An Empirical Analysis of the Impact of Data Augmentation on Distillation

Deepan Das [*1] **Haley Massa** [*1] **Abhimanyu Kulkarni** [*1] **Theodoros Rekatsinas** [*2]

Abstract

Generalization Performance of Deep Learning models trained using Empirical Risk Minimization can be improved significantly by using Data Augmentation strategies such as simple transformations, or using Mixed Samples. We attempt to empirically analyze the impact of such strategies on the transfer of generalization between teacher and student models in a distillation setup. We observe that if a teacher is trained using any of the mixed sample augmentation strategies, such as MixUp or CutMix, the student model distilled from it is impaired in its generalization capabilities. We hypothesize that such strategies limit a model's capability to learn example-specific features, leading to a loss in quality of the supervision signal during distillation. We present a novel Class-Discrimination metric to quantitatively measure this dichotomy in performance and link it to the discriminative capacity induced by the different strategies on a network's latent space.

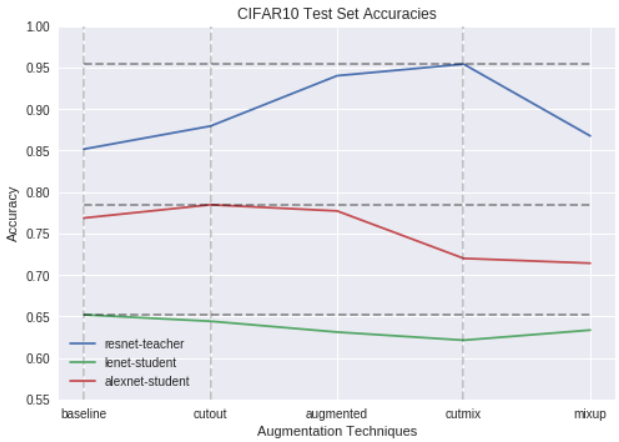

Figure 1. Impact of the different augmentation strategies on teacher and student models tested on CIFAR-10. Note that Mixed Sample Augmentation strategies help improve teacher performance, but corresponding student performance is impaired.

1. Introduction

A deeper analysis of implicit regularization techniques has shown that as neural networks increase in size, they are actually able to create solutions with lower complexity (Neyshabur, 2017), corresponding to better generalization performance. However, the size of such complex neural networks proves to be a hindrance when being deployed on more reasonable hardware. Several techniques such as model quantization (Zhou et al., 2017), model pruning (Han et al., 2015), and more recently, lottery tickets (Jonathan Frankle, 2018) enable the extraction of a subnetwork from the original network that approximates the performance of the original model. However, knowledge based distillation can also be used to separately train a smaller, lightweight model, possibly without even using any external annotations on given data. In a distillation setup, where a smaller parameter space(student) attempts to mimic the softened softmax output of a larger space(teacher), student performance improves considerably, when compared to standalone training, but there still exists a significant generalization gap between the two models. Besides trying to minimize this generalization gap, one should also consider the viability of using one or many of the implicit regularizers available during the cumbersome training process. One such regularizer draws inspiration from the varying degree of noise in training data in natural settings, and is often referred to as Vicinal Risk Minimization. Contemporary augmentation strategies such as Random Flip, Random Rotate, Cut Out (DeVries & Taylor, 2017), etc. are widely popular, but the scope of a vicinity is not explicitly defined in such strategies. Mixed Sample Data Augmentation Techniques like Mix Up (Zhang et al., 2017) and Cut Mix (Yun et al., 2019) or FMix (Harris et al., 2020) provide a new outlook on the concept of vicinity. Considered as standalone techniques, they can lead to state-of-the-art results in standard Deep Learning tasks, but it is interesting to note the impact each technique has on the supervision signal from

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of Wisconsin-Madison, USA [2]Department of Computer Science and Statistics, University of Wisconsin-Madison, USA. Correspondence to: Deepan Das <ddas27@wisc.edu>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

the teacher model to the student model. We hypothesize that even though Data Augmentation techniques provide good regularization, they impair the distillation process because of several implicit qualitative biases in the techniques. This impairment is much more pronounced in the Mixed Sample Data Augmentation Techniques. Our contributions are thus summarized as follows:

