Controlling the Quality of Distillation in Response-Based Network Compression

Vibhas Vats, David Crandall

Indiana University Bloomington
vkvats@iu.edu, djcran@indiana.edu

Abstract

The performance of a distillation-based compressed network is governed by the quality of distillation. The reason for the suboptimal distillation of a large network (teacher) to a smaller network (student) is largely attributed to the gap in the learning capacities of a given teacher-student pair. While it is hard to distill all the knowledge of a teacher, the quality of distillation can be controlled to achieve better performance. Our experiments show that the quality of distillation is largely governed by the quality of teacher's response, which in turn is heavily affected by the presence of similarity information in its response. A well-trained large capacity teacher loses similarity information between classes in the process of learning fine-grained discriminative properties for classification. The absence of similarity information causes the distillation process to be reduced from *one example-many class* learning to *one example-one class* learning, thereby throttling the flow of diverse knowledge from the teacher. With the implicit assumption that only the instilled knowledge can be distilled, instead of focusing only on the knowledge distilling process, we scrutinize the knowledge inculcation process. We argue that for a given teacher-student pair, the quality of distillation can be improved by finding the sweet spot between batch size and number of epochs while training the teacher. We discuss the steps to find this sweet spot for better distillation. We also propose *the distillation hypothesis* to differentiate the behavior of the distillation process between knowledge distillation and regularization effects. We conduct our experiments on three different datasets.

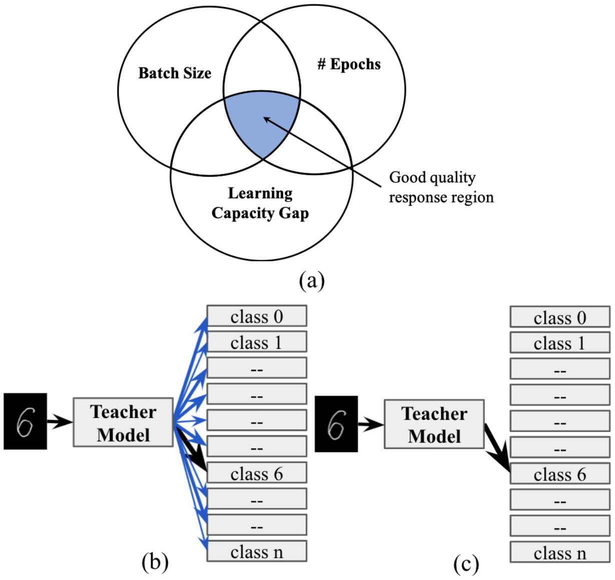

Figure 1: (a) Sweet spot between batch size and number of epochs for a given teacher-student pair for effective KD. (b) and (c) Difference in teacher's response in presence and absence of similarity information, respectively.

1 Introduction

Wider and deeper deep learning (DL) models have helped us build sophisticated AI systems across a broad range of areas (LeCun, Bengio, and Hinton 2015). But the huge computation and memory requirements of these models create hurdles in deploying them on edge devices (Simonyan and Zisserman 2015; Howard et al. 2017). A number of network compression techniques have been developed to compress large DL models into more efficient models with comparable performance, like model quantization (Wu et al. 2016), model binarization (Courbariaux, Bengio, and David 2016), parameter sharing (Han et al. 2015; Wang and Yoon 2021), low-rank factorization (Denton et al. 2014; Yu et al. 2017), convolution filter compression (Zhai et al. 2016), and knowledge distillation (KD) (Buciluǎ, Caruana, and Niculescu-Mizil 2006; Hinton, Vinyals, and Dean 2015). KD has two basic steps, knowledge extraction and distillation. Knowledge extraction can be response-based, feature-based, or relation-based and its distillation can be offline, online, or self-distillation (Gou et al. 2021). In this paper, we explore the KD process that uses a teacher's response to distill the knowledge in an offline manner to achieve network compression.

Recent advances suggest that the performance of the response-based offline knowledge distillation (RBKD) process is largely affected by the gap in learning capacities of the teacher-student pair (Yuan et al. 2020; Mirzadeh et al. 2020; Gao et al. 2020). However, we argue that this gap is not the root cause of the poor distillation.

