Report on:
Does Knowledge Distillation Really Work?

Emirhan Bilgiç   Aaron Weissberg   Amruth Srivathsan   Biruk Abere Ambaw

Abstract

Knowledge Distillation (KD) is a widely used technique for training a small neural network to copy the behavior of a larger network. The paper ”Does Knowledge Distillation Really Work?” (Stanton et al., 2021a), published at NeurIPS 2021, challenges the common understanding of how Knowledge Distillation functions. The authors show that a large gap in agreement between student and teacher predictions on unseen data can arise even if the student has the capacity to perfectly mimic the teacher. They present experimental evidence that optimization difficulties are a key reason for this discrepancy. Additionally, they investigate the role of dataset properties in distillation fidelity, revealing that increasing fidelity does not always correlate with improved generalization.

In this report, we start by summarizing the key insights of the paper, including the authors' exploration of optimization challenges, dataset augmentation effects, and the impact of ensemble teachers. We then discuss our attempt to replicate their experiments, analyzing both the successes and limitations of our reproduction. Our work aims to provide further empirical validation of the paper's claims, as well as insights into the nuances of Knowledge Distillation. The steps for setting up the required environment to run the experiments, the commands for each experiment, all models, and all results can be found in our GitHub repository: https://github.com/emirhanbilgic/knowledge_distillation_replication

1. Presentation of the Paper

The common understanding of KD is as follows: given a large, pre-trained teacher network and a smaller student network, KD aims to teach the student to effectively replicate the teacher's behavior, thereby transferring its learned representations.

The authors' main observation is that while KD does improve the student's generalization, it often fails to achieve high fidelity—i.e., strong agreement between the student and teacher's predictions.

1.1. Knowledge Distillation Transfers Knowledge Poorly

The authors begin with a simple experiment. They first train a LeNet-5 teacher network on 200 samples from MNIST, achieving a test accuracy between 84% and 85%. They then perform self-distillation—where the student and teacher share the same architecture—using the full MNIST dataset, along with additional distillation data from EMNIST.

As shown in Figure 2, this experiment follows the expected behavior under the common understanding of KD: fidelity increases as more distillation data is added, reaching up to 99% agreement. So far, nothing surprising.

Next the authors do an analogous self-distillation experiment with CIFAR-100 instead of MNIST, with synthetic images generated by an SN-GAN playing the role of EMNIST, as well as using a ResNet-56 as the network. As can be seen in Figure 4, now we see a large fidelity gap, even for the largest distillation set and the student outperforms the teacher with the generalization getting worse as we distill more. When we switch from self-distillation to the teacher being an ensemble of three ResNet-56, the latter effect disappears but the fidelity is still quite poor.

This illustrates the claim at the end of the last section, we get good generalization but with a large fidelity gap. The authors argue that even if KD improves generalization, closing this fidelity gap remains crucial for three reasons:

(1) Outside of self-distillation, increasing fidelity should help narrow the generalization gap as well. (2) High-fidelity distillation could enable the transfer of large black-box models into smaller, more interpretable, and reliable ones. (3) The existence of this fidelity gap highlights a fundamental issue in our understanding of KD itself.

They identify six potential reasons for the fidelity gap, which we will present with their given rebuttals.

1. **Student Capacity**: The experiments on self-

distillation already make this implausible. The authors additionally show experimentally that increasing the capacity does not help with the fidelity.

2. **Architecture**: Changing the architecture from ResNet to VGG produces similar results.
3. **Dataset Scale and Complexity**: They observe analogous results for the much bigger Imagenet dataset.
4. **Domain**: The authors also perform experiments on the IMDB dataset using LSTM, showing similar results in Natural Language Processing
5. **Identifiability**: This is referring to the idea that the issue is due to the distillation data.
6. **Optimization**: Lastly the issue might be that the KD optimization problem is not solved well enough.

The authors discuss the last two points in greater detail in the main text.

1.2. Identifiability: Are We Using the Right Distillation Dataset?

The authors test two hypotheses, the data distribution support hypothesis (they do not use this term specifically) and the Data Recycling Hypothesis.

The former can be stated in the following way: The fidelity gap is primarily caused by the support of the distillation data distribution being too small. To test this, they increase the support by applying extensive data augmentation to the distillation data from the CIFAR-100 experiment (section 1.1). As can be seen in Figure 6, the fidelity does increase somewhat, but the gap is still large and most importantly changing the temperature to $\tau = 4$ produced a larger agreement improvement than changing the support, which the authors argue rules out the hypothesis.