- We demonstrate that popular data augmentation techniques, and especially Mixed Sample techniques, such as MixUp and CutMix when applied on a teacher model, can impair the transfer of generalization capabilities onto a student model in a distillation setting.
- We present a novel similarity-based metric to help explain some qualitative traits inherent in the latent representations of such models. These findings are also backed by a KL-Divergence based Similarity metric, presented in the Appendix.
- We analyze the adversarial impact of Mixed Sample Augmentation strategies on the distillation objective when presented with data under distributional shift.
- We present empirical proof that data augmentation techniques tend to increasingly make models more discriminative and regularizes on example-specific features pertinent to the image.

2. Experimental Setup

Comparison Methods: Using the standard Empirical Risk Minimization principle, the loss objective is optimized only on the training samples, whereas in Vicinal Risk Minimization, virtual data points, and possibly labels, are also sampled from the vicinity of the real data points. It is hard to replicate an estimate of density around available data points, but the augmentation strategies discussed below can be thought of as extended VRM techniques as they provide a natural improvement over the existing empirical distribution. Moreover, a lot of recent work attempts to understand the qualitative abilities of such techniques (He et al., 2019; Gontijo-Lopes et al., 2020). We consider standard transformations such as flipping, splitting, scaling, rotating, cropping. It is important to note that these augmentations are data set dependent and require domain expertise. In this work, we test random cropping and random flipping along the horizontal axis. Cut Out is a generalization technique inspired by Dropout (Srivastava et al., 2014). In addition to these strategies, we consider Mixed sample strategies such as Mixup(Zhang et al., 2017), where a convex combination of data samples and their labels are generated as follows.

$$\hat{x} = \lambda x_i + (1-\lambda)x_j$$
$$\hat{y} = \lambda y_i + (1-\lambda)y_j$$

Here, $\lambda \sim \beta(\alpha, \alpha)$ is the mixture percentage of each image. This creates a target value that is a mix of the two original target values. By using an implicit bias that linear interpolations of data should lead to predictions that are linearly interpolated in the target space, Mix Up enables generation of well-calibrated models whose generalization performance is slightly better (Thulasidasan et al., 2019). Capturing a similar idea, (Yun et al., 2019) introduced Cut Mix that blends the classes of images (Yuji Tokozume, 2018) like Mix Up, but uses a mask to drop and fill the two different images.

$$\hat{x} = \mathbf{M} \odot x_i + (1-\mathbf{M}) \odot x_j$$
$$\hat{y} = \lambda y_i + (1-\lambda)y_j$$

where $\mathbf{M}$ is a binary mask that contains the information of where to drop and fill the image, $\odot$ denotes element-wise multiplication and $\lambda \sim \beta(1, 1)$ is the combination ratio.

Knowledge Distillation: As mentioned previously, we attempt to understand the impact of different augmentation strategies in a simple distillation setup, where a smaller model is trained by forcing it to mimic the temperature-treated probability distributions of a larger, cumbersome model. One can carefully tune an α value to appropriately weigh the Cross-Entropy loss against the labeling, and the distillation loss against the teacher's predictions.

$$\mathcal{L}(\mathbf{x}, y) = (1-\alpha)\mathcal{L}_{CE}(f_S(\mathbf{x}), y) + \alpha\tau^2\mathcal{D}_{KL}(f_S^\tau(\mathbf{x}), f_T^\tau(\mathbf{x}))$$

Datasets: Measuring Generalization is a cumbersome and ill-defined task and it is important to analyse performance of the different candidate models on not only unseen test data, but even test data with some natural variations in it. In our setup we consider the **CIFAR-10** test set to measure the model's performance on unseen data lying within the seen distribution. We also use the **CINIC-10** dataset as an out-of-sample generalization test. This data set is collected by (Darlow et al., 2018) and contains both CIFAR-10 and ImageNet images in its test fold. However, we just use the 70,000 ImageNet images that have been bucketized into CIFAR-10 classes. Another interesting test set is the **CIFAR-10H** dataset(Peterson et al., 2019) , that contains the exact same images as the CIFAR-10 images, but additionally provides the original probability distributions based on the labelings provided by the human annotators for each image. We use this to measure the closeness of a model's predictions with human beliefs about a data sample. We also run a similar set of experiments on **MNIST** data.