Our experiments suggest that the quality of distillation

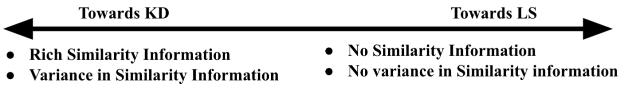
Towards KD
Towards LS
Rich Similarity Information
Variance in Similarity Information
No Similarity Information
No variance in Similarity information

Figure 2: The distillation hypothesis

is largely controlled by the amount of "similarity information" about classes that a teacher provides in its responses. In other words, teachers provide a distribution over the possible class labels for any given example. A highly-confident teacher produces a "hard" result that has a peak response for exactly one class, whereas a less-confident teacher provides a more spread, higher-entropy distribution. We find, somewhat counter-intuitively, that the less-confident teacher provides the student more information about the relationships between classes through the non-zero responses to the other classes. We find that careful training of the teacher can largely offset the negative effect of the gap between models. Fig. 1 (a) presents the idea. For a given teacher-student pair of any capacity gap, we can find a balance between batch size and number of epochs of training a teacher to retain similarity information in its response. Fig. 1 (b) and (c) show the fundamental difference in the quality of a teacher's response in the presence and absence of similarity information, respectively. A good response has relevant information about similarity between two classes and a poor response lacks this information. The RBKD process is fast and effective with the similarity-rich response as it learns about more than one class from a single input - *one example-many class* learning. It requires fewer examples per class to learn about the whole distribution of examples. In contrast, the process is slow, less effective, and requires more examples per class when similarity information is absent.

We also observe that for a fixed student network, as the teacher model becomes more powerful, and wider and deeper in size, it loses the rich similarity information present in the soft labels and behaves like a label smoothing (LS) (Müller, Kornblith, and Hinton 2019; Szegedy et al. 2016) process. We use this observation to differentiate the nature of the distillation process between KD and regularization effect through *the distillation hypothesis* shown in Fig. 2. We argue that any distillation process with little or no similarity information in the teacher's response brings more of a regularization effect than the distillation effect and when it completely lacks similarity information, the KD process is equivalent to LS.

Our contributions are as follows: (i) We show that rich similarity information in teacher's response can facilitate *one example-many classes* learning and accelerates KD. (ii) We argue that the gap in teacher-student pair is not the root cause or the dominating factor of poor distillation, and provide a method to carefully train any teacher to retain more similarity information in its response to achieve better distillation. (iii) We also propose *the distillation hypothesis* to understand the underlying nature of the distillation process.

2 Related Work

Buciluă, Caruana, and Niculescu-Mizil (2006) devise a method of network compression by distilling the knowledge of a teacher network, input to the softmax layer, to a student network. Hinton, Vinyals, and Dean (2015) popularized this idea by introducing a temperature term T, $q_i = \frac{exp(\frac{z_i}{T})}{\sum_j exp(\frac{z_j}{T})}$ class probabilities (q_i) and logits (z_i), in softmax that controls the softness of response. Increasing T decreases q_i for correct class, indicating decrease in the confidence of network's response, while increasing q_i for incorrect classes, indicating increase in similarity information in network's response.

The response-based network compression is a very effective tool to instill the knowledge of a cumbersome model into an efficient model to achieve uncompromised network compression. A common understanding is that a more powerful teacher should be able to provide more knowledge to its student. But it is observed that wider and deeper networks do poor distillation (Mirzadeh et al. 2020; Yuan et al. 2020; Gao et al. 2020). Kim and Kim (2017) and Müller, Kornblith, and Hinton (2019) ascribe the effectiveness of RBKD being similar to the effectiveness of LS. Ding et al. (2019) explain RBKD as the regularization effect brought about by the response of the teacher model. However, the RBKD process relies on the teacher's response and can not properly explain the hidden-layer supervision (Gou et al. 2021). We argue that whether RBKD behaves as regularization, distillation, or LS, depends on the presence of similarity information in the teacher's response. This, in tern, affects the extent of network compression that can be achieved by this method.

Mirzadeh et al. (2020) attribute the gap in learning capacities of teacher-student pair as the only reason for poor distillation, and propose a teacher assistant (TA) model, with a learning capacity in between teacher and student, to address this problem. The knowledge from the teacher is routed through one or more TA(s) to the student. While TAs improve the distillation performance but it also makes the whole process of network compression computationally expensive.

With a similar hypothesis, Yuan et al. (2020) propose a teacher-free KD (TFKD) process to reduce the gap between teacher-student pairs. The TFKD process distills its knowledge to itself during training. Theoretically, TFKD reduces the gap to zero, but its knowledge is limited by its learning capacity and extent of training. It also deviates from the network compression task.

Gao et al. (2020) propose residual KD to distill the knowledge by introducing an assistant model to learn the residual error between teacher and assistant models. This method also tries to reduce the gap between teacher-student pairs. This method is also effective, but it does bring additional computational costs with added networks to achieve network compression.

All the above methods try to find a solution to reduce the gap between the teacher-student pair. They focus specifically on the distillation process and not on the knowledge inculcation process. In this paper, we show that the quality of knowledge distillation can be controlled by controlling the quality of knowledge inculcation in a teacher. We also ex-

Input class	Response of LS process									
	0	1	2	3	4	5	6	7	8	9
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

Table 1: The LS labels for different input images of same class at $\alpha_{LS} = 0.6$. It shows that the distribution of soft-labels in the absence of similarity information. This distribution promotes *one example-one class* learning.

plain how the nature of KD changes with a change in similarity information in the teacher's response with *the distillation hypothesis*.

