Self Regulated Learning Mechanism for Data Efficient Knowledge Distillation

1st Sourav Mishra  
Department of Aerospace Engineering  
Indian Institute of Science  
Bangalore, India  
srvmishra832@gmail.com

2nd Suresh Sundaram  
Department of Aerospace Engineering  
Indian Institute of Science  
Bangalore, India  
vssuresh@iisc.ac.in

Abstract—Existing methods for distillation use the conventional training approach where all samples participate equally in the process and are thus highly inefficient in terms of data utilization. In this paper, a novel data-efficient approach to transfer the knowledge from a teacher model to a student model is presented. Here, the teacher model uses self-regulation to select appropriate samples for training and identifies their significance in the process. During distillation, the significance information can be used along with the soft-targets to supervise the students. Depending on the use of self-regulation and sample significance information in supervising the knowledge transfer process, three types of distillations are proposed - significance-based, regulated, and hybrid, respectively. Experiments on benchmark datasets show that the proposed methods achieve similar performance as other state-of-the-art methods for knowledge distillation while utilizing a significantly less number of samples.

I. INTRODUCTION

Deep learning models have shown remarkable performance in several fields such as image classification [1], object detection [2], etc. However, deploying them on edge devices such as mobile phones and an on-board computer is not feasible due to their larger memory footprint. Therefore several methods have been proposed in the literature to address model compression without compromising generalization performance. Based on their assumption about knowledge representation, the methods can be divided into model compression-based methods and knowledge distillation based methods.

Model compression-based methods assume that knowledge is contained in the model’s weights and reduce the redundancies present in deep models. LeCun introduced neural network pruning in his paper on Optimal Brain Damage [3]. Various other methods for compressing neural networks have been proposed in the literature [4]–[8]. Model compression-based methods involve iterative pruning and fine-tuning of networks and are often time-consuming processes.

On the other hand, knowledge distillation based methods assume that the knowledge of a model is captured in its intermediate activations and outputs. Hence, smaller models known as students receive supervision from larger models called teachers and the ground truths. The probabilities assigned by the teacher to the incorrect classes constitute ‘dark knowledge,’ and it has been shown to improve the generalization ability of student models [9]. Based on the type of knowledge being transferred, knowledge distillation methods fall into one of two families - response-based and feature-based. Response based methods [9]–[12] transfer the knowledge from the teacher to the student by matching the outputs of their last layers, whereas feature-based methods [13]–[15] supervise the students by matching the activations of the intermediate layers of the teacher and the student models. [9] uses the original training data to perform distillation, [10]–[12], [16] propose methods to construct synthetic samples for distillation. These methods are based on the conventional way of training models where all the samples participate equally in learning the input-output mapping. Due to their inability to discriminate between samples based on their importance towards learning, these methods are very inefficient in terms of data usage and require large amounts of data.

However, in machine learning literature, it has been shown that metacognitive neural network achieve better generalization by employing self-regulation to select appropriate training samples for learning from stream-of-training data [17]–[19]. The heuristic strategies help in accounting for the different levels of knowledge present in different samples resulting in improved overall generalization and data-efficiency.

In this work, we address the data-efficiency issue of current state-of-the-art distillation methods by employing self-regulation. The teacher network uses an adaptive threshold to maximize the inter-class posterior probability difference during training. In this process, the samples on which it learns faster get filtered out from further training. The participation of each sample in training is monitored and is used to compute its significance value, which is a measure of its contribution to the teacher model’s knowledge. During knowledge transfer, the student model’s learning is driven by applying the computed sample significance information (sample significance based distillation), or by self-regulation alone (regulated distillation), or by a combination of both (hybrid distillation). The proposed methods (summarized in Figure 1) are data-efficient as they utilize significantly fewer training samples than other methods for knowledge distillation. The proposed distillation methods are evaluated on three benchmark data sets - MNIST, Fashion-MNIST, and CIFAR10. The results establish the data efficacy of the proposed distillation methods and their competitive performance with current state-of-the-art results reported in
the literature.

