KNOWLEDGE DISTILLATION USING UNLABELED MISMATCHED IMAGES

Mandar Kulkarni(*), Kalpesh Patil(**), Shirish Karande(*)
TCS Innovation Labs, Pune, India (*), IIT Bombay, Mumbai, India(**)
{mandar.kulkarni3,shirish.karande}@tcs.com, kalpeshpatil@iitb.ac.in

ABSTRACT

Current approaches for Knowledge Distillation (KD) either directly use training data or sample from the training data distribution. In this paper, we demonstrate effectiveness of 'mismatched' unlabeled stimulus to perform KD for image classification networks. For illustration, we consider scenarios where this is a complete absence of training data, or mismatched stimulus has to be used for augmenting a small amount of training data. We demonstrate that stimulus complexity is a key factor for distillation's good performance. Our examples include use of various datasets for stimulating MNIST and CIFAR teachers.

1 INTRODUCTION AND RELATED WORKS

Knowledge Distillation (KD) is the process of transferring the generalization ability of a teacher model (usually large neural network) to a student model (usually a small neural network). Hinton et al. (2015) demonstrated that the student network can be trained using an input data with combination of hard labels as well as soft labels. The hard labels are the ground truth labels (one-hot vectors) available for the training data while the soft labels are the output of the teacher network on the input data. Most of the distillation approaches either directly use training data or device a strategy to learn the training data distribution and then sample from it. However, the assumption of availability of labeled training data may not always hold true, due to various reasons.

In this paper, we investigate an effectiveness of 'mismatched' unlabeled stimulus for KD when the training data of the teacher is not available. Specifically, for the MNIST teacher, we utilize mismatched stimulus such as CIFAR(Krizhevsky & Hinton (2009)), STL(Coates et al. (2010)), Shape(Bengio et al. (2009)) and Noise. For a CIFAR teacher, we use stimulus such as 120k Tiny Imagenet(Torralba et al. (2008),tin), MNIST (LeCun et al. (1998)), Shape, SVHN (Netzer et al. (2011)), DTD-Texture (Cimpoi et al. (2014)). We observe that for CNN architectures these stimuli provide a surprisingly efficient distillation to student networks. We study an effect of complexity of the stimulus dataset on the distillation performance by using stimulus of varied complexity. Experimental results clearly demonstrate that the complexity of stimulus plays an important role in distillation, where more complex dataset appear to give better generalization performance. We also consider a scenario where a small labeled training set is available. In such cases, unlabeled stimulus can be very effective for data augmentation.

2 PRELIMINARIES

2.1 METHODOLOGY

The training objective of the distillation process Romero et al. (2014) can be written as follows

$$L(W_S) = H(P_T, P_S) + \beta H(y_{true}, P_S) \tag{1}$$

where W_S indicates weights of the student network, H indicates the cross entropy and β is the relative weights of two terms. The second term in the equation corresponds to a traditional cross entropy loss between output of a student network and labels (y_{true}). Let D denotes the data used for distillation. P_T is the posterior output of the teacher network on D while P_S is the posterior output of the student. The first term in the equation attempts to make the posterior of the student similar to that of teacher for the input data D. If the training data is not available, second term in the Eq. 1

cannot be used. A student is trained to optimize only the first term in Eq. 1. In case of availability of a small labeled training set, we use combination of this labeled set and unlabeled stimulus to train a student. For the unlabeled stimulus, we assume an uniform distribution over all classes while using it in the second term of Eq. 1.

2.2 Description of datasets and networks

We use teachers trained on two well known image classification datasets, MNIST and CIFAR-10.

2.2.1 MNIST

A CNN is trained on the dataset which has approx. 478k params. We experiment with two student architectures. In the first experiment, we use a student with the same architecture as the teacher. In the second experiment, we attempt to distill the teacher network into a relatively smaller student CNN. The student network has approx. 35% less parameters than the teacher. The details of the teacher and student architectures are provided in appendix. For MNIST teacher, we use mismatched stimulus such as CIFAR-10 , STL-10, Shape dataset and uniform random noise [-0.3,0.7]. In case of CIFAR and STL data, images are converted to grayscale and resized appropriately.