3 Response Based Techniques

Label Smoothing

The LS process uses a less confident form of one-hot label to train a model. It decreases the model confidence about the correct class and treats all other classes as equal, providing no similarity information (Müller, Kornblith, and Hinton 2019; Szegedy et al. 2016). The LS labels are computed as,

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

where, y_c is the one-hot vector, c is the number of classes, and $1/c$ is a uniform distribution. Examples are shown in Table 1. To calculate the cross-entropy over these labels, we denote the true soft-label distribution $q(c|x)$ (x is input) as $q^{'}$ for LS process, q for RBKD process and use p to denote distribution $p(c|x)$ generated by model (Yuan et al. 2020). The loss function is then,

$$\mathcal{L}_{LS} = (1 - \alpha_{LS})H(q,p) + \alpha_{LS}D_{KL}(u,p) \tag{2}$$

where $H(u)$ is a fixed entropy value of uniform distribution and D_{KL} is the Kullback-Leibler divergence (KL),

$$\begin{aligned} H(q^{'},p) &= -\sum_{c=1}^{c} q^{'}log(p) \\ &= (1 - \alpha_{LS})\ H(q,p) + \alpha_{LS}H(u,p) \\ &= (1 - \alpha_{LS})\ H(q,p) + \alpha_{LS}(D_{KL}(u,p) \\ &+ H(u)) \end{aligned} \tag{3}$$

The soft-labels on the LS process is shown in Table 1. The LS process use a fixed distribution to produce "hard" result that has a peak response for exactly one class, and assigns all other classes equal value without considering the similarity between classes. This paradigm promotes *one example-one class* learning (Fig. 1 (c)) i.e. each example provides useful information only about its class and treat all other classes equally.

Input class	Soft-labels by small learning capacity teacher									
	0	1	2	3	4	5	6	7	8	9
Digit 6	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
	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

Table 2: Soft-labels generated by small capacity teacher model with 3 hidden layers on MNIST data.

Input class	Soft-labels by large learning capacity teacher									
	0	1	2	3	4	5	6	7	8	9
Digit 6	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
	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

Table 3: Soft-labels generated by a large capacity teacher model with 6 hidden layers on MNIST data

Response-Based Knowledge Distillation

The RBKD process has the teacher produce probabilistic distribution for each class at a temperature T. It contains rich similarity information between classes to support *one examples-many classes* learning (Fig. 1 (b)) (Hinton, Vinyals, and Dean 2015). These information-rich labels are used to distill the knowledge to the student by minimizing the weighted sum of KL divergence and cross-entropy losses as,

$$\mathcal{L}_{KD} = (1 - \alpha_{KD})H(p,q) + \alpha_{KD}D_{KL}(p_T^t, p_T) \tag{4}$$

where $H(p,q)$, p, p_T, and p_T^t denote the cross-entropy loss for student with true labels (q), output of student model, output of student model softened at T and output of teacher model softened at T, respectively. α_{KD} is the contribution factor for loss function.

The loss functions of RBKD and LS processes have similar formulations as shown in Equation (2) and (4), but they use different methods for generating soft-labels. The only difference is the p_T^t in $D_{KL}(p_T^t, p_T)$, which is generated by a teacher model, and u in $D_{KL}(u,p)$, which is a uniform distribution (Yuan et al. 2020). This difference decides the extent of similarity information in the response, and that governs the nature of the distillation process. We discuss this later in Section 4. It is safe to conclude that LS is a special case of the RBKD process in which soft labels are generated by a constant distribution as prior knowledge instead of learned-knowledge of a pre-trained teacher.

Quality of Teacher's Response

The student mimics the teacher's response by minimizing the KL divergence between their response at T. The RBKD loss, Equation (4), is weighed higher α_{KD}, 0.99, for KD. Each example generates two types of soft-labels, for itself - *confidence label*, and for all other classes - *similarity labels*. The confidence label shows the confidence of a teacher for the correct class and the similarity labels provide a probabilistic value of similarity of all other classes with input class. Tables 1, 2, and 3 show the confidence label in bold text and similarity labels in plain text. The hyper-parameter α_{LS} is 0.6 for LS process in Table 1, while temperature T is 9 for RBKD process in Tables 2 and 3.

The quality of distillation is controlled by the similarity information in the teacher's response, which can be con-

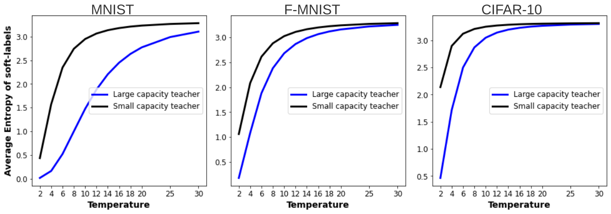

Figure 3: Average entropy of teacher's response for MNIST, Fashion-MNIST and CIFAR-10 datasets.