The main contributions of the paper is summarized below:

- For the first time in distillation literature, the data-efficiency issue is addressed.
- Self-regulation is proposed as a technique to improve data-efficiency as it accounts for the different levels of knowledge present in different samples.
- Three types of data-efficient approaches for knowledge transfer are proposed - sample significance based, regulated and hybrid. In sample significance based distillation, the significance information computed during teacher training is used to guide the student model’s learning. In regulated distillation, the student model employs self-regulation to learn from the soft targets produced by the teacher. In the hybrid strategy, both the above mechanisms are combined to guide the student.
- The proposed distillation schemes are evaluated on the benchmark datasets - MNIST, Fashion-MNIST, and CIFAR10. The proposed methods achieve similar or slightly better generalization performance than the current state-of-the-art distillation methods while utilizing much less data samples in the process.

II. RELATED WORKS

The idea of distillation was proposed in [20] and gained momentum in [9]. The student model is found to generalize better if supervised by the soft targets obtained at a high temperature from a bigger teacher model instead of the conventional way of training. This provides an easy method for transferring most of the generalization capacity of larger models to smaller models. Research in distillation is motivated by this observation. Apart from model compression, distillation has been successfully used in other applications as well. Recently, distillation has been applied in face recognition [21], cross-modal hashing [22] and collaborative learning [23]. Several methods have been proposed for distillation and, a comprehensive review is provided in [24].

A. Knowledge Distillation

[20] was one of the first papers to be proposed in knowledge distillation literature. It proposed to use the original training data as the transfer set for distillation. In addition to learning from the ground truths, the student also receives supervision from the teacher model in the form of soft targets computed at a high softmax temperature. Subsequently, [10], [12], [16] proposed methods for knowledge transfer wherein the transfer set was not available. [10] models the softmax space of the teacher network using a Dirichlet distribution and constructs synthetic data instances by inverting samples drawn from this distribution. [12] uses the teacher model as a fixed discriminator and trains a generator to produce synthetic samples. [16] relies on the availability of activation statistics of the teacher model and uses these as metadata for constructing a synthetic transfer set through inversion. In these works, a synthetic transfer set is constructed from the teacher model and it is used for knowledge transfer. The process of distillation is the same as in [9]. [25] proposed a conditional distillation mechanism. It incorporates the fact that the teacher model can sometimes be wrong in its predictions and only in that case the student must learn from the ground truths.

[13] propose to guide the feature extraction process in students by using supervision from intermediate layers of a teacher model. In case the output sizes of the layers involved in the transfer process do not match, a learnable convolutional regressor network is used to match the sizes. [14] propose to transfer the activations of the hidden neurons rather than the actual response values. They show that the generalization ability is better encoded by the decision boundaries formed by the hidden neurons rather than the actual response magnitudes. [15] propose to transfer knowledge between teacher and student models by matching their attentions instead of activations. They define different types of attentions for CNNs. These are computed at certain layers for the teacher and student models and are matched by minimizing the $L_p$ norm of their difference.

[26] propose to use the same network as the teacher and the student models and name the process as self distillation. The model from the previous epoch is used as the teacher. The proposed work is different from self distillation methods as teacher and student models are different networks of different sizes. Hence, self distillation methods are not used as baselines for comparison.

B. Self Regulation

In the conventional deep neural network training and knowledge distillation methods, all the training samples participate equally in capturing the input-output relationship. However, in machine learning literature, it has been shown that regulating the sample participation during training can lead to better generalization [17]–[19]. The conventional method for training disregards the relative importance of each sample in the dataset towards the knowledge of the model. Different samples contain different levels of knowledge. For example, some portions of a book are easier to grasp than others. The reader spends more time on those portions that he finds difficult and less on those portions that he finds easy. Self-regulation emulates this aspect of human learning in neural network training. In this way, the self-regulated learning process is much more efficient in terms of data usage than the conventional training method.