2.2.2 CIFAR

We train a 12 layer CNN on the CIFAR dataset which has approx.3M params. Here as well, we experiment with two student architectures, one which is same as the teacher and other smaller CNN which has approx. 10 times less parameters than the teacher. For CIFAR teacher, we use mismatched stimulus such as MNIST, SVHN, Shape, Texture, uniform noise and a slightly similar dataset, 120k TinyImagenet. As pointed out by the reviewer, a subset of 80M unlabeled TinyImagenet is been used for CIFAR distillation Ba & Caruana (2014). However, we demonstrate the result on 120k TinyImagenet which contain labeled examples belonging to 200 classes. We observe that there is no significant overlap between the classes of CIFAR and 120k TinyImagenet.

3 Experimental results

3.1 Performance of mismatched stimulus and effect of its complexity

3.1.1 MNIST

Fig. 1(a) shows the performance of mismatched stimulus on the student network which has same architecture as the teacher. The teacher accuracy on the test is 99.1% (90 errors). The best accuracy obtained with CIFAR, STL and random stimulus is 98.1% (190 test errors), 97.7% (228 test errors) and 84.5% respectively. The result seems interesting because, though MNIST is a digit dataset, a mismatched object dataset CIFAR works very well as the stimulus. Fig. 1(b) shows the test accuracy performance for a smaller student network.

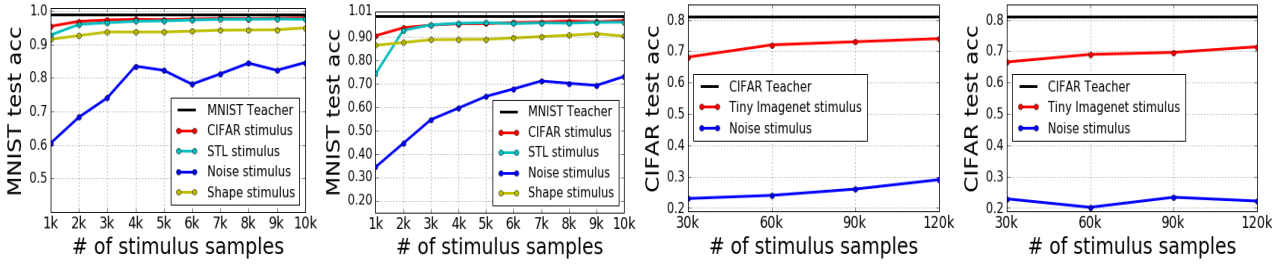
Shape dataset was previously used for demonstrating an effectiveness of Curriculum learning Bengio et al. (2009). The dataset consist of 10k examples of simple shape images. Due to small variability, the dataset is simpler than CIFAR or STL. From the plots in Fig. 1(a)(b), it can be seen that Shape stimulus performs inferior to CIFAR and STL.

For the sake of completeness, we also performed experiment with DNN teacher-student for MNIST. The details of the experiment are given in Appendix.

3.1.2 CIFAR

Fig. 1(c) shows the result when teacher and student are identical. Teacher accuracy is 81.1% while the best accuracy with 120k TinyImagenet stimulus was 74%. Fig. 1(d) shows the performance with a (10 times) smaller student CNN. The maximum accuracy here with TinyImagenet is 71.4%.

For CIFAR teacher, we additionally perform experiment with MNIST, Shape, SVHN, Texture datasets on the smaller student CNN. We use 5k samples from each dataset as the stimulus. Table 1 shows the result of the experiment. The datasets are of varied 'complexity' (variations). Visually, the order of complexity is MNIST $<$ Shape $<$ SVHN $<$ Texture $<$ Tiny Imagenet. A trend similar to



(a) (b) (c) (d)

Figure 1: Distillation result with MNIST and CIFAR teachers.(a) MNIST test accuracy where teacher and student architectures (CNNs) are identical, (b) MNIST test accuracy where student is smaller than the teacher (CNNs), (c) CIFAR test accuracy when student is same as teacher, (d) CIFAR test accuracy when student is 10 times smaller than teacher.

