Disagreement Augmentation for Distillation

Judah Goldfeder
Columbia University
New York City, USA
jag2396@columbia.edu

Yau-Meng Wong
Columbia University
New York City, USA
yw3809@columbia.edu

Benjamin Pace
Columbia University
New York City, USA
bfp2113@columbia.edu

Hod Lipson
Columbia University
New York City, USA
hod.lipson@columbia.edu

I. Introduction

Disagreement is a catalyst for learning. This principle, rooted in the Socratic method, highlights the role of productive conflict in refining ideas and uncovering deeper truths. Socrates, through his method of dialectical questioning, often encouraged his students to confront contradictions in their beliefs, leading to a richer understanding of complex concepts. This pedagogical approach, centered on optimizing the interplay between opposing perspectives, inspires a new direction in knowledge distillation. We propose that, much like in Socratic dialogue, fostering disagreement between the teacher and student models can drive learning and enhance model performance.

Knowledge distillation traditionally aims to minimize the divergence between a large, well-trained teacher model and a smaller student model, transferring the teacher's expertise to create a compact, deployable version of the original system [1]. This approach focuses on alignment, where the student learns to emulate the teacher's soft predictions, thereby inheriting its generalization capabilities. However, this paradigm overlooks the potential benefits of disagreement—particularly as a mechanism to explore underrepresented or ambiguous aspects of the data distribution [2]. By intentionally optimizing for areas where the student and teacher disagree, we aim to emulate the Socratic process, leveraging conflict as a driver of more robust learning.

In this work, we introduce a novel method of data augmentation rooted in disagreement. Our approach, Disagreement Augmentation (DA), augments training samples to maximize divergence between the student and teacher models. These disagreement-optimized examples challenge the student to reconcile conflicting predictions, encouraging it to develop a more nuanced approximation of the teacher. This method of structured disagreement offers a complementary perspective to traditional distillation methods.

We demonstrate the effectiveness of this approach across multiple benchmarks, showing that DA improves generalization and robustness. See the implementation on Github.

II. Literature Review

Knowledge distillation, introduced by Hinton et al [1], has traditionally focused on minimizing the divergence between a teacher model's outputs and a student model's predictions. While this approach has proven effective for model compression, recent work suggests that the standard paradigm for direct output matching may be suboptimal [2], [3]. The field has evolved to recognize that valuable information exists not just in the teacher's predictions, but in the learning process itself. The efficiency of knowledge transfer remains a central challenge, particularly with limited data or computational resources.

Recent studies have shown that the process of knowledge transfer can be enhanced by considering specific aspects of model behavior beyond simple output matching [8]. Work by Maroto et al. [5] demonstrated that knowledge distillation can enhance adversarial robustness, while Goldblum et al. [6] showed how adversarially robust teachers can produce more robust student networks. Of particular relevance to our work is the approach of using decision boundary information to improve distillation [7], which shares conceptual similarities with our augmented disagreement method. Recent work has explored using distillation specifically for adversarial defense. Methods like Adversarial Diffusion Distillation [11] and Adversarially Robust Distillation [6] show promising results in transferring robustness properties from teacher to student. These approaches suggest that careful consideration of decision boundaries during distillation can improve both robustness and efficiency.

Our work builds directly on the Committee Disagreement sampling approach introduced by Goldfeder et al [9]. While their work focused on exact parameter reconstruction, we adapt this technique for efficient knowledge distillation. Disagreement between multiple student models is shown to identify regions of the input space where knowledge transfer is most needed [9]. This relates to work in adversarial sample generation where the model disagreement often indicates decision boundary regions susceptible to adversarial attacks [10]. By combining insights from both disagreement-based learning and adversarial approaches, our method provides a novel framework for enhancing knowledge transfer through structured exploration of model differences.

III. Methodology

A. Experimental Setup

We conducted our experiments on the CIFAR-100 dataset [13], with three configurations of student/teacher pairs: Resnet8x4/Resnet32x4, VGG8/VGG13, and ShuffleNet-V2/Resnet32x4 [4] [8]. We used the original knowledge distillation method proposed by Hinton et al. [1], though our augmentation should be compatible with more modern techniques as well. The student model was trained to minimize

a weighted sum of the knowledge distillation loss and cross-entropy loss. Typical image augmentations were performed in both the baseline and DA experiments, such as random cropping and horizontal flipping. Experiments were run on either a NVIDIA RTX 4090 or distributed across 4 NVIDIA Tesla T4s.