Generalization Measures: We evaluate each model on the basis of standard confusion metrics, such as *Accuracy, F-1 Score, Precision and Recall*, however we observed consistent behavior across the different confusion metrics. Moreover, since we have access to the probability distributions for

Table 1. Performance of the different Teacher and Student Models on CIFAR-10 Test Set. The KLD Metric is the distance between Human labeled confidence scores and Model prediction probabilities. Expected Calibration Error(ECE) measures prediction quality

MODELS	TEACHER			LENET STUDENT			ALEXNET STUDENT		
	CIFAR-10		CINIC	CIFAR-10		CINIC	CIFAR-10		CINIC
STRATEGY	ACC.	KLD	ACC.	ACC.	KLD	ACC.	ACC.	KLD	ACC.
BASELINE	0.852	0.656	0.600	**0.652**	1.002	**0.457**	0.769	0.710	0.544
AUGMENT	0.940	0.466	0.687	0.631	0.951	0.439	0.777	0.735	0.549
CUTOUT	0.880	0.220	0.630	0.644	0.987	0.451	**0.785**	0.726	**0.555**
MIXUP	0.868	0.641	0.614	0.633	0.991	0.444	0.714	0.836	0.498
CUTMIX	**0.954**	0.524	**0.716**	0.621	0.987	0.439	0.720	0.776	0.499

Table 2. Performance on the MNIST Test Set.

MODELS	ALEXNET TEACHER		LENET STUDENT	
	ACCURACY	LOSS	ACCURACY	LOSS
BASELINE	0.9958	0.020	0.9868	**0.039**
MIXUP	0.9960	**0.010**	0.9886	0.074
CUTMIX	0.9938	0.104	0.9861	0.107

each CIFAR test sample, we can easily compare a model's softmax score vector with the human-labeled distribution using a Cross Entropy measure, which is nothing but the KL Divergence across the two distributions.

$$\mathcal{L}_{CE} = \sum_{i=1}^{N} \mathcal{D}_{\mathcal{KL}}(\mathbf{p}_f(x_i), \mathbf{p}_h(x_i))$$

We also propose a novel metric to explain the discriminative power of the different models we train. This operates on the penultimate layer embeddings generated by the model and can be thought of as a measure of how well-separated the embedding manifold is. It takes into account both the intra-class and the inter-class similarity and defines the discriminative power of the model as the difference between them. If intra-class similarity is high, class representations are cohesive and more compressed. If inter-class similarity is low, class representations are less adhesive and are far away from each other. An optimal classifier will tend to have high cohesion C and low adhesion A, and thus higher discriminative power D. To compute this metric, we first standardize the embeddings, and define the cohesion and adhesion metrics as inter- and intra-class similarities, respectively using a cosine similarity function S. We define intra-class cohesion C as the following, where we deal with embeddings from just a single class i.

$$C^{(i)} = \frac{1}{N_C(N_C-1)} \sum_{i=1}^{N_C-1} \sum_{j=i+1}^{N_C} S(d_i, d_j)$$

and, inter-class Adhesion as the following, which is computed across all possible pairs of classes i and j:

$$A^{(i,j)} = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} S(d_k, d_l)$$

Where, N_k represents the number of instances in Class k, and d_i represents the embedding vector corresponding to image I_i. Furthermore we define the class-discrimination score as:

$$D = \frac{1}{\sqrt{d}}[\frac{1}{K} \sum_{i=1}^{K} C^{(i)} - \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} A^{(i,j)}]$$

Where, K represents the total number of classes and d represents the dimensionality of the embeddings.