The Data Recycling Hypothesis states: The fidelity gap is primarily explained by us reusing the teachers' training data for distillation. The authors argue that recycling the data might cause problems since it violates the independency assumption of Empirical Risk Minimization. To test it, they split CIFAR-100 into D_0 which they use to train the teacher and D_1, distilling on D_0, D_1 and $D_0 \cup D_1$ (we call the students $s_.$ accordingly). Their experiments show s_0 getting a higher test accuracy than s_1 while having a worse fidelity. So the hypothesis seems to be formally right with respect to the fidelity but similar to the previous one, the increase in fidelity is too small to have a significant impact on the accuracy or to explain the gap.

With both hypotheses being rejected, the authors argue that Identifiability is ruled out as a key reason for the fidelity gap, only leaving optimization as a possible explanation.

1.3. Optimization: Does the Student Match the Teacher on Distillation Data?

The authors revisit the experiment from section 1.1. As we can see in figure 4 increasing the distillation data helped with the test agreement. At the same time figure 8 shows that for the agreement on the distillation set the opposite is the case, with data augmentation the effect gets even larger, going as low as 60%.

In conclusion, larger distillation datasets lead to a lower train agreement. However, since they showed that larger distillation datasets lead to better test agreement this suggests a trade-off between train and test agreement (in the following train agreement will always refer to agreement on the distillation data).

To better understand this, the authors look at a version of the CIFAR-100 experiment with a deterministic version (using LayerNorm instead of BatchNorm) of ResNet-20. This network cannot achieve 100% training agreement (with baseline augmentation) even when switching from SGD to Adam and increasing the number of epochs to 5000.

With this model in a self-distillation setting the optimal teacher weights are also optimal for the student. The authors show that initializing the student at a convex combination of teacher weights θ_t and random weights θ_r, namely $\theta = \lambda\theta_t + (1 - \lambda)\theta_r$ leads to the train agreement quickly dropping from 100% to about 80% when λ crosses 0.375. This explicitly demonstrates that the problem lies in the optimization.

The authors also experiment with the effect of choosing the same initialization for teacher and student, which overall seems to help fidelity only very slightly if at all.

2. General Discussion

In this section, we will make some general comments about the paper's strengths and weaknesses. We have put more emphasis on the latter as we see it as more insightful and because we can engage with points of criticism in our in-depth discussion in the replication section in some cases (This is not due to we think this is not a good paper).

2.1. Strengths of the Paper

The authors focus on fidelity in KD and show that it is lower than one might intuitively expect. We especially like the way in which possible explanations for this phenomenon are systematically ruled out through extensive experiments. The ultimate conclusion that optimization is the root cause, is insightful and may hopefully lead to improvements in KD in the future.

In general, this paper presents an extensive experimental study on KD which is interesting in and of itself. There are many smaller practical insights that can be taken from it. To name a few: Larger teacher ensembles are easier to emulate for students, data augmentation (in their experiments MixUp worked best) can help with fidelity and accuracy as can temperature tuning, ideally distilling on a mix of recycled data (easier to fit) and new in distribution data (helps identifiability) if one wants to balance accuracy and fidelity.

The paper is structured and written well. It also features an extensive appendix with a lot of additional details and experiments.

2.2. Points of Criticism

In this section we will discuss some of the paper's weaknesses. We have aggregated three general points of criticism that were both noticed by us and also raised by official reviewers.

1. **How to improve fidelity?**: The authors identify the optimization problem as not being solved sufficiently as the root cause of the fidelity gap, but they do not provide a way to remedy this issue. This means that the paper mostly does not have any direct big practical implications for KD.

2. **Why is fidelity important?**: Since the authors show that good fidelity does not necessarily imply good student accuracy (even outside of self-distillation), one is inclined to ask why one should care about the fidelity gap.

   We briefly mentioned the three reasons the authors gave for this focus in section 1.1 (1. High fidelity is the most obvious way to increase generalization, 2. Interpretability, 3. Understanding of KD). While these points are not invalid, they can also be criticized.

   Regarding the first point one might argue that if the ultimate goal is to increase generalization, it might make more sense to focus on finding ways to increase it directly since the authors show that the two are not always connected directly.