III. METHODS

In this section, the underlying mathematical and algorithmic details of the proposed data efficient knowledge distillation methods are described.

A. Self-Regulated Training and Sample Significance Computation

1) Self-Regulated Teacher Training: Training is made more data-efficient by controlling the participation of samples on which the model is able to learn faster. For this purpose, the regulation strategy proposed in [27] is modified.
While employing self-regulation, the model need not learn on a sample again if it is already too confident on it. So the model is able to distinguish between easy and hard samples based on an epoch dependent threshold and discards the easy samples from the process. In this way, the model learns to focus more on the difficult samples (which contribute more to its knowledge) than easy ones. The self-regulation process is explained below.

Given a dataset \( \mathbb{D} \) containing labeled samples \((x, y)\) and a model \(M\), the following quantities are monitored for all samples in all epochs \(N\):

- The predicted label, \(\hat{y}\):
  \[\hat{y} = \arg \max_x M(x)\]
- The difference between the maximum and the second maximum predicted probabilities, \(\delta\):
  \[\delta = \max_x M(x) - \max\{s \in M(x), s \neq \max M(x)\}\]

As the model learns to classify properly, the difference \(\delta\) gradually increases with the number of epochs \(n\). A sample is included in training if the predicted class is incorrect or if \(\delta\) is less than an epoch dependent adaptive threshold, \(\eta\). \(\delta\) will increase faster for easy samples compared to difficult samples. The purpose of the epoch dependent threshold function \(f(n)\) is to filter out such samples from training. Since \(\delta\) is the difference between the maximum and the second maximum probabilities, it is in the range \([0, 1]\). So the function \(f(n) : \mathbb{N} \rightarrow [0, 1]\) must be an increasing function of \(n\). So \(f(n) = 1 - \exp(-\alpha n)\) is chosen as threshold predictor, where \(\alpha\) is a hyperparameter. It maximizes the difference in the predicted posterior probabilities by allowing samples with a smaller growth rate of \(\delta\) to participate more in training. The method is shown in Figure 2.

2) Computation of Sample Significance: As explained earlier, all the samples in the dataset will not contribute equally to the knowledge of the model. The training process must distinguish samples accordingly to enhance generalization ability of the model. Self-regulation introduced in the previous subsection is a method of doing this. In the conventional training scheme, all the samples present in the dataset \(\mathbb{D}\) participate equally. That is, if the model is trained for \(N\) epochs, then each sample participates exactly \(N\) times in the training. However, with self-regulation, each sample participates \(\leq N\) times in the process, with the difficult ones participating more often than the easy ones. In such a scheme, the number of times a sample participates in the training process can be seen as a measure of its contribution to the knowledge of the model.

The significance of a sample is defined its class-wise min-max normalized participation. It is a number between 0 and 1. Let the significance of a sample be denoted by \(\hat{n}\) and the number of times it participates in the training be \(n\). Let \(S_i\) denote the subset of samples in the dataset \(\mathbb{D}\) that belong to the class \(i\) out of a total of \(C\) classes. That is,

\[S_i = \{x | (x, y) \in \mathbb{D}, \text{ and } y = i, i \in \{0, 1, ..., C - 1\}\} \quad (1)\]

For a given sample \((x, y)\), the significance \(\hat{n}\) is defined by:

\[\hat{n} = \frac{n - \min_{x \in S_i} n}{\max_{x \in S_i} n - \min_{x \in S_i} n} \quad (2)\]

The sample significance information is computed during the self-regulated training of the teacher model. It serves as a measure of the importance of the sample to the teacher’s knowledge and is used during distillation to transfer the different levels of knowledge present in the different samples.

The self-regulated teacher training and the sample significance computation processes are described in algorithms 1 and 2 respectively.