Missing	**KD by small capacity teacher**						**KD by large capacity teacher**					
class	Student accuracy(%) on removed class						Student accuracy(%) on missing class					
	T=3	T=6	T=9	T=12	T=15	T=20	T=3	T=6	T=9	T=12	T=15	T=20
Digit 0	99.18	99.28	99.48	99.18	98.67	98.57	2.8	0.20	0.61	0.0	0.0	0.0
Digit 1	98.67	98.67	98.67	98.59	98.50	98.23	0.0	0.26	0.0	0.26	0.08	0.0
Digit 2	95.73	96.80	96.12	95.83	94.86	94.76	2.5	0.38	0.58	0.48	1.25	0.19
Digit 3	98.11	98.41	98.31	98.11	98.01	97.92	0.0	0.79	1.38	6.23	2.47	2.27
Digit 4	98.47	98.67	98.57	98.67	98.37	96.94	0.10	0.10	0.0	0.0	0.0	0.0
Digit 5	96.86	97.53	97.30	97.19	96.41	97.86	0.67	0.56	1.35	2.80	2.46	4.26
Digit 6	97.39	97.49	97.28	97.18	97.18	96.65	17.32	2.08	12.83	7.09	18.58	7.41
Digit 7	96.10	96.30	96.40	95.91	94.84	95.03	0.29	0.29	3.50	0.58	0.77	1.84
Digit 8	97.12	96.71	97.22	96.61	96.40	96.09	0.00	0.10	0.0	0.0	0.0	0.0
Digit 9	96.13	96.63	96.23	95.83	95.63	94.25	0.00	0.69	0.10	1.48	1.28	2.18

Table 4: Student accuracy on missing class (MNIST) when distilled with small and large learning capacity teacher

Missing	**KD by small capacity teacher**						**KD by large capacity teacher**					
class	Student accuracy(%) on removed class						Student accuracy(%) on missing class					
	T=3	T=6	T=9	T=12	T=15	T=20	T=3	T=6	T=9	T=12	T=15	T=20
T-shirt	84.70	77.20	62.8	46.70	32.69	23.70	0.6	6	7.6	7	8.4	9
Trouser	94.90	94.19	92.90	91.39	88.09	89.99	0.0	4.2	15.4	19.2	23.2	16.8
Pullover	64.89	53.79	44.20	28.99	17.49	15.00	0.0	0.0	0.1	0.1	0.1	0.1
Dress	84.50	82.09	76.70	69.99	60.50	46.90	0.8	0.7	1.8	1.0	1.7	1.4
Coat	79.19	71.49	59.79	44.49	28.40	16.20	0.3	0.4	0.4	1.4	1.0	1.6
Sandal	91.50	92.00	90.49	87.30	84.89	82.30	0.2	0.4	0.4	0.4	0.4	0.5
Shirt	55.80	42.30	30.30	17.49	6.40	2.70	0.2	0.7	1.1	1.7	2.1	1.9
Sneaker	91.79	87.69	83.30	69.99	63.89	57.09	0.1	0.0	0.0	0.0	0.0	0.0
Bag	95.20	95.20	94.19	91.60	88.80	86.19	0.0	0.5	0.4	0.4	0.9	0.4
boot	91.29	89.99	87.99	83.39	79.69	74.00	1.1	2.6	4.4	3.6	3.7	3.7

Table 5: Student accuracy on missing class (Fashion-MNIST) when distilled with small and large learning capacity teacher

Missing	**KD by small capacity teacher**						**KD by large capacity teacher**					
class	Student accuracy(%) on removed class						Student accuracy(%) on missing class					
	T=3	T=6	T=9	T=12	T=15	T=20	T=3	T=6	T=9	T=12	T=15	T=20
plane	70.99	66.6	52.1	42.6	39.7	31.5	1.79	0.89	0.49	0.70	0.80	0.60
Auto	80.8	77.5	70.6	63.8	55.8	45	2.09	0.70	0.30	0.20	0.30	0.10
Bird	48.1	41.8	35	26.9	25.7	14	0.89	0.60	0.30	0.40	0.30	0.20
Cat	50.8	42.1	33	22.6	8.7	4.6	0.80	0.70	0.49	0.30	0.30	0.10
Deer	61.59	53.2	55.7	43.1	26.2	16.2	0.99	1.49	0.89	0.60	0.80	0.40
Dog	67.4	55.6	29.1	20.2	16	10.8	0.70	0.70	0.40	0.20	0.30	0.20
Frog	79.5	76.3	65.9	60.5	47.1	37.7	2.30	1.89	1.60	1.99	1.4	2.09
Horse	67.69	63.8	63.2	54.2	52.7	40	0.80	0.20	0.40	0.40	0.20	0.49
Ship	73.79	70.3	67.5	59.7	54	46.3	3.99	2.09	1.49	1.09	1.09	0.49
Truck	79.4	75.4	73.3	66.8	55.1	43.3	3.09	2.99	1.09	0.80	0.09	0.80