   The use of KD for interpretability is not a typical application. If the goal was to improve this use of KD, it might again make more sense to study it directly.

   For the final point that fidelity is key to understanding how KD actually works, we will refer to a point made by reviewer oZSf (which we will do again later). The authors seem to claim that the general understanding of KD in the ML community is that one really transfers ”dark knowledge” (like it is expressed in the original paper (Hinton et al., 2015)) from one model to another. Reviewer oZSf argues that the success of self-distillation already showed that this is too literal an interpretation of the name and that KD is instead connected to easier optimization, regularization, and self-supervised learning (Stanton et al., 2021b).

3. **Use of GAN generated data**: As discussed in section 1.1 the authors use GAN-generated images to extend the CIFAR-100 when doing distillation. This is meant to be more in distribution data like in the MNIST experiment they did before.

   We find it questionable if these synthetic images can really be considered in-distribution. It is generally known that augmenting a dataset with synthetic data is quite different than adding real data when it comes to training a model.

   This has a direct effect on the interpretation of the experimental results, which we will discuss in section 3.3.

3. Replication of Experiments

3.1. General Difficulties with the Replication

3.1.1. Environment Setup

Using the Paper's Official GitHub repository (https://github.com/samuelstanton/gnosis) was challenging. The absence of a Docker container required approximately 1.5 days to set up the environment and execute their commands. More critically, the number of commands provided was insufficient. The repository only included example commands, necessitating a thorough investigation of the codebase to recreate the experiments. As mentioned in the Abstract, to assist future researchers, we created a comprehensive GitHub repository (https://github.com/emirhanbilgic/knowledge_distillation_replication) containing all necessary commands for reproducing the experiments described in this report.

3.1.2. Reproducibility Issues

Certain experiments were impossible to replicate using their repository. For instance, Figure 6(c) in the actual paper, could not be recreated using the provided repository. A user raised this issue on GitHub (https://github.com/samuelstanton/gnosis/issues/18), but the authors did not respond.

Additionally, seeds were not provided by the authors, which caused inconsistencies. For example, seed 3 failed entirely for EMNIST-350K, throwing an error, while seed 4 worked.

3.1.3. GAN-Related Problems

The GAN experiment did not work with `synth_ratio=1`, as it threw an assertion error despite the provided GitHub explanation suggesting otherwise. We attempted using `synth_ratio=0.9999`, but this led to batch size issues. Ultimately, we set `synth_ratio=0.8`. This adjustment was based on a misexplanation in their GitHub documentation, which stated that a ”1:4 synthetic to real ratio” could be achieved with `synth_ratio=0.2` (actually corresponding to a 1:5 ratio).

To use 50k GAN samples, we calculated that `synth_ratio=0.8` (80% of 60k CIFAR100 data) approximated 50k samples. However, for earlier steps, we unintentionally used slightly more data due to unclear documentation. For example:

- With `synth_ratio=0.25`, we expected to use 12.5k additional GAN samples but actually used 15k ($0.25 \times 60k = 15k$).
- Despite these discrepancies, we increased the number of examples at each step which is crucial for meaningful results.

Another issue arose when training the GAN model: the generator checkpoint name needed to be `generator_100000.ckpt` (indicating 100k steps). However, since we could only train for 3500 steps (15 epochs), our checkpoint was named `generator_3500.ckpt`. We had to manually rename it to `generator_100000.ckpt`.

3.1.4. Optimization and Normalization Limitations

The code did not support Adam as an optimizer directly. To use Adam, we had to disable `trainer.optimizer.momentum` and `trainer.optimizer.nesterov`. Additionally, there were no options for layer normalization or batch normalization in the codebase, which posed challenges in replicating Figure 12.

3.1.5. Dataset Management

Organizing datasets into correct folders for replication of Figure 2 proved difficult due to issues downloading MNIST/EMNIST datasets. To address this, we created a separate Google Docs guide accessible at: https://docs.google.com/document/d/13QAJhrD0JMBOcTkTE_uzNtptfxvKLOSEjmIWhiD2VBM/edit?tab=t.

3.1.6. Computational Resources

Another main difficulty we encountered was limited computational resources. While the original paper does not say anything about the amount of used computing, the supplementary material https://proceedings.neurips.cc/paper/2021/file/376c6b9ff3bedbbea56751a84fffc10c-Supplemental.pdf containing the NeurIPS checklist responses give a rough estimate of 1000 GPU hours, which is obviously way out of our available resources.