B. Data Efficient Distillation Methods

1) Sample Significance Based Knowledge Distillation: In conventional knowledge distillation, the soft targets computed from the teacher model at a temperature \(\tau\) are used. Given a teacher network \(T\) parametrized by \(\theta_T\) and a student network \(S\) parametrized by \(\theta_S\), distillation minimizes the following objective over all samples \((x, y)\) in the transfer set \(\mathbb{D}\):

\[\eta = f(n) = 1 - \exp(-\alpha n)\]

\[n = 0, 1, 2, \ldots, N - 1\]
Algorithm 1 Teacher model training with self-regulation

Input: Teacher network $T$, with parameters $\theta_T$ (without output softmax), dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^t$, epochs $N$, parameter $\alpha$ for self-regulation

Output: Parameters of trained Teacher network $\theta_T$, array of sample participations $\nu$, size of $\nu$ is same as $|\mathcal{D}| = t$

1: Initialize participations $\nu$
2: for $k$ in range($t$) do
3: $\nu[k] = 0$
4: end for
5: Train teacher and record participations
6: for $i$ in range($N$) do
7: $\eta = 1 - \exp(\alpha i)$
8: for $j$, $(x, y) \in \text{enumerate}(\mathcal{D})$ do
9: $z_T = T(x)$
10: $\hat{y}_T = \text{softmax}(z_T)$
11: $\hat{y} = \arg\max y_T$
12: compute $\delta$ from $y_T$
13: if $\hat{y} \neq y$ or ($y == y$ and $\delta < \eta$) then
14: $L = L_{CE}(y, \hat{y})$
15: update teacher’s parameters: $\theta_T' = \theta_T - \nabla_{\theta_T} L$
16: $\theta_T = \theta_T'$
17: $\nu[j] = \nu[j] + 1$
18: else
19: continue
20: end if
21: end for
22: end for
23: return $\theta_T, \nu$

Algorithm 2 Computation of Sample Significance

Input: Sample Participation Statistics recorded during teacher training $\nu$, The Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^t$ for which the participation is recorded. Dataset $\mathcal{D}$ has a total of $C$ classes labeled as $0, 1, ..., C - 1$

Output: Sample significance vector $\hat{\nu}$. The sizes of $\hat{\nu}, \nu$ and $\mathcal{D}$ are the same.

1: for $i$ in range($t$) do
2: $(x, y) = \mathcal{D}[i]$
3: $n = \nu[i]$
4: compute $\hat{n}$ from equation (2)
5: $\hat{\nu}[i] = \hat{n}$
6: end for
7: return $\hat{\nu}$

$$L = \sum_{(x,y) \in \mathcal{D}} L_{KD}(S(x, \theta_S, \tau), T(x, \theta_T, \tau)) + \lambda L_{CE}(\hat{y}_S, y)$$ (3)

where, $L_{KD}$ is the distillation loss which is minimized at a temperature $\tau$. It can be the cross entropy loss for classification or the $L_2$ loss for regression. $L_{CE}$ is the cross entropy loss which is minimized at a temperature of $1$. $\hat{y}_S$ is the prediction of the student network on the sample $x$ and $\lambda$ is a hyperparameter to balance the two losses.

In sample significance based distillation, the sample significance information computed above is used to direct the student model’s learning along with the soft targets. The loss function thus becomes sample specific and accounts for the different levels of knowledge to be transferred from the teacher model for the different samples in the dataset. The loss incurred on each sample is scaled by its significance computed during teacher training. For sample significance based distillation, the sample significance $\hat{n}$ is also included as a part of the dataset. The loss function is given by:

$$L_{new} = \sum_{(x,y,\hat{n}) \in \mathcal{D}} \hat{n} L_{KD}(S(x, \theta_S, \tau), T(x, \theta_T, \tau)) + \lambda \hat{n} L_{CE}(\hat{y}_S, y)$$ (4)

In this distillation process, the student receives maximum guidance from the teacher - in the form of soft targets and the sample significance information. As the teacher model is of larger capacity than the student model, it is expected that the samples which were difficult for the teacher will be difficult for the student model as well. So the student must put more focus on such samples during the knowledge transfer process.