Experimental Details: The models were implemented in PyTorch, with different parameter values for the different augmentation techniques. Knowledge distillation was performed with a temperature value of 20 and $\alpha = 0.5$. When performing augmentation, baseline and standard data augmentation did not require explicit parameters. MixUp, and CutMix used $\beta = 1$ and $\alpha = 0.5$. Finally, Cut Mix used alpha and beta parameters of 0.5 and 1, respectively. These values were chosen based on original paper implementation of the techniques, and could be fine tuned in later work. Five ResNet18 models (Kaiming He, 2015) were trained using the data augmentation techniques discussed above, on the CIFAR-10 training dataset. Two sets of student networks were built from the LeNet(Y. LeCun & Haffner, 1998) and AlexNet(Krizhevsky et al., 2012) architectures, with no data augmentation techniques added during the distillation setup.

3. Results and Discussion

As can be seen in Table 1, there is an impairment of generalization for student models derived from teachers trained using Mixed Sample Augmentation. This is consistent across

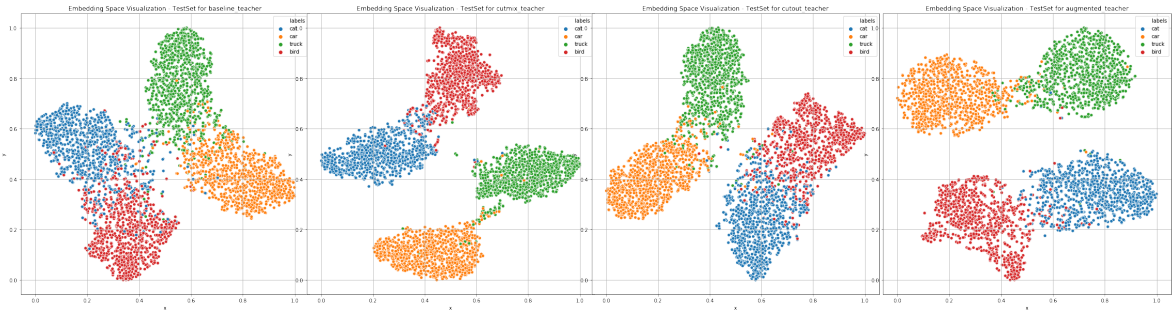

Figure 2. Latent Space Visualization from Penultimate Layer, dimensionally reduced using tSNE. From L-R: Baseline, CutMix, CutOut, Augmented. Notice the well-separated class representations in the augmented models as compared to the baseline model

Table 3. Class Discrimination Evaluation for Teacher Models

METRICS	COHESION	ADHESION	DISCRIMINATION
BASELINE	**0.739**	**-0.041**	**0.246**
AUGMENT	0.886	-0.049	0.296
CUTOUT	0.783	-0.043	0.261
MIXUP	0.793	-0.042	0.255
CUTMIX	0.917	-0.050	0.306

both the sets of student networks. Moreover, out of distribution performance of such models is also much worse than the baseline models and standard augmentation models. This trend holds in the MNIST dataset as well as seen in Table 2. This reversal of behavior is quite counter-intuitive, as we expect better generalized teacher models to transfer their capabilities to their students in a corresponding fashion. In an attempt to explain this behavior, we visualize the penultimate layer embeddings in Figure 2 for the different teacher models. We choose four classes that can be paired up using semantic similarity. We note that for CutMix, that uses a linear interpolation of both images and labels, the clusters are much more compressed and each semantic group lies much farther away from the other than any other strategy. Moreover, there is little interaction between the classes, as not a lot of points lie between the clusters. On the other hand, the baseline model manifold presents itself as much more uniform, wherein a lot of points lie on the boundaries of the class clusters and there is a gradual change in representational capability of the model between the classes. The two different semantic groups are closer in the baseline, when compared to any other augmentation strategy. This separability between classes and semantic groups exists in Cutout and transformation based Augmentations, but is not as pronounced as the interpolative Cut Mix.