Additionally, we did not have access to GPUs outside of the limited use provided by the Google Colab free tier. Google Colab free tier supplies NVIDIA Tesla T4 GPUs, which have 16 GB GDDR6. The GPUs used by the authors include NVIDIA Tesla K80, which has 24 GB GDDR5, and NVIDIA Titan RTX, which has 24 GB GDDR656.

With these constraints in place, we focused on a few central experiments and reduced the number of epochs, as we would not be able to conduct them otherwise. Subsequently, our focus is not to replicate the precise numerical values given by the authors, but more to see if we can observe the same behavior qualitatively.

3.2. Knowledge Distillation on MNIST

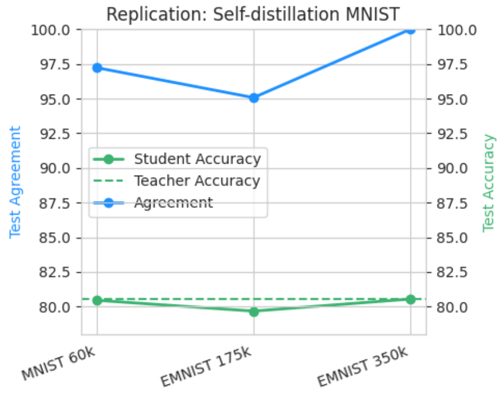


Figure 1. Our experiment results for MNIST KD.

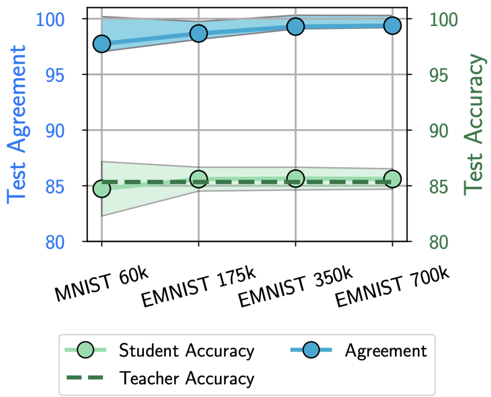


Figure 2. Actual paper’s experiment results for MNIST KD.

We started with the experiments they did on MNIST. They trained a LeNet-5 teacher on 200 random samples from MNIST, after which they performed self-distillation on MNIST and different subsets of EMNIST.

With this experiment, we quickly ran into some problems due to gaps in their code as well as in the implementation description given in the paper.

The different datasets they use come with default settings for their preprocessing. Specifically for the parameters mean, standard deviation, minimum value, and maximum value used for normalizing the data, 3 values are given by default, corresponding to the three color channels. Since MNIST is in grayscale this caused an error, since the default here also gave three values for each parameter. To work around this, we explicitly set the variables in our command.

Another issue was that the authors only provided detailed instructions for the hyperparameters when training the ResNets, not the LeNet. In the main text, the authors state explicitly that the LeNet-5 teacher network was trained for 100 epochs. In the appendix in A.3, where the authors describe the hyperparameters used for training the teachers, it is stated that 200 epochs. They also only mention there that they trained ResNets this way, not LeNet-5. Overall these inconsistencies seem to suggest, that the authors did not use these hyperparameter settings for training the LeNet-5 (at least not fully).

The command for the MNIST 60k setting and EMNIST 175k setting gave plausible results, but when we extended it to EMNIST 350k and EMNIST 700k, the command working for MNIST 60k and EMNIST 175k failed completely, resulting in a test accuracy of around 10%. When then cut the Figure 1 at EMNIST 350k, and plotted the train accuracy from EMNIST 350k. This failure of the command at 350k and 700k settings is concerning, though likely because of the lack of experiment parameters.

3.3. Knowledge Distillation on CIFAR-100

Now we move to the analogous experiment with CIFAR-100. Although we now have a complete description of the hyperparameters used during training, there are still some gaps in the documentation. Specifically, it is not stated for how many epochs the GAN they use to generate the synthetic images is trained, but only the number of gradient steps is stated. Since we would likely not be able to match their training time anyway with our limited resources, this is not a practical issue.

Training the teacher and performing distillation afterward is an effective approach for time efficiency. We also generally followed this method, except for the optimization experiment.

For GAN training, we set the `checkpoint_period` to 1, `evaluation_period` to 100, and set the logger directories manually.