2) Regulated Knowledge Distillation: The student model is trained by using the self-regulation strategy proposed in the first subsection but it does not use the sample significance information. In this scheme, the student model is given freedom to discriminate between the samples on its own through self-regulation just like the teacher model. The student model may find a different set of easy and difficult samples compared to the teacher model. The teacher is used to supervise the student through soft targets just like in conventional distillation [9].

3) Hybrid Knowledge Distillation: The student model is trained by using the sample significance information as well as by using the proposed self-regulation strategy. In this scheme, two effects are taking place simultaneously. The student model is trying to learn independently through self-regulation and at the same time it receives additional guidance in the form of sample significance information to focus more on the samples that the teacher model found tough during its training. Algorithm 3 shows the implementation of these distillation methods. The distillation methods are summarized in Figure 1.

IV. EXPERIMENTS

This section describes the implementation of the proposed algorithms. The generalization performance and sample efficiency of the proposed methods are evaluated. The MNIST,
Algorithm 3 Distillation Algorithms

Input: Pre-trained Teacher network \( T \), Student network \( S \) with parameters \( \theta_S \) (without output softmax), dataset \( \mathbb{D} = \{(x_i, y_i)\}_{i=1}^N \), epochs \( N \), parameter \( \alpha \) for self-regulation, temperature \( \tau \) for distillation, hyperparameter \( \lambda \), distillation mode - significance, regulated, hybrid, sample significance information (in case of significance based distillation) \( \hat{v} \)

Output: Parameters of trained Student network \( \theta_S \)

1: for \( i \) in range(\( N \)): do
2: \( \eta = 1 - \exp(\alpha i) \)
3: for \( j, (x, y) \in \text{enumerate}(\mathbb{D}) \): do
4: \( z_T, z_S = T(x), S(x) \)
5: \( \hat{y}_T, y_S = \text{softmax}(z_T/\tau), \text{softmax}(z_S/\tau) \)
6: \( \hat{y}_S = \text{softmax}(z_S) \)
7: \( \hat{y} = \arg \max y_S \)
8: compute \( \delta \) from \( y_S \)
9: \( \hat{n} = \hat{v}[j] \)
10: if mode == regulated: then
11:   if \( \hat{y} \neq y \) or (\( \hat{y} == y \) and \( \delta < \eta \)): then
12:     \( L = L_{KD}(\hat{y}_T, \hat{y}_S) + \lambda \hat{n} L_{CE}(y, y'_S) \)
13:     update student’s parameters:
14:     \( \theta'_S = \theta_S - \nabla \theta_S L \)
15:     \( \theta_S = \theta'_S \)
16:   else
17:     continue
18: end if
19: if mode == significance: then
20:     \( L = \hat{n} L_{KD}(\hat{y}_T, \hat{y}_S) + \lambda \hat{n} L_{CE}(y, y'_S) \)
21:     update student’s parameters:
22:     \( \theta'_S = \theta_S - \nabla \theta_S L \)
23:     \( \theta_S = \theta'_S \)
24: end if
25: if mode == hybrid: then
26:   if \( \hat{y} \neq y \) or (\( \hat{y} == y \) and \( \delta < \eta \)): then
27:     \( L = \hat{n} L_{KD}(\hat{y}_T, \hat{y}_S) + \lambda \hat{n} L_{CE}(y, y'_S) \)
28:     update student’s parameters:
29:     \( \theta'_S = \theta_S - \nabla \theta_S L \)
30:     \( \theta_S = \theta'_S \)
31:   else
32:     continue
33: end if
34: end for
35: return \( \theta_S \)

FashionMNIST, and the CIFAR10 datasets are used for the experiments. The batch size is set to 512, the distillation temperature \( \tau \) is set to 20, the normal temperature is set to 1, the hyperparameter \( \lambda \) is set to 0.3. The Adam optimizer [28] is used to train the models. Models are evaluated at the normal temperature. The algorithms are evaluated based on their accuracy on the test set. All implementations are done in pytorch. Two NVIDIA GeForce RTX 2080Ti cards are used for the experiments. The following sections describe the experiments and the results.