To quantify the behavior observed in the latent space visualiztions, we present the novel Discrimination scores for the different teacher models in Table 3. We find that data augmentation, in general creates a more discriminant space(high cohesion, low adhesion). Intuitively, a model with loose, yet well-separated class representation space(low cohesion, low adhesion) will make a better teacher than one with tight representations. This is consistent with the ideas presented in (He et al., 2019) wherein data augmentation strategies are believed to regularizers that focus more on class-specific major features, and simultaneously regularizing nuanced, example specific features. We interpret this as a great attribute in a standalone performance perspective, but distillation performance depends largely on the amount of information encoded in all the latent features encoded by the teacher model. Thus, if the generated features within any given class have greater variance, they are able to encode more information about the class' relationship with other classes and are expected to generate more generalized probability distributions. This is a key factor in generating better quality students. This also helps explaining the superior performance of Cutout and Augmentation that add new information by retaining the same label and transforming the image. In addition to this, we analyze the quality of predictions by generating visual heatmaps corresponding to KL Divergence across the prediction and human labeled distribution, and create a KL-Divergence based class separability measure, presented in the Appendix.

Some recent works attempt to explain the relationship between implicit regularizers and knowledge distillation, such as (Müller et al., 2019), which models the adversarial impact of label smoothing on distillation, while (Arani et al., 2019) analyses the beneficial impact of using trial-to-trial variability during distillation. (Cho & Hariharan, 2019) analyses the impact of Early Stopping on Distillation, but none refer to data augmentation strategies. We believe that by conducting this empirical study, we have been able to set a precedent about training Neural Networks with a certain objective. The analysis can be made more holistic by including results using data from a different modality, or by considering student performance under adversarial attack. Nonetheless, this work provides a novel outlook on data augmentation strategies and sets an order of preference when a practitioner may want to use Distillation as a downstream task.

References

Arani, E., Sarfraz, F., and Zonooz, B. Improving generalization and robustness with noisy collaboration in knowledge distillation, 2019.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801, 2019.

Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. Cinic-10 is not imagenet or cifar-10, 2018.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout, 2017.

Gontijo-Lopes, R., Smullin, S. J., Cubuk, E. D., and Dyer, E. Affinity and diversity: Quantifying mechanisms of data augmentation, 2020.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks, 2017.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2015.

Harris, E., Marcu, A., Painter, M., Niranjan, M., Prügel-Bennett, A., and Hare, J. Understanding and enhancing mixed sample data augmentation, 2020.

He, Z., Xie, L., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. Data augmentation revisited: Rethinking the distribution gap between clean and augmented data, 2019.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015.

Jonathan Frankle, M. C. The lottery ticket hypothesis: Finding sparse, trainable neural networks, 2018.

Kaiming He, Xiangyu Zhang, S. R. J. S. Deep residual learning for image recognition, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Müller, R., Kornblith, S., and Hinton, G. When does label smoothing help?, 2019.

Neyshabur, B. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Peterson, J., Battleday, R., Griffiths, T., and Russakovsky, O. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. doi: 10.1109/iccv.2019.00971. URL http://dx.doi.org/10.1109/ICCV.2019.00971.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks, 2019.

Y. LeCun, L. Bottou, Y. B. and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, November 1998.

Yuji Tokozume, Yoshitaka Ushiku, T. H. Between-class learning for image classification, 2018.

Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., and Choe, J. Cutmix: Regularization strategy to train strong classifiers with localizable features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. doi: 10.1109/iccv.2019.00612. URL http://dx.doi.org/10.1109/ICCV.2019.00612.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization, 2017.

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., and Glynn, P. On the convergence of mirror descent beyond stochastic convex programming, 2017.