We adopted the GAN training parameters provided by the authors for synthetic image generation. Specifically used SN-GAN (Spectral Normalization GAN). We set the batch size to 128 and used Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.9$. For the data augmentation, random horizontal flips with $p = 0.5$ (for the discriminator) are used. The authors achieved an FID score of 74.2617 after training for 100,000 gradient steps. In our case, we trained the GAN for 3,500 steps and obtained an FID score of 389.91.

We train three teacher ResNet-56 networks on CIFAR-100 for the ensemble distillation, one of which we use for self-distillation. The training is done with the hyperparameters stated in the paper: SGD with momentum 0.9, batch size 256, weight decay of 0.0001, cosine annealing learning rate scheduling ($\eta_{min} = 0$ and $\eta_{max} = 0.1$)and random horizontal flips ($p = 0.5$) and cropping (`padding_width` = 4). These settings are used for all following teachers we train unless we specify differences. We reduce the number of epochs from 200 to 50 to reduce computation time.

The three teacher networks achieved test accuracies of 63.2%, 63.58% and 63.59%. That they are so similar might in part be due to us using a smaller number of epochs. Networks tend to be more similar during early training dynamics (Calvo-Ordoñez et al., 2025). In consequence, our ensemble does not differ much from the single networks which we will see in a moment.

We perform self-distillation and ensemble distillation using the given hyperparameters so: (`batch_size` = 128), SGD with momentum and weight decay like with the teacher training, cosine annealing with $\eta_{min} = 10^{-6}$ and $\eta_{max} = 5 \cdot 10^{-2}$ and data augmentation also like before. These settings are used for all following students we distill, unless we specify differences. We again reduce the number of epochs from 300 to 20.

The results are displayed in figure 3. The analogous results from the original paper are in figure 4. Both agreement and accuracy are generally lower since we used shorter training times. Let us start by discussing the self-distillation results. Like in the paper, we see the agreement increase as we add more synthetic images to the distillation dataset, and unlike with MNIST the agreement does not get close to 100% (in accordance with the paper). We also see that the student accuracy is in some cases higher than the teachers, which is also observed in the paper and is specific to self-distillation.

The authors also state that they observe the student accuracy decrease because more distillation data is added. Looking at their results we do not find this conclusion necessarily follows. Notice how in their results the accuracy first increases, then decreases, and finally increases again, all on a rather small scale of about 1% and with only three trials. Reviewer oZSF argues that the decrease in accuracy should rather be explained by the GAN images being OOD and thereby making the teacher give noisy signals (Stanton et al., 2021b). The authors correctly note that if this was the case we should also be seeing a decrease for the ensemble distillation case, which we do not. Still, we think that the use of GAN might have exterior effects on student accuracy.

Looking at our results we see the student accuracy increase slowly (and later stay the same) as we add more GAN images. While this differs from their observation, it does not really contradict their interpretation, as our accuracy was not above the teachers' for the smallest distillation set (fewer epochs).

If we now compare our self-distillation results to the ensemble results, we see that they are rather similar. As we explained earlier our three teachers are quite similar resulting in an ensemble that does not outperform the individual teachers (reaching the performance of the best teacher). As such our ensemble distillation is effectively quite similar to self-distillation, which explains the similarity of the results.

We want to make some additional comments regarding the ensemble distillation results of the original paper. In Section 4.2 the authors state that their results show fidelity and student accuracy being positively correlated. While there is a slight increase from the first to the second point, the accuracy does not change noticeably afterward (note the estimated variance), so the positive trend is quite weak. Coming back to the comment of Reviewer oZSF, it does seem plausible that while having a larger distillation dataset did increase agreement overall, adding GAN images that are (weakly) OOD, caused the student to not improve at the CIFAR-100 image classification task much.

3.4. Identifiability Experiments

In this section, we discuss the experiments that were used to rule out the choice of distillation dataset as the central reason for the fidelity gap. As discussed in section 1.2, the authors perform experiments with data augmentation and with data recycling to this end. Unfortunately, the authors' code does not implement the experiments performed around the recycling hypothesis (we assume they used custom code not available on the repository), meaning that we will focus on the data augmentation part.

Due to computational constraints, we will not test all the possibilities, but restrict us to the baseline at temperature $T = 1$ (the baseline includes random horizontal flips and crops), the Baseline at temperature $T = 4$, adding rotation and adding MixUp. We chose the MixUp since it was the best-performing augmentation in their experiments (highest accuracy, very high agreement) and the rotation as a simple augmentation that did not perform particularly well in their results for comparison.