A. Self-Regulated Teacher Training and Data Efficient Distillations

1) Training Teacher Model with Self-Regulation: First, the teacher models are trained. The results are shown in Table 1. Conventionally training the models is equivalent to setting \( \alpha = \infty \) in algorithm 1. It is observed that the self-regulated training performs comparably to the conventional method of training. In the Fashion-MNIST case, the self-regulated training performs better than the conventional training method for all values of \( \alpha \) considered.

2) Distillation Results on MNIST: It is a dataset of handwritten digits. Each sample is a 28 x 28 grayscale image. The training set has 60000 labeled samples and the test set has 10000 labeled samples. Lenet-5 model [29] is used as the teacher network and Lenet-5 half model as the student network. The models are trained for 200 epochs. The teacher model is trained with a learning rate of 0.001 and \( \alpha = 0.02 \) at the normal temperature. A learning rate of 0.01 is used for distillation. The same value of \( \alpha \) is used in the ‘regulated’ and ‘hybrid’ distillation processes. Table 2 compares the results of the proposed methods against other response based knowledge distillation methods available in the literature.
3) Distillation Results on FashionMNIST: It is a dataset of 10 fashion items. Each sample is a 28×28 grayscale image. The training set has 60000 labeled samples and the test set has 10000 labeled samples. Lenet-5 model [29] is used as the teacher network and Lenet-5 half model as the student network. The models are trained for 200 epochs. The teacher model is trained with a learning rate of 0.001 and α = 0.04 at the normal temperature. A learning rate of 0.01 is used for distillation. The same value of α is used in the ‘regulated’ and ‘hybrid’ distillation processes. Table 3 compares the results of the proposed methods against other response based knowledge distillation methods available in the literature.

4) Distillation Results on CIFAR10: It is a dataset of 10 items as classes and each class contains 6000 samples. Each sample is a 32×32 colour image. The training set contains 50000 labeled samples and the test set contains 10000 labeled samples. Alexnet [1] model is used as the teacher network and Alexnet half model as the student network. The models are trained for 1000 epochs. The teacher model is trained with a learning rate of 0.001 and α = 0.04 at the normal temperature. A learning rate of 0.001 is used for distillation. The same value of α is used in the ‘regulated’ and ‘hybrid’ distillation processes. Table 4 compares the results of the proposed methods against other response based knowledge distillation methods available in the literature.

On simpler datasets such as MNIST, there isn’t much difference in performance across the distillation methods. On more realistic datasets like CIFAR10, hybrid distillation might perform slightly better than others. Regulated and hybrid distillations are expected to perform better on realistic scenarios because the student model is given the freedom to discriminate between the samples from the dataset through self-regulation. In this way, it can learn in a better way. The proposed methods perform better than most of the state-of-the-art methods (Tables 2-4). However, [9] performs better than the proposed methods because it uses all the samples available in the dataset. The advantage of the proposed methods is that they do not use all the samples. They are highly efficient in terms of data usage, as explained in the next section.

B. Evaluation of Sample Efficiency

To establish the data efficiency of the proposed self-regulated training method, the sample significance ˆn extracted during teacher training is visualized as a histogram for each class. Figures 3-5 show these. These numbers are used as weights in the ‘significance based distillation’ and the ‘hybrid distillation’ processes.

Most samples are insignificant towards learning as indicated by the large frequency bars in the 0.0-0.25 bin on each plot. This is because the teacher model learns fast on these samples, so they participate less often in training. For the FashionMNIST dataset, we observe that classes 0, 2, 4, and 6 have a similar shape of the sample significance histograms. These class indices correspond to T-Shirt, Coat, Pullover, and Shirt classes. Since these objects have similar appearances, the model needs to see them more often to be able to classify them properly.