A. More Generalization Measures

A.0.1. Class Separability

As yet another way to quantify the behavior observed in the latent space visualiztions, we propose the use of a very basic *class-separability* metric that simply measures the difference in a model's confidence structure for different classes. In an ideal scenario where the model is supremely confident of each image, each class' confidence distribution is well separated and this can theoretically happen if a model is let to run for long enough using gradient descent algorithm. But, in real scenarios, models also assign probabilities to incorrect classes and as claimed by (Hinton et al., 2015), this relative probability structure is crucial in the generalization performance during distillation. The mathematical

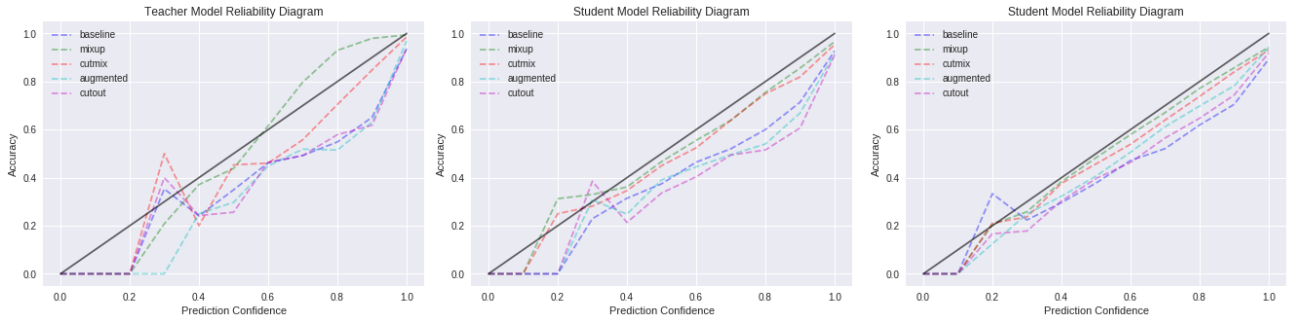

Figure 3. Reliability curves for different models. From L-R: Teacher Models, AlexNet Models, LeNet Models. It is observed that data augmentation generates calibrated predictions

formulation of this *class-separability* metric can be written as:

$$S_f = \frac{1}{C^2} \sum_{i=1}^{C} \sum_{j=1}^{C} \mathcal{D}_{\mathcal{KL}}(p_{f,i}, p_{f,j})$$

Where, $p_{f,k} \in \mathrm{R}^C$ represents the average model prediction distribution for the class k, and C represents the number of classes.

Table 4. Class Separability Score S_f

MODELS	TEACHER	LENET	ALEXNET
BASELINE	**3.73**	3.21	3.07
AUGMENT	5.19	3.14	3.23
CUTOUT	4.30	3.14	3.41
MIXUP	5.32	2.92	**2.17**
CUTMIX	5.33	**2.89**	2.27

As the class separability between probability distributions of different teacher models grows, one can imagine that it tends to generate well-knit tight clusters for each class in the data. This adversely impacts the generalization performance of the students distilled from such teachers.

Prediction Quality: To test the idea of generating high quality probability distributions, we make use of the CIFAR-10 Human labeled datasets, and measure the quality of the average model confidence distributions against the human confidence estimates. This can again be represented as a KL-Divergence between the probability distributions of a model's average prediction probabilities for a given class against the average human distribution. We present this information in a Confusion Matrix like visualization in 5. Each cell represents the KL-Divergence between the human estimate on that class and the model's estimate. The KL-Divergences are all scaled uniformly and are color coded,