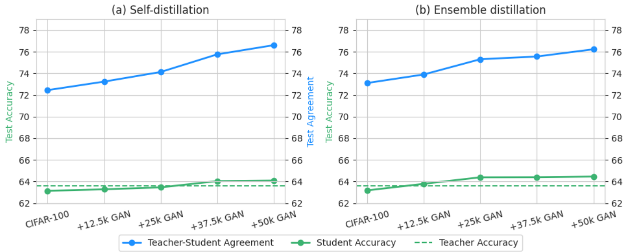
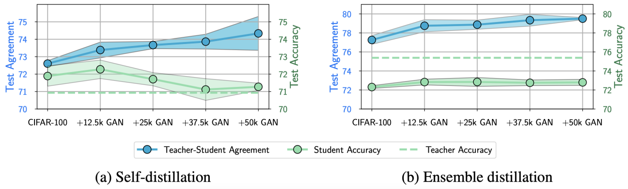


Figure 3. Our experiment results for CIFAR KD.



(a) Self-distillation

(b) Ensemble distillation

Figure 4. Actual paper's experiment results for CIFAR KD.

We also picked the higher temperature Baseline since it performed comparably to their best augmentations, suggesting that data augmentation is not key to closing the gap. Again to comply with our resource constraints we will use three instead of five networks as the teacher ensemble. Like previously we performed all experiments with 20 epochs.

We display our results and the authors for comparison in figures 5 and 6. Firstly, we note that both agreement and accuracy are lower overall for our results since we used a lower number of epochs. Having used a smaller ensemble might also be a reason for the smaller agreement as the authors showed that agreement increases with the teacher ensemble size. We also notice that both baseline versions performed better than the other two (except for MixUp having a slightly higher Agreement than $T = 1$ Baseline). This is not surprising as the student networks are likely underfitted due to the shorter training time, leading to better results with the ”smaller” training dataset (distillation set here specifically) without additional data augmentation.

Besides that, the relative differences mirror those in the paper. Rotation did noticeably worse than the other three, reaching an agreement of around 55% and an accuracy of 47.5%. Although the difference is more pronounced in our results, in the paper it also had the lowest agreement and second-lowest accuracy among all data augmentation schemes (besides Baseline). As in the paper, MixUp did

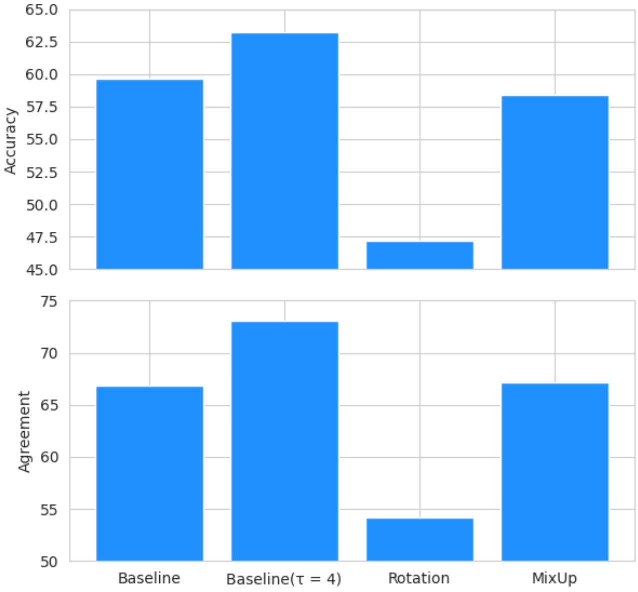


Figure 5. Our experiment results comparing different data augmentation techniques.

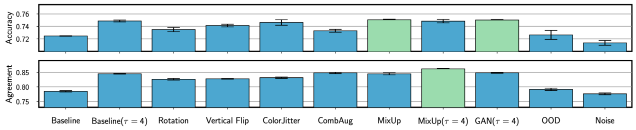


Figure 6. Actual paper's experiment results comparing different data augmentation techniques.

better than Rotation, achieving similar performance to the Baseline. Finally, we also observed the Baseline with temperature $T = 4$ doing better than the regular Baseline. The percentage differences are also similar with an increase of $\sim 5\%$ in agreement and $\sim 2.5\%$ in accuracy.