Since the ‘significance based distillation’ process is similar to the conventional distillation process [9], its data efficiency is not evaluated. The total sample participation across all epochs for the ‘regulated distillation’ and the ‘hybrid distillation’ processes are reported. It is also reported as a percentage of all samples available for distillation across all epochs. This helps to compare the data efficiency of the proposed methods relative to the conventional distillation process [9]. The results are tabulated in Table 3. Sample participation is relatively higher for the hybrid method. This is because the student model does not learn in the same way through self-regulation as the teacher model and the sample significance data ˆn, used as weights) are obtained during teacher training. Mathematically, the sample efficiency ζ is defined as:

\[
\zeta = \frac{\sum_{i=1}^{t} v[i]}{N/|D|} = \frac{\sum_{i=1}^{t} v[i]}{N_t}
\]

where v is the array of sample participations used in Algorithms 1 and 2.

For example, we observe that the total sample participation across all epochs is 85528 for the MNIST dataset in the ‘regulated distillation’ process. However, we perform distillation for 200 epochs and 60000 samples are available for it in every epoch, making a total of 12000000 samples. This would be the total sample participation across all epochs for a normal distillation process [9]. So, we report 85528 as a percentage of 12000000. This is the sample efficiency ζ.

In addition to being data-efficient, the proposed methods perform comparable to other state-of-the-art data-free methods (as shown in Tables 2-4) for distillation. The original training data is used as the transfer set and the sample participation shows that the proposed methods use much less data (<20%) for distillation and training in general while achieving similar or better performance compared to other state-of-the-art methods.
Fig. 3. MNIST: Classwise sample significance extracted during teacher training with self regulation at $\alpha = 0.02$. x axis denotes the sample significance $\hat{n}$ and y axis denotes the frequency.

Fig. 4. FMNIST: Classwise sample significance extracted during teacher training with self regulation at $\alpha = 0.04$. x axis denotes the sample significance $\hat{n}$ and y axis denotes the frequency.

Fig. 5. CIFAR10: Classwise sample significance extracted during teacher training with self regulation at $\alpha = 0.04$. x axis denotes the sample significance $\hat{n}$ and y axis denotes the frequency.
TABLE V
SAMPLE EFFICIENCY θ OF PROPOSED DISTILLATIONS. THE FIRST NUMBER
DENOTES THE TOTAL SAMPLE PARTICIPATION IN THE DISTILLATION
PROCESS ACROSS ALL EPOCHS. THE SECOND NUMBER DENOTES ALL THE
SAMPLES AVAILABLE FOR DISTILLATION ACROSS ALL EPOCHS. THE
NUMBER IN THE PARENTHESES IS THE FIRST NUMBER EXPRESSED AS A
PERCENTAGE OF THE SECOND NUMBER.

| Dataset | Regulated Distillation | Hybrid Distillation |
|---------|------------------------|---------------------|
| MNIST   | (~0.713%)              | (~2.085%)           |
| FMNIST  | (~11.54%)              | (~17.474%)          |
| CIFAR10 | (~9.106%)              | (~9.211%)           |

V. CONCLUSIONS

Existing methods for knowledge distillation are not efficient in terms of data utilization, whether it is data from the original training set or synthetically constructed. Self regulated training mechanism was proposed for training the teacher model to identify the samples which contributed significantly to its knowledge. It is observed that self-regulation results in data-efficient training and achieves similar or better generalization performance than the conventional method of training models. Based on the sample significance information computed during teacher training and self-regulation, three types of distillation schemes are proposed. It is observed that regulated and hybrid variants of distillation are better suited to the knowledge transfer process in more realistic scenarios as the student has the freedom to learn on its own through self-regulation. Experiments on benchmark datasets establish the data efficacy of the proposed distillation methods (these use < 20% of the training data during distillation) and their competitive performance with other state-of-the-art distillation methods.

ACKNOWLEDGMENT

The authors would like to acknowledge WIRIN (WIPRO IIsc Research Initiative) for the financial support.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Advances in Neural Information Processing Systems, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012, pp. 1097–1105.