B. Disagreement Augmentation Algorithm

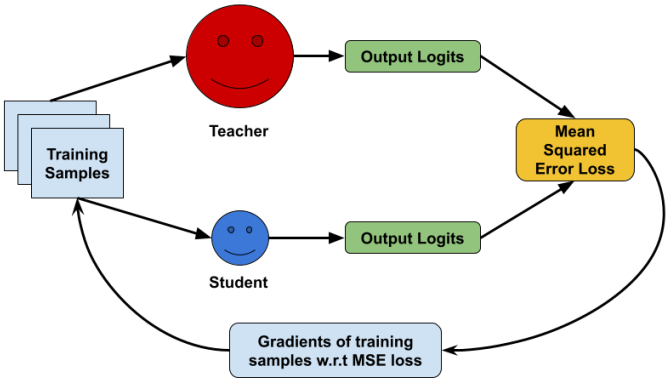

Fig. 1: Schematic of the recursive DA algorithm.

The Disagreement Augmentation algorithm is designed to optimize input data by emphasizing areas of disagreement between a teacher model and a student model. The process begins by freezing the weights of both the teacher (T) and student (S) models to ensure that the augmentation process only modifies the input batch (I).

For each iteration of augmentation, the input batch is forward-propagated through the teacher and student models to compute their respective output logits, denoted as L_T and L_S. These logits are then normalized to ensure they are on a comparable scale. The algorithm computes a disagreement loss l as the negative Mean Squared Error (MSE) between the normalized logits of the teacher and student: $l = -\text{MSE}(L_S, L_T)$. This loss function incentivizes maximizing the discrepancy between the models' predictions.

The disagreement loss is backpropagated to compute gradients with respect to the input batch I. These gradients are then used to update I directly, employing a fixed learning rate α. This process is repeated for a predefined number of epochs e, iteratively refining the input batch to amplify disagreement between the models.

Once the iterations are complete, the optimized input batch I is returned as the final output of the algorithm, and used to train the student in typical knowledge distillation fashion. This approach ensures that the augmented data emphasizes areas where the teacher and student models diverge, challenging the student model to learn more robust and generalizable features.

Algorithm 1 Disagreement Augmentation Algorithm

Require: Student S, teacher T, input batch I, learning rate α, epochs e
procedure DA(I, S, T, α, e)
  Freeze weights of S and T
  for each epoch i in 1 to e **do**
    Forward-propagate I through S and T
    Compute logits: $L_S = S(I)$, $L_T = T(I)$
    Normalize logits: $L_S \leftarrow \text{Normalize}(L_S)$, $L_T \leftarrow \text{Normalize}(L_T)$
    Compute disagreement loss: $l = -\text{MSE}(L_S, L_T)$
    Back-propagate l and compute gradient w.r.t. I
    Update I using α
  end for
  Return I
end procedure

IV. Experimental Results

A. Hyperparamater Search

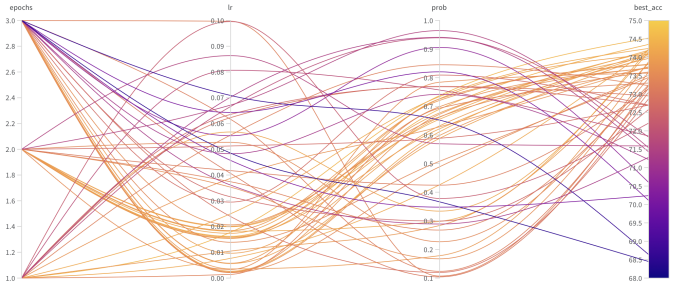

Fig. 2: Visualization of hyperparameter search over epochs e, learning rate α and probability of augmentation p. The column on the right is the student's best validation accuracy during training. Each line represents a training run.

To optimize DA for our task, we conducted a Bayesian hyperparameter search with Hyperband early stopping [15] over three hyperparameters: the number of epochs of augmentation per batch e, the learning rate of augmentation α, and the probability of augmentation per batch p. The search was conducted using a Resnet32x4 teacher and a Resnet8x4 student, with the goal of maximizing student validation accuracy. It found the ideal parameters to be $e = 1$, $\alpha = 0.01778$, and $p = 0.7374$. These are the parameters used in all subsequent experiments.

B. Validation Results

TABLE I: Validation accuracy of baseline student models and student models trained with DA.

Teacher	Student	Baseline (%)	DA (%)
Resnet32x4	Resnet8x4	73.66 ± 0.26	74.59 ± 0.24
VGG13	VGG8	73.33 ± 0.25	73.76 ± 0.29
Resnet32x4	ShuffleNet-V2	71.67 ± 0.34	73.70 ± 0.19

All training runs used an SGD optimizer [16], a batch size of 64, 240 training epochs, an initial learning rate of 0.05, and

learning rate decay at epochs 150, 180 and 210. The learning rate here refers to the normal student learning rate, not the DA learning rate α. Both DA experiments and baselines without DA were run 5 times each to ensure statistical reliability, with results reported as the mean and standard deviation across these runs.