Table 6: Student accuracy on missing class (CIFAR 10) when distilled with small and large learning capacity teacher

trolled through T. But the increase and decrease in similarity information are not linear with T, but instead dependent on the similarity relation learned by the teacher during knowledge inculcation. The teacher perceives each example differently, even within the same class, and assigns a different confidence and similarity labels in its response. This introduces a variation in confidence and similarity labels —*a variance in response*, as shown row-wise and column-wise in Tables 2 and 3. More variance in similarity is desirable for KD as it provides knowledge about which classes are most similar to others. A highly confident teacher produces less variance as compared to a less confident teacher in its response, degrading the quality of distillation.

The behavior of a highly confident teacher model is analogous to the LS process. The similarity labels are treated equally with a little or no variance in response, providing neither similarity information nor variance in response. We describe this observation through the distillation hypothesis.

The Distillation Hypothesis

The distillation hypothesis, Figure 2, defines the nature of distillation by observing the quality of similarity labels and variance in response of a teacher. It states that for a given student network, as the learning capacity of the teacher network increases, the nature of the distillation process starts to move away from a similarity label-based RBKD process to a non-similarity-based LS process. The shift towards the LS process is caused by the loss of similarity information and variance in response of a teacher and leads to poor distillation. In other words, the nature of the knowledge distillation process shifts away from *one example-many classes* learning to *one example-one class* learning.

Since the quality of distillation is directly dependent on the knowledge inculcation process of a teacher. We argue that any teacher model can be trained to retain more similarity information by finding a sweet spot between the batch size and the number of epochs for a given teacher-student pair (see Figure 1 (a)). The blue region symbolizes the right balance between the batch size and the number of epochs for better knowledge distillation.

Every DL model has an optimal batch size for optimum learning. The optimal batch size is the number of examples that the model is effectively able to process. If the batch size is larger than its optimal value, then the model is not able to process all the information at once and takes more epochs to achieve high confidence. These two factors can be iteratively balanced to reach the optimal batch size and number of epochs for knowledge inculcation of a teacher.

4 Experiments and Results

First, we present empirical results to support the distillation hypothesis, and then show the results of improved distillation achieved using our proposed method on MNIST, Fashion-MNIST, and CIFAR-10 datasets.

Similarity Information in Teacher's Response

The entropy of the soft labels is directly proportional to the presence of similarity information in the teacher's response. It is calculated as $E_{Soft-labels} = -\sum_{i=1}^{C} p_i log(p_i)$ where C is the number of classes in the dataset, p_i is the probabilistic value from teacher. We use two different capacities of teachers, one with large learning capacity and the second with small learning capacity, to study this behavior. We argue that small capacity teachers having fewer parameters are not able to learn the fine-grained discriminative properties between classes and retain more similarity information in their response as compared to the large capacity teacher which can learn fine-grained properties and lose similarity information in the process.

We show two experiments to support our argument. First, we compare the average entropy in teachers' responses for

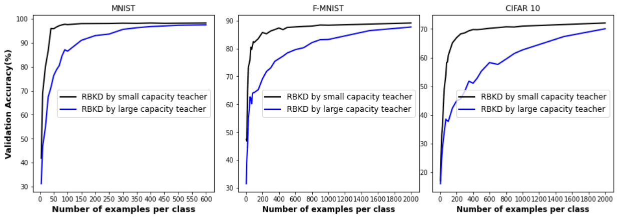

Figure 4: Entropy based example selection for knowledge distillation. RBKD process is more efficient if it requires lesser number of examples per class to distill the knowledge.

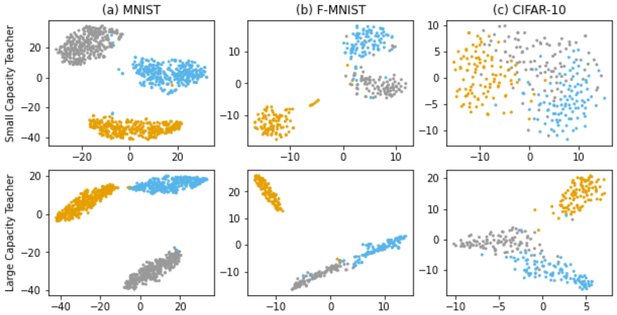

Figure 5: Penultimate layer representation of small and large capacity teacher models.

both small and large capacity teachers at different T, see Figure 3. At all T the average entropy of small capacity teachers remains higher than large capacity teachers. In other words, it is safe to say that a moderately confused teacher is better for knowledge distillation as compared to a well-trained large capacity teacher.