MNIST is observed where a more complex dataset performs better than the relatively less complex stimulus. We explored one of the quantification approach for complexity and results are reported in Appendix.

	Noise	MNIST	Shape	SVHN	Texture	TinyImagenet
CIFAR Test acc	0.125	0.161	0.228	0.304	0.371	0.429

Table 1: CIFAR test acc with 5k samples from different stimuli.

3.2 Unlabeled stimulus for data augmentation

Though we have shown results under the assumption of no training data, mismatched stimulus can also be effective for data augmentation. To validate this, we performed following experiments. For the MNIST teacher, we used 500 labeled samples from MNIST and (optionally) augmented it with 3k unlabeled samples from various stimuli. Results are given in the Table 2.

	No augmentation	Noise(3k)	CIFAR(3k)	Shape(3k)
MNIST Test acc	0.955	0.956	0.972	0.973

Table 2: MNIST teacher data augmentation results.

For CIFAR teacher, we used 5k labeled samples from CIFAR dataset and (optionally) augmented it with 5k samples from various stimuli. The Table 3 shows the results. It can be seen that, in both the cases data augmentation helps the student to generalize better.

	No augmentation	Noise	MNIST	Shape	SVHN	Texture	TinyImagenet
CIFAR Test acc	0.548	0.583	0.594	0.593	0.586	0.632	0.634

Table 3: CIFAR test acc with data augmentation.

4 Discussion and Conclusion

As mentioned in Bucilu et. al. Bucilu et al. (2006), though collecting synthetic stimulus is easy for images, it is crucial that the data should match the training data distribution. If the training data is not available, mismatched images surprisingly, turn out to be a good stimulus. Experimental results demonstrate that the complexity of the stimulus plays a major role and a more complex dataset provides better performance. We explored a quantification approach for dataset complexity. When a small training set is available, an unlabeled stimulus is also effective for data augmentation .

REFERENCES

Tiny imagenet 120k: http://cs231n.stanford.edu/tiny-imagenet-100-a.zip, http://cs231n.stanford.edu/tiny-imagenet-100-b.zip.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pp. 2654–2662, 2014.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48. ACM, 2009.

Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541. ACM, 2006.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Adam Coates, Honglak Lee, and Andrew Y Ng. An analysis of single-layer networks in unsupervised feature learning. *Ann Arbor*, 1001(48109):2, 2010.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 5, 2011.

George Papamakarios. Distilling model knowledge. *arXiv preprint arXiv:1510.02437*, 2015.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328, 2014.

5 APPENDIX

5.1 DETAILS OF TEACHER AND STUDENT ARCHITECTURES

5.1.1 MNIST

The architecture of the teacher network is [Conv1(32,5,5)-MaxPool(2)-Conv2(64,5,5)-MaxPool(2)-FC(128)-Softmax(10)]. The architecture of smaller student network is [Conv1(16,5,5)-Conv2(16,5,5)-Conv3(16,5,5)-MaxPool(2)-Conv4(16,5,5)-MaxPool(2)-FC(128)-Softmax(10)].

5.1.2 CIFAR

The architecture of the teacher network is [Conv1(32,3,3)- Conv2(32,3,3)-MaxPool(2)-Conv3(64,3,3)- Conv4(64,3,3)- MaxPool(2)-Conv5(128,3,3)- Conv6(128,3,3)-MaxPool(2)-FC(1024)-FC(512)-Softmax(10)]. The architecture of smaller student network is [Conv1(32,5,5)-Conv2(32,5,5)-MaxPool(2)-Conv3(32,5,5)- Conv4(32,5,5)- MaxPool(2)-Conv5(32,5,5)-Conv6(32,5,5)-MaxPool(2)- Conv7(32,3,3)- FC(1000)-Softmax(10)].