[2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[3] Y. LeCun, J. S. Denker, and S. A. Solla, “Optimal brain damage,” vol. 2, pp. 598–605, 1990. [Online]. Available: http://papers.nips.cc/paper/250-optimal-brain-damage.pdf

[4] S. Han, J. Pool, J. Tran, and W. J. Dally, “Learning both weights and connections for efficient neural networks,” in Advances in Neural Information Processing Systems, 2015.

[5] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” in 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, 2016.

[6] A. Sandryhaila and J. M. Moura, “Discrete signal processing on graphs,” IEEE Transactions on Signal Processing, 2013.

[7] E. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, “Exploiting linear structure within convolutional networks for efficient evaluation,” in Advances in Neural Information Processing Systems, 2014.

[8] Y. Wang, C. Xu, S. You, D. Tao, and C. Xu, “CNNpack: Packing convolutional neural networks in the frequency domain,” in Advances in Neural Information Processing Systems, 2016.

[9] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in NIPS Deep Learning and Representation Learning Workshop, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[10] G. K. Nayak, K. R. Mopuri, V. Shaj, R. Venkatesh Babu, and A. Chakraborty, “Zero-shot knowledge distillation in deep networks,” 36th International Conference on Machine Learning, ICML 2019, vol. 2019-June, pp. 8317–8325, 2019.

[11] A. Kimura, Z. Gharahramani, K. Takeuchi, T. Iwata, and N. Ueda, “Few-shot learning of neural networks from scratch by pseudo example optimization,” British Machine Vision Conference 2018, BMVC 2018, pp. 1–12, 2019.

[12] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, “Data-free learning of student networks,” Proceedings of the IEEE International Conference on Computer Vision, vol. 2019-Octob, no. Iccv, pp. 3513–3521, 2019.

[13] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “FitNets: Hints for thin deep nets,” 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–13, 2015.

[14] B. Heo, M. Lee, S. Yun, and J. Y. Choi, “Knowledge transfer via distillation of activation boundaries formed by hidden neurons,” 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, pp. 3779–3787, 2019.

[15] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings, pp. 1–13, 2017.

[16] R. G. Lopes, S. Fenu, and T. Starner, “Data-Free Knowledge Distillation for Deep Neural Networks,” LLDB Workshop at Neural Information Processing Systems (NIPS), 2017. [Online]. Available: http://arxiv.org/abs/1710.07535

[17] G. Sateesh Babu and S. Suresh, “Meta-cognitive Neural Network for classification problems in a sequential learning framework,” Neurocomputing, 2012.

[18] R. Savitha, S. Suresh, and N. Sundararajan, “Metalearning in a fully complex-valued radial basis function neural network,” 2012.

[19] G. S. Babu and S. Suresh, “Meta-cognitive RBF Network and its Projection Based Learning algorithm for classification problems,” Applied Soft Computing Journal, 2013.

[20] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, “Model compression,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

[21] J. Feng, H. Wang, H. R. Hu, L. Yu, W. Wang, and S. Wang, “Triplet distillation for deep face recognition,” in 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 808–812.

[22] H. Hu, L. Xie, R. Hong, and Q. Tian, “Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3120–3129.

[23] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, and P. Luo, “Online knowledge distillation via collaborative learning,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11017–11026.

[24] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” 2020. [Online]. Available: http://arxiv.org/abs/2006.05525

[25] Z. Meng, J. Li, Y. Zhao, and Y. Gong, “Conditional Teacher-student Learning,” ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2019-May, pp. 6445–6449, 2019.

[26] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, “Born-Again Neural Networks,” 2019. [Online]. Available: http://arxiv.org/abs/1710.07535

[27] S. Han, M. Pedram, W. J. Dally, and Y. Sang, “Low-power neural network,” in Proceedings of the IEEE Symposium on VLSI Circuits, 2015.

[28] G. S. Babu and S. Suresh, “Meta-cognitive Neural Network for classification problems in a sequential learning framework,” Neurocomputing, 2012.

[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, 1998.