C. Robustness to Disagreement Augmented Samples

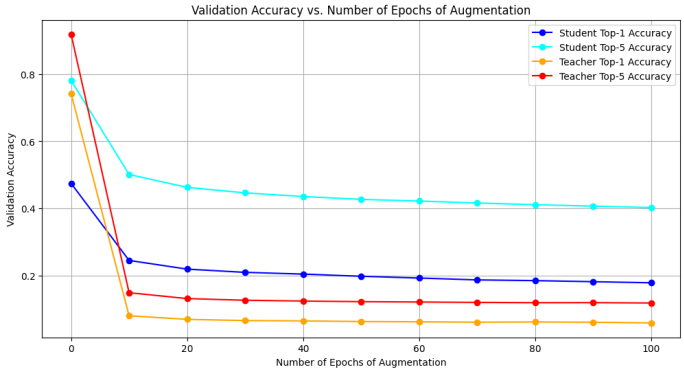

Fig. 3: Investigation into .

We hypothesized that training a student model with disagreement augmented samples would result in a more robust model. To investigate, we evaluated the validation accuracy of a pre-trained Resnet32x8 teacher and a DA-trained Resnet8x4 student under varying levels of augmentation intensity, measured by the number of augmentation epochs. Here, augmentation occurs on the validation set to ensure that the evaluation reflects whether training with disagreement-augmented samples leads to improved robustness against such perturbations. Additionally, we compared the performance of the student model with its teacher to assess whether the knowledge distillation process, combined with disagreement-based augmentation, enables the student to achieve similar or superior resilience in handling these adversarial-like inputs. This approach allowed us to validate the hypothesis that disagreement-driven training fosters a more adaptable and robust student model.

V. Discussion

A. Interpretation of Results

The results of our experiments demonstrate that incorporating DA into the knowledge distillation process significantly improves the generalization and robustness of student models. Across all tested configurations, models trained with DA consistently outperformed their baseline counterparts in terms of validation accuracy. This suggests that the structured introduction of disagreement during training helps the student model better learn nuanced representations of the teacher's decision boundaries.

Furthermore, the robustness evaluation confirmed our hypothesis that disagreement-driven training fosters resilience to adversarial-like inputs. By augmenting the validation set to contain disagreement-optimized samples, we observed that DA-trained students were better equipped to reconcile these challenging inputs, achieving performance levels comparable to or surpassing their teachers.

B. Comparison with Previous Studies

Our findings align with and expand upon prior work that has explored the role of adversarial robustness in knowledge distillation. Although earlier studies, such as Goldblum et al. [6], demonstrated the benefits of robust teachers for improving student resilience, our method extends this concept by actively incorporating disagreement between models as a training signal. Compared to approaches like adversarially robust distillation, DA introduces a more generalizable framework that does not rely on predefined attack methods but instead leverages natural divergences between teacher and student predictions. This positions DA as a complementary and scalable strategy for enhancing robustness in distillation tasks.

C. Challenges and Limitations

Despite its promising results, DA is not without challenges. One limitation is the additional computational cost incurred during the augmentation process, as optimizing input batches over multiple epochs introduces overhead. While this cost was manageable in our experiments with CIFAR-100, scaling to larger datasets or models may require further optimization of the augmentation procedure.

Another limitation is the reliance on hyperparameter tuning to achieve optimal performance. As shown in our hyperparameter search, the number of augmentation epochs (e), learning rate (α), and probability of augmentation (p) are critical to the success of DA. Automating or simplifying this tuning process could improve the accessibility of the method for broader applications.

D. Future Directions

Future work could address the computational challenges of DA by exploring methods to reduce augmentation overhead, such as adaptive augmentation strategies. Additionally, extending DA to other domains, such as natural language processing or reinforcement learning, could reveal its potential for tasks beyond image classification. Finally, investigating the theoretical underpinnings of disagreement as a learning signal, particularly in the context of decision boundary exploration, could further refine and justify the approach.

VI. Conclusion

A. Summary of Findings

This work introduced Disagreement Augmentation, a novel method for improving knowledge distillation through the intentional optimization of disagreement between teacher and student models. Inspired by the Socratic method, DA leverages structured conflict to challenge the student model, encouraging it to develop more robust and generalizable representations. Experimental results on CIFAR-100 demonstrated that DA-trained students consistently outperformed baseline

models, both in terms of validation accuracy and robustness to disagreement-augmented samples.

B. Contributions

Our primary contributions are as follows:

- The introduction of Disagreement Augmentation as a data augmentation strategy for knowledge distillation.
- Empirical validation of DA's effectiveness across multiple teacher-student configurations, demonstrating improved generalization and robustness.
- A conceptual shift in knowledge distillation, emphasizing the role of structured disagreement as a catalyst for learning.