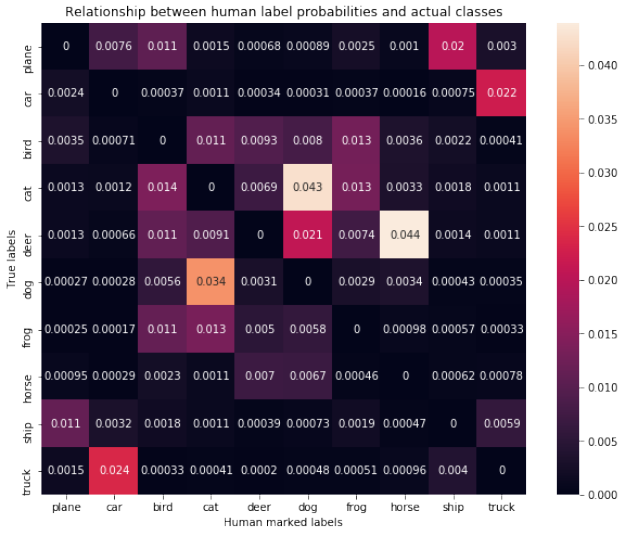

Figure 4. Average human confidence distribution across ground truth classes in CIFAR-10. Diagonal elements, that have the highest confidence have been masked to reveal implicit patterns between different classes

so higher values correspond to brighter pixel values. We are more interested in the non-diagonal elements as they reveal the mutual information between classes and note that the Cut Mix Matrix is much brighter in those pixels when compared to the Baseline. This points out the fact that the mutual information between different classes is better encoded in the Baseline than Cut Mix-trained model. This enables the creation of a superior representation manifold for distillation to take place.

Model Calibration: We also measure calibration performance across the several models, and note that interpolative techniques generate better calibrated models consistently across teacher and student models. This is evident from the reliability diagrams in 3 and Figure 6. However, it has

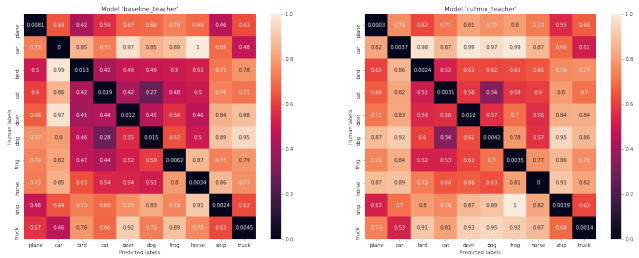

Figure 5. Relative KL-Divergence Confusion Matrices for Baseline Teacher(L) and CutMix Teacher(R)

been hypothesized in (Guo et al., 2017) that smaller models like LeNet generally tend to exhibit better Calibration performance than overparameterized, modern models like ResNet. This trend is visible as the reliability curves for all models tend to hug the ideal straight line closer as the model complexity decreases. However, no direct relationship can be found between the calibration performance of a student and the augmentation applied on its teacher model.

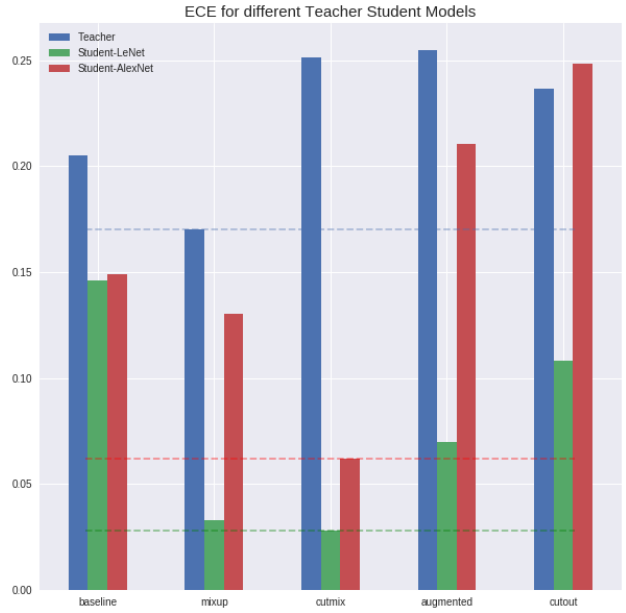

Figure 6. Expected Calibration Error for the different Models. Smaller models like LeNet usually have better calibration performance and previous studies also show that linear interpolation of data samples always leads to better model calibration

B. Augmenting the Students

Apart from the Accuracy, we also analyzed the F-1 Score, Precision and Recall metrics for all of the models trained. We found these values consistent with our findings, that augmentation techniques provide an adversarial impact on all stages of training. The lack of improvement from augmentation is shown through both AlexNet and LeNet students trained from augmented ResNet teachers. Additionally, we used a small ablation study with ResNet teachers and LeNet students to compare how augmentation effected each stage of the distillation process: no augmentation, augmented teacher with a baseline student, baseline teacher with an augmented student, and augmented teacher and student. The results are compiled in tables 5 and 6.