Second, we scrutinize the impact of similarity information for KD. We remove one out of N classes from distillation set during the KD process, and then check the performance of the student on the removed class, see Tables 4, 5, and 6. For the student network, the missing class is something it has never seen during training. It learns about the removed class only through the probabilistic value of the similarity information in the teacher's response.

The impact of similarity information on the quality of distillation is visible in Table 4, 5 and 6. Each table is divided into two columns, KD by the small and large capacity teachers. For each teacher, we perform KD on the same student and check its accuracy on the removed class on all three datasets. KD by the small teacher is very effective throughout but KD by the large teacher significantly under-performs on the removed class test. This shows that the quality of KD is controlled by the presence of similarity information in the teacher's response. In presence of similarity-rich soft responses to classes, the student can learn about more than one class from a single input example (one example- many class learning), while it can learn about only one class from one input example if the response is "hard". We further argue that this critical information in soft labels defines the nature of distillation to be KD, regularization, or LS. We use entropy as an indicator of similarity information in the rest

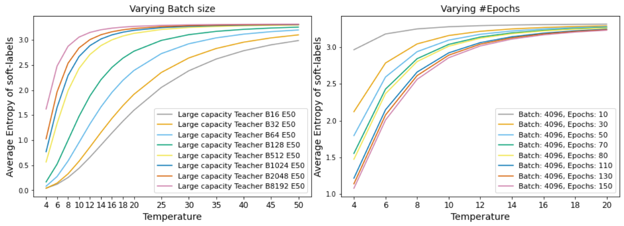

Figure 6: Variation in average entropy of soft-label vs temperature. Left, for different batch sizes at epoch 50. Right, for different epochs at batch size 4096 on MNIST dataset.

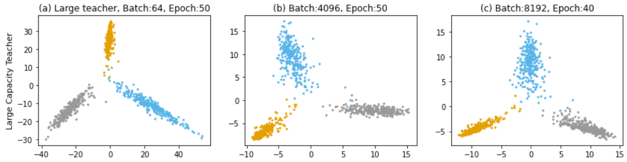

Figure 7: Improved penultimate layer representation for large capacity teach on MNIST. (a) shows the tight clusters, (b) and (c) shows similar clustering for different batch sizes and number of epochs.

of the paper.

KD is one example-many class learning

We established that similarity labels are the most critical information for KD. Now, we show that similarity information accelerates the KD process and only requires a handful of examples to distill knowledge from the teacher. With two different capacity teacher models and using entropy as an indicator at $T = 9$, we select different number of examples varying from 5 to 2000 per class for KD. We compare the highest validation accuracy achieved by the student network at each step as shown in Figure 4. The validation accuracy of the student network distilled by a large capacity teacher always remains smaller than by a small capacity teacher on each dataset.

For MNIST, distillation by a large capacity teacher requires 10 times more examples per class to achieve similar accuracy, and this gap increases for more complex datasets (15 times for Fashion-MNIST and 18 times for CIFAR-10). This shows that a large capacity teacher with less similarity information in its response fails to distill its knowledge using one example-many class learning and requires a much larger transfer set to achieve the same performance. This also establishes that one example-many class learning is a very important component of a good KD process.

Penultimate Layer Representation

Müller, Kornblith, and Hinton (2019) presents a visualization technique to understand the effects of the LS process on penultimate layer representations of a model. This method uses the linear projection of the activation of the penultimate layer to understand the change in representation. Müller, Kornblith, and Hinton (2019) argue that clusters that are relatively spread carry more similarity information in their response as compared to compact clusters.

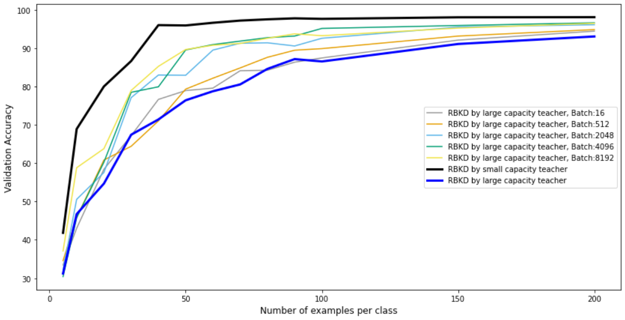

Figure 8: Improved number of examples required per class for distillation on MNIST

The distillation hypothesis says that with the increase in learning capacity of a teacher model it loses similarity information in its response. This loss of similarity indicates that the nature of distillation shifts towards the nature of the LS process. We use this experiment to corroborate the distillation hypothesis by showing the change in clustering of penultimate layer projection which directly relates to similarity information in soft labels.

We randomly select three classes and then choose 300 examples from each class randomly. We extract its activation from the penultimate layer of both small and large capacity models and plot its projection, see Figure 5. We observe that the penultimate layer representation of the stronger teacher model forms tighter clusters as compared to small capacity teachers for each dataset. The spread-out 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 provides strong evidence for the distillation hypothesis.