5.2 RESULT OF DNN TEACHER-STUDENT FOR MNIST TEACHER

We experimented with DNN teacher-DNN student scenario similar to Papamakarios (2015). We train a single DNN teacher on MNIST data [784 - Dense(500) - Dense(300) - Softmax(10)]. The student is a DNN with only 10% hidden nodes as compared to teacher. We perform distillation using CIFAR, STL and a normal Gaussian noise stimulus. Fig. 2 shows the result of the experiment.

Note that, though natural image stimulus works better than noise, the difference in the performance is not significant in case of DNN. It is known that the initial layers of CNN learn generic features such as edges, blobs Yosinski et al. (2014). Since such features are easily found in natural images as compared to noise images, natural images turn out to be a be a better stimulus than noise. However, since DNN employs full connection between hidden layers, its hidden nodes get activated even with random noise and hence noise may be performing well for DNN teacher as reported in Papamakarios (2015).

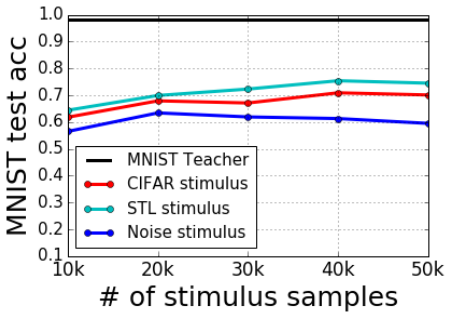
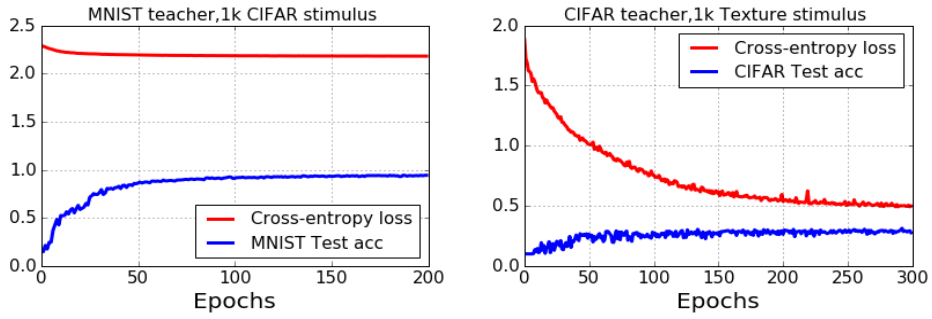


Figure 2: DNN teacher-student for MNIST.

5.3 PLOT OF CROSS-ENTROPY LOSS AND TEST ACCURACY

Our objective function is to minimize the cross entropy loss between soft targets of the teacher and the student on the unlabeled stimulus. We terminate the iterations when the cross entropy loss cease to change. To visualize possibility of overfitting (if any), we plotted the cross entropy loss and the test accuracy for two cases: MNIST teacher using 1k CIFAR stimulus and CIFAR teacher using 1k Texture stimulus. The plots is shown below.



(a) (b)

Figure 3: Plot of cross entropy loss and test accuracy for each epoch.

Note that, the test accuracy and the cross entropy loss settles down with more iterations. Even with small training size (1k), overfitting is not observed. This could be because of soft labels used in the optimization.

5.4 QUANTIFICATION OF COMPLEXITY

We explored one of the quantification approach for complexity. We suspect that a stimulus dataset which matches the convolution filters as well as have more variations, work better for distillation. To validate this, we performed an experiment with CIFAR teacher and various 5k stimulus datasets. For the first convolution layer feature maps, we calculate a mean across the dataset. We also calculate an average of feature map-wise std. dev. for the dataset. A higher mean value indicates that convolution filters has better match with the dataset while high value of std.dev. indicates more variations.

	SVHN	Shapes	Noise	MNIST	TinyImagenet	Texture
Mean	0.117	0.134	0.139	0.148	0.165	0.174
Std Dev	0.06	0.08	0.02	0.01	0.09	0.12

Table 4: Mean and avg standard deviation for first layer filter map of CIFAR-teacher for various datasets.

We observe that a dataset with high value of mean as well as std. dev. works better for distillation.