What ultimately rules out the identifiability hypothesis with respect to data augmentation in the paper, is that the $T = 4$ Baseline achieved a similar increase in agreement to applying data augmentation. Since our agreement is lower for both data augmentation schemes, likely due to the lower number of epochs, we can not fully replicate the observations necessary for this. Since we see a similar gain from the temperature tuning, we would need to see a much larger increase in the agreement of MixUp or Rotation to contradict the authors' argument, making it implausible.

3.5. Optimization Experiments

Next we look at the experiments that were used to deduce that issues with optimization are in fact the root cause of the fidelity gap. We discussed these in section 1.3.

3.6. Train Agreement

In this section, we cover the experiments on larger distillation datasets leading to smaller train agreement. The authors presented results on train agreement based on the initial CIFAR-100 experiments (section 3.3), the data augmentation experiments (section 1.2) and the data recycling experiments. As we did not replicate the last one we will only cover the first two.

We start with the first one. The train agreement results taken from the experiments we performed in section 3.3, can be seen in figure 7, the corresponding results from the paper are in 8. Again due to our shorter training time, we did not reach a full train agreement on CIFAR-100 like the original authors did, and the agreements are also lower overall.

Still, like in the original results we see a clear decrease in the train agreement as the distillation set is enlarged. Our decline is actually even stronger going from 74% to about 60%, compared to the authors' 100% to 95% (or 90% for the ensembles). This difference is likely again due to us having a shorter training time.

Curiously going from +37.5k GAN to +50k GAN did not decrease the train agreement and the decrease in the previous step was also already smaller. It is possible that the train agreement decrease slows down at some point, but we do not have sufficient evidence for or a good explanation for this hypothetical behavior.

Unlike in the actual results, the curves for ensemble and self distillation align with ours. This is consistent with our discussion in section 3.3, where we argued that our ensemble teacher does not differ much from the single teacher network effectively.

Next, we will discuss the train agreement for the data augmentation experiments we did in section 1.2. In figure 9 we display our results and in figure 10 the original results. We used the results for the data augmentation schemes we actually implemented, so Rotation and MixUp, the latter of which is not actually shown in the original results. We also show the $T = 4$ Baseline which is also not shown in the original results. Note again that the agreement values are lower due to the smaller amount of epochs. Also note that we only show the ensemble distillation results, since they did not differ much from the self-distillation ones (as we saw in the previous paragraph).

We can see that in both our and the original results there is a clear drop in the train agreement when we go from

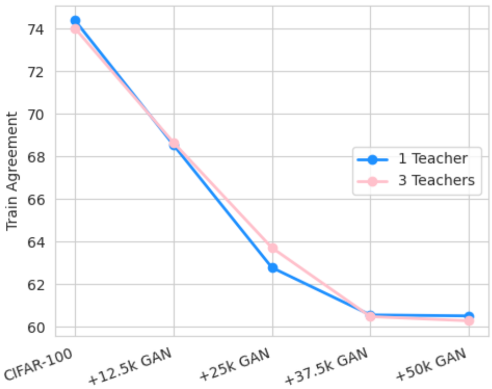


Figure 7. Our experiment results for the train agreement for teacher ensembles.

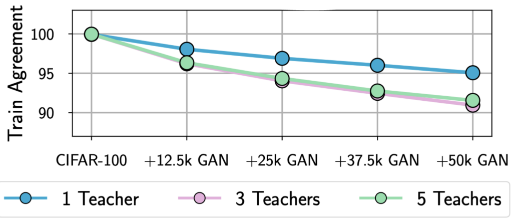


Figure 8. Actual paper's experiment results for the train agreement for teacher ensembles.

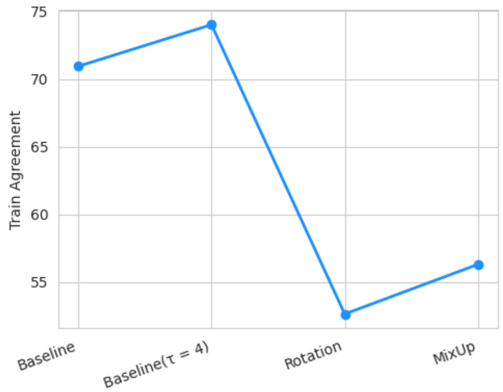


Figure 9. Our experiment results for train agreement with different data augmentation schemes.