5 Similarity-Rich Knowledge Inculcation

We already established the importance of similarity information in a teacher's response. But instead of trying to improve the KD process, we focus on the knowledge inculcation process - training the teacher. We carefully calibrate the optimum value of batch size and number of epochs to train the teacher described in section 3. We show improvement for each experiment shown before.

Based on our hypothesis, we show the change in average entropy of teachers' responses for different batch sizes and number of epochs in Figure 6. It shows that to increase the entropy of soft labels, we can either increase in batch size at a fixed epoch or we can decrease the number of epochs for given batch size. By carefully and iteratively balancing these two factors, we can find one of the many possible sweet spots suited for KD. This controls the extent of similarity information in the teacher's response, which is responsible for the quality of KD and network compression in this process.

Next, we show the improvement in penultimate layer representation of the large capacity teacher by adapting to our method of knowledge inculcation in Figure 7. Figure 7(a)

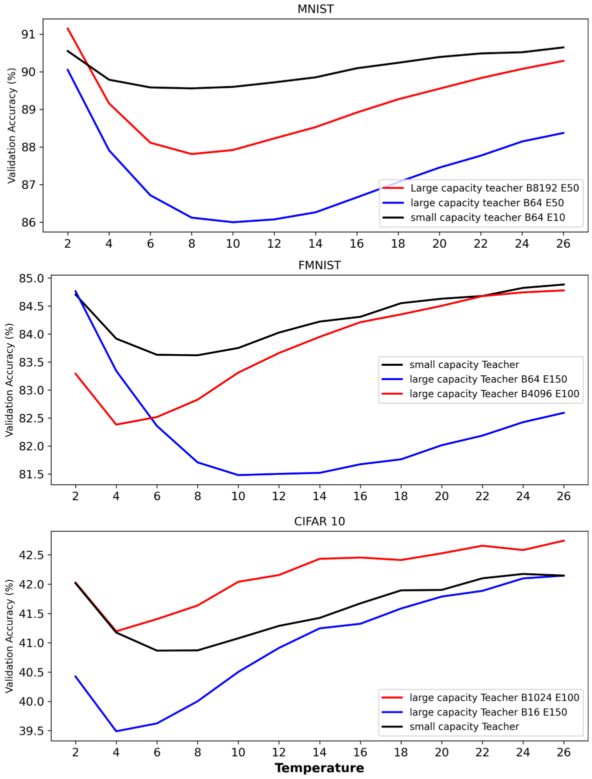

Figure 9: Improved performance on training with suitable batch size and epochs,

shows tight clusters, indicating a lack of similarity information. These clusters spread out for many combinations of batch sizes and number of epochs indicating more similarity information in the teacher's response. A similar clustering is observed for batch size 4096 - epochs 50 and for batch size 8192 - epochs 40 in Fig. 7(b) and (c), respectively. We can conclude that there can be many combinations of batch size and number of epochs for a given teacher-student pair for a more efficient KD.

With all these improvements in the quality of teachers' responses, the minimum number of examples needed to achieve optimal performance should also reduce for large teachers. As shown in Fig. 8, we improve around 5 times in the minimum number of examples required per class to achieve similar performance, decreased from 500 to 100 examples per class, for efficient KD. This provides further support to our idea of focusing on the knowledge inculcation process for achieving better and faster distillation, leading to better network compression.

Even with the above empirical results supporting our hypothesis, we still wanted to test our hypothesis for extreme teacher-student pair situations by maximizing the learning capacity gap between them. To achieve this, we use a single-layer student (baseline student) model with only a classification layer (we provide detail of models used in this paper in Table 7). We retrain the same large capacity teacher models for each dataset but train it differently - with different batch

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

Table 7: Details of teacher and student models for different datasets. Conv, Dense, Dropout, and BN denote the number of convolution layers, number of Dense layers, dropout layers, and Batch Normalization layers, respectively.

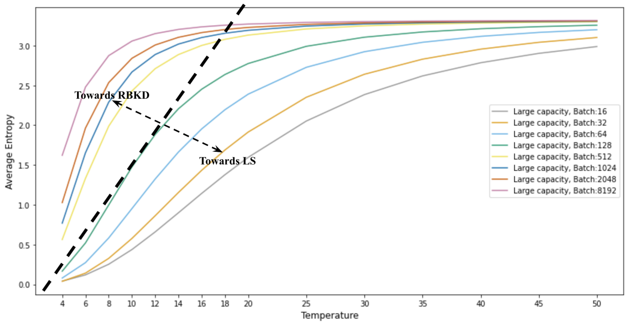

Figure 10: The distillation hypothesis in terms of average entropy of soft-labels.

sizes and number of epochs, to accommodate for the larger gap in model capacities and use it for KD to the baseline student. We also use the small capacity teacher model to show the relative improvements in Fig. 9. The black, blue, and red lines in Fig. 9 shows the performance of baseline students on RBKD by small, and large capacity teachers trained with different batch size and epochs at different temperatures. The gap between the red and the blue line shows the improvement in KD performance by a single-layer student.