Model	Test Set	Accuracy	Precision	Recall	F1-Score	Baseline LeNet Accuracy	Baseline LeNet Precision	Baseline LeNet Recall	Baseline LeNet F1-Score	Baseline AlexNet Accuracy	Baseline AlexNet Precision	Baseline AlexNet Recall	Baseline AlexNet F1-Score
Baseline	CIFAR-10	0.852	0.851	0.852	0.851	0.652	0.648	0.652	0.650	0.768	0.767	0.769	0.768
	CINIC	0.600	0.602	0.600	0.593	0.456	0.452	0.457	0.449	0.544	0.550	0.544	0.529
Augmentation	CIFAR-10	0.940	0.940	0.940	0.940	0.631	0.632	0.631	0.474	0.777	0.776	0.777	0.766
	CINIC	0.630	0.620	0.628	0.629	0.439	0.434	0.435	0.435	0.549	0.557	0.549	0.535
Cut Out	CIFAR-10	0.879	0.878	0.880	0.879	0.644	0.643	0.644	0.643	0.785	0.783	0.785	0.784
	CINIC	0.629	0.620	0.628	0.630	0.451	0.447	0.451	0.445	0.554	0.562	0.554	0.540
Mix Up	CIFAR-10	0.867	0.954	0.867	0.953	0.633	0.631	0.633	0.632	0.714	0.712	0.714	0.712
	CINIC	0.603	0.606	0.603	0.598	0.444	0.435	0.443	0.433	0.497	0.501	0.498	0.477
Cut Mix	CIFAR-10	0.954	0.869	0.954	0.868	0.621	0.617	0.621	0.619	0.720	0.717	0.720	0.718
	CINIC	0.710	0.707	0.720	0.710	0.439	0.434	0.439	0.435	0.498	0.502	0.498	0.478

Table 5. ResNet Teacher metrics with different augmentation strategies, and LeNet and AlexNet Baseline Student Metrics distilled from corresponding teachers

Student Model	Test Set	Baseline Teacher Accuracy	Baseline Teacher Precision	Baseline Teacher Recall	Baseline Teacher F1-Score	Augmented Teacher Accuracy	Augmented Teacher Precision	Augmented Teacher Recall	Augmented Teacher F1-Score
Baseline	CIFAR-10	0.652	0.648	0.652	0.650	-	-	-	-
	CINIC	0.456	0.452	0.457	0.449	-	-	-	-
Augmentation	CIFAR-10	0.641	0.639	0.641	0.639	0.641	0.637	0.641	0.638
	CINIC	0.451	0.446	0.451	0.444	0.445	0.441	0.445	0.441
Cut Out	CIFAR-10	0.635	0.630	0.635	0.632	0.636	0.643	0.636	0.638
	CINIC	0.446	0.446	0.439	0.441	0.440	0.451	0.440	0.440
Mix Up	CIFAR-10	0.647	0.645	0.647	0.647	0.626	0.622	0.626	0.624
	CINIC	0.457	0.455	0.457	0.452	0.603	0.606	0.603	0.598
Cut Mix	CIFAR-10	0.629	0.626	0.629	0.627	0.635	0.631	0.635	0.633
	CINIC	0.439	0.433	0.439	0.434	0.446	0.443	0.446	0.441

Table 6. We mirrored the augmentation strategy in the teacher model to train the corresponding student models and noted no major difference in performance. LeNet Student Metrics trained with Augmentation from a) Baseline Teacher, and b) Teacher with same augmentation as student.