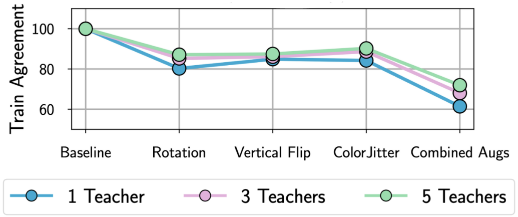


Figure 10. Actual paper's experiment results for train agreement with different data augmentation schemes.

the baseline to adding data augmentation. For Rotation the decrease is about 20% in both results. For MixUp the decrease is a bit smaller in our results, the authors did not provide the train agreement for it.

For the higher temperature baseline we actually see an increase in the train agreement (we again have no comparison to the paper). This is interesting as it emphasizes that heuristics like temperature tuning can be used to increase both train and test agreement (and test accuracy in our experiments), while increasing the distillation dataset with data augmentation introduces the trade-off we discussed in section 1.3.

3.7. Optimizer Effect

Finally, we cover the experiment on the optimizer effect, where the authors self-distilled a deterministic version of ResNet-20 on CIFAR-100 for up to 5000 epochs with SGD and Adam to see if a full train agreement is achievable. Due to our compute constraints we were only able to reach a maximum of 150 epochs. The SGD version is trained like before, for the Adam version we change nothing besides the optimizer.

Our results are in figure 11, the original results can be seen in figure 12. The authors observed the increase in train agreement slowing down with the number of epochs. With SGD an increase of about 3% from epoch 300 with agreement 79% took 700 epochs, while the next similar increase to about 83% took 4000 epochs. They argue that it is unlikely one would reach a full train agreement like this and even under the implausible assumption of the behavior continuing linearly, one would need tens of thousands of epochs, which is much longer than what is typically done.

We already see this slowing down of improvement at the lower epochs we replicated. We plotted the results at equal intervals of 50 epochs to make this more apparent.

We also see in both our and the original results that SGD with momentum did better than Adam for this task.

4. Conclusion

In this work, we aimed to replicate the experiments in the ”Does Knowledge Distillation Really Work” paper, and explored the challenges and insights surrounding KD focusing on fidelity and optimization issues. As discussed in Section 2, our findings mostly align with the original paper's conclusions, highlighting that optimization difficulties are a key factor contributing to the fidelity gap. While data augmentation and other techniques provide marginal improvements, they do not fully address the underlying challenges.

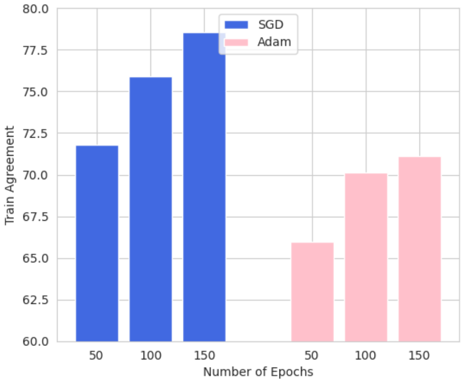


Figure 11. Our experiment results for different optimizers with different epochs.

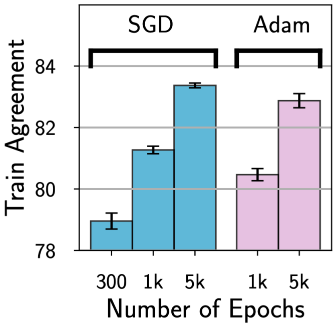


Figure 12. Actual paper's experiment results for different optimizers with different epochs.

The inadequacy of the provided GitHub repository and the lack of training details were the biggest challenges we faced, which, combined with our limited resources, led to partially different results than reported.

References

Calvo-Ordoñez, S., Plenk, J., Bergna, R., Cartea, A., Hernandez-Lobato, J. M., Palla, K., and Ciosek, K. Observation noise and initialization in wide neural networks. *arXiv preprint arXiv:2502.01556*, 2025.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., and Wilson, A. G. Does knowledge distillation really work? *Advances in neural information processing systems*, 34: 6906–6919, 2021a.

Stanton, S. D., Izmailov, P., Kirichenko, P., Alemi, A. A., and Wilson, A. G. Openreview: Does knowledge distillation really work? In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021b. URL https://openreview.net/forum?id=Oa9RlXNggGy.