6 Summary and Conclusion

For building intuition about the nature of distillation, we present the entropy-version of the distillation hypothesis (previously shown in Fig. 2) to approximate the nature of the distillation process in Figure 10. For all the processes lying on the extreme ends in Fig. 2, we can imaging multiple lines with different entropy spread across, as shown in Fig. 10. We can draw an imaginary line (the dashed line in the figure) that separates the RBKD process and the LS process. For efficient distillation from any teacher to any student, the corresponding entropy line should lie in the RBKD region. This can be achieved using our proposed method of knowledge inculcation during teacher's training. We emphasize that this representation is to develop an intuitive understanding of the KD process.

We discuss that the quality of distillation can be controlled at various stages of the KD process. While most works focuses on perfecting the distillation step of the KD process, we focus on controlling the quality of distillation by improving the knowledge inculcation process of the teacher model. Our experiments suggest that similarity information in a teacher's response plays a dictating role in determining the quality of distillation. We show the role of similarity information in accelerating the distillation process through one example-many class learning, making the KD process more effective and efficient. The distillation hypothesis can help in approximating the nature of the KD process to be RBKD, regularization, or LS process and to help make the process more effective. While we argue that the distillation performance of any teacher-student pair can be improved by our method of knowledge inculcation, we do not explore the extent to which the KD process can be improved. We believe that the factor of improvement should be different for each teacher-student pair. We leave this for future exploration.

7 Acknowledgements

This work was supported in part by the Naval Engineering Education Consortium (NAVSEA) through grant N00174-19-1-0010 with support of NSWC Crane Division, and by the Office of Naval Research under award N00014-19-1-2655.

References

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, 535–541. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-59593-339-3.

Courbariaux, M.; Bengio, Y.; and David, J.-P. 2016. BinaryConnect: Training Deep Neural Networks with binary weights during propagations. *arXiv:1511.00363 [cs]*. 01928 arXiv: 1511.00363.

Denton, E.; Zaremba, W.; Bruna, J.; LeCun, Y.; and Fergus, R. 2014. Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation. *arXiv:1404.0736 [cs]*. 01206 arXiv: 1404.0736.

Ding, Q.; Wu, S.; Sun, H.; Guo, J.; and Xia, S.-T. 2019. Adaptive Regularization of Labels. *arXiv:1908.05474 [cs, stat]*. 00004 arXiv: 1908.05474.

Gao, M.; Shen, Y.; Li, Q.; and Loy, C. C. 2020. Residual Knowledge Distillation. arXiv:2002.09168.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129: 1789–1819.

Han, S.; Pool, J.; Tran, J.; and Dally, W. J. 2015. Learning both Weights and Connections for Efficient Neural Networks. *arXiv:1506.02626 [cs]*. 03320 arXiv: 1506.02626.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861 [cs]*. 07974 arXiv: 1704.04861.

Kim, S. W.; and Kim, H.-E. 2017. Transferring knowledge to smaller networks with class-distance loss. In *ICLR*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature*, 521(7553): 436–444. 36268.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved Knowledge Distillation via Teacher Assistant. *AAAI*, 34(04): 5191–5198.

Müller, R.; Kornblith, S.; and Hinton, G. 2019. When Does Label Smoothing Help? In *NeurIPS*, 13.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*. ArXiv: 1409.1556.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. Las Vegas, NV, USA: IEEE. ISBN 978-1-4673-8851-1. 12443.

Wang, L.; and Yoon, K.-J. 2021. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–1.

Wu, J.; Leng, C.; Wang, Y.; Hu, Q.; and Cheng, J. 2016. Quantized Convolutional Neural Networks for Mobile Devices. *arXiv:1512.06473 [cs]*. 00693 arXiv: 1512.06473.

Yu, X.; Liu, T.; Wang, X.; and Tao, D. 2017. On Compressing Deep Models by Low Rank and Sparse Decomposition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 67–76. Honolulu, HI: IEEE. ISBN 978-1-5386-0457-1. 00168.

Yuan, L.; Tay, F. E.; Li, G.; Wang, T.; and Feng, J. 2020. Revisiting Knowledge Distillation via Label Smoothing Regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3902–3910. Seattle, WA, USA: IEEE. ISBN 978-1-72817-168-5.

Zhai, S.; Cheng, Y.; Lu, W.; and Zhang, Z. 2016. Doubly Convolutional Neural Networks. *arXiv:1610.09716 [cs]*. ArXiv: 1610.09716.