C. Recommendations for Future Research

We recommend future research focus on scaling DA to larger datasets and more complex models, as well as exploring its applicability to domains beyond computer vision. Additionally, investigating hybrid approaches that combine DA with adversarial training or other robustness-enhancing techniques could further improve its efficacy. Lastly, theoretical studies on the dynamics of disagreement during distillation could provide valuable insights into the underlying mechanisms driving its success.

Disagreement Augmentation represents a promising step toward more effective and robust knowledge distillation, opening new possibilities for model compression and deployment in challenging real-world environments.

Acknowledgment

This work was built on this codebase [12].

References

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[2] X. Xu et al., "A survey on knowledge distillation of large language models," arXiv preprint arXiv:2402.13116, 2024.
[3] C.-Y. Hsieh et al., "Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes," arXiv preprint arXiv:2305.02301, 2023.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, 2015
[5] J. Maroto, G. Ortiz-Jim'enez, and P. Frossard, "On the benefits of knowledge distillation for adversarial robustness," arXiv preprint arXiv:2203.07159, 2022.
[6] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein, "Adversarially robust distillation," arXiv preprint arXiv:1905.09747, 2019.
[7] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge Distillation with Adversarial Samples Supporting Decision Boundary," arXiv preprint arXiv:1805.05532, 2018.
[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, 2015.
[9] J. Goldfeder, Q. Roets, G. Guo, J. Wright, and H. Lipson, "Sequencing the neurome: Towards scalable exact parameter reconstruction of black-box neural networks," arXiv preprint arXiv:2409.19138, 2024.
[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. Int. Conf. Learning Representations, 2018.
[11] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," arXiv preprint arXiv:2311.17042, 2023.
[12] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled Knowledge Distillation," arXiv preprint arXiv:2203.08679, 2022.
[13] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., Univ. of Toronto, 2009.
[14] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," arXiv preprint arXiv:1807.11164 , 2018.
[15] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization," arXiv preprint arXiv:1603.06560, 2016.
[16] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, 2017.

VII. Appendix

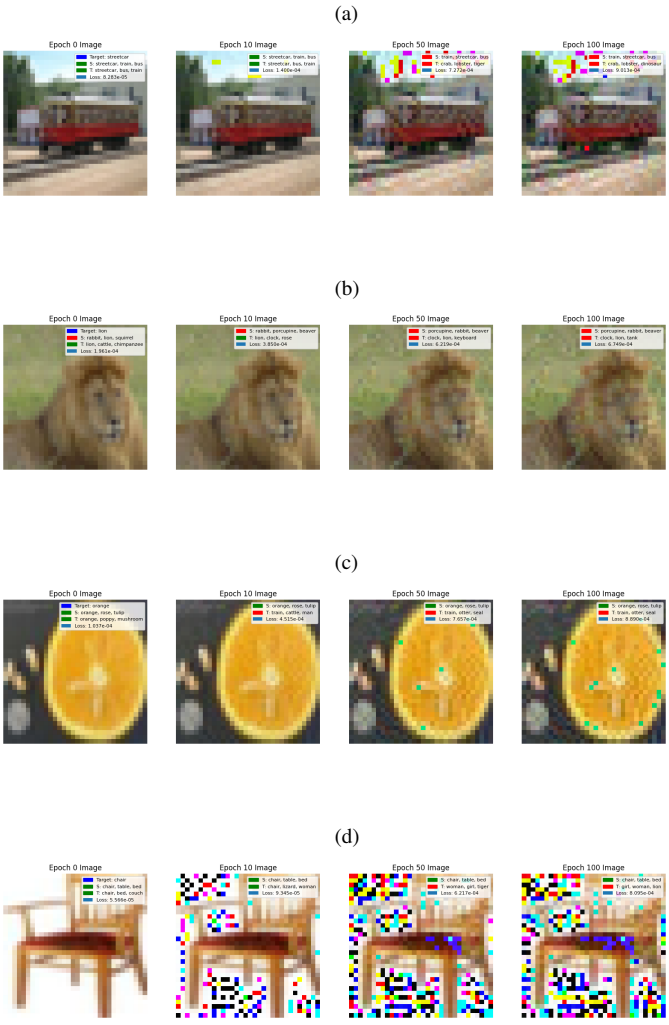

Fig. 4: Examples of CIFAR-100 images undergoing various epochs of DA. The legends shows the ground truth target label, the Resnet8x4 student model's top 3 prediction, the Resnet32x4 teacher model's top 3 predictions, and the MSE loss between the student and teacher logits.