolution layer, number of Dense layer, dropout layer and Batch Normalization layer, respectively.

Dataset	Conv	Dense	Dropout	BN	Total Parameters
MNIST	×	1	×	×	7,850
F-MNIST	×	1	×	×	7,850
CIFAR-10	×	1	×	×	30,730

REFERENCES

[1] Cristian Buciluǎ, 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. New York, NY, USA: Association for Computing Machinery, Aug. 2006, pp. 535–541. ISBN: 978-1-59593-339-3.

[2] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. “BinaryConnect: Training Deep Neural Networks with binary weights during propagations”. In: *arXiv:1511.00363 [cs]* (Apr. 2016). 01928 arXiv: 1511.00363.

[3] Emily Denton et al. “Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation”. In: *arXiv:1404.0736 [cs]* (June 2014). 01206 arXiv: 1404.0736.

[4] Qianggang Ding et al. “Adaptive Regularization of Labels”. In: *arXiv:1908.05474 [cs, stat]* (Aug. 2019). 00004 arXiv: 1908.05474.

[5] Mengya Gao et al. “Residual Knowledge Distillation”. In: (), p. 9.

[6] Jianping Gou et al. “Knowledge Distillation: A Survey”. In: (), p. 36.

[7] Song Han et al. “Learning both Weights and Connections for Efficient Neural Networks”. In: *arXiv:1506.02626 [cs]* (Oct. 2015). 03320 arXiv: 1506.02626.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the Knowledge in a Neural Network”. In: (Mar. 2015).

[9] Andrew G. Howard et al. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *arXiv:1704.04861 [cs]* (Apr. 2017). 07974 arXiv: 1704.04861.

[10] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *arXiv:1502.03167 [cs]* (Mar. 2015). 26372 arXiv: 1502.03167.

[11] Seung Wook Kim and Hyo-Eun Kim. “TRANSFERRING KNOWLEDGE TO SMALLER NET- WORK WITH CLASS-DISTANCE LOSS”. In: (2017), p. 7.

[12] Alex Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”. In: (). 10621, p. 60.

[13] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (Nov. 1998). 35206, pp. 2278–2324.

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (May 2015). 36268, pp. 436–444.

[15] Seyed Iman Mirzadeh et al. “Improved Knowledge Distillation via Teacher Assistant”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020), pp. 5191–5198.

[16] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. “When Does Label Smoothing Help?” In: (), p. 13.

[17] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (Oct. 1986). 24512, pp. 533–536.

[18] C E Shannon. “A Mathematical Theory of Communication”. In: (). 84389, p. 55.

[19] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *arXiv:1409.1556 [cs]* (Apr. 2015). 55622 arXiv: 1409.1556.

[20] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: (). 27411, p. 30.

[21] Christian Szegedy et al. “Rethinking the Inception Architecture for Computer Vision”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 12443. Las Vegas, NV, USA: IEEE, June 2016, pp. 2818–2826. ISBN: 978-1-4673-8851-1.

[22] Yunhe Wang et al. “CNNpack: Packing Convolutional Neural Networks in the Frequency Domain”. In: (). 00117, p. 9.

[23] Jiaxiang Wu et al. “Quantized Convolutional Neural Networks for Mobile Devices”. In: *arXiv:1512.06473 [cs]* (May 2016). 00693 arXiv: 1512.06473.

[24] Han Xiao, Kashif Rasul, and Roland Vollgraf. “Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms”. In: *arXiv:1708.07747 [cs, stat]* (Sept. 2017). 02123 arXiv: 1708.07747.

[25] Xiyu Yu et al. “On Compressing Deep Models by Low Rank and Sparse Decomposition”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 00168. Honolulu, HI: IEEE, July 2017, pp. 67–76. ISBN: 978-1-5386-0457-1.

[26] Li Yuan et al. “Revisiting Knowledge Distillation via Label Smoothing Regularization”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2020, pp. 3902–3910. ISBN: 978-1-72817-168-5.

[27] Shuangfei Zhai et al. “Doubly Convolutional Neural Networks”. In: *arXiv:1610.09716 [cs]* (Oct. 2016). arXiv: 1610.09716.