Knowledge Distillation: The Functional Perspective

Israel Mason-Williams[1*] **Gabryel Mason-Williams**[2*] **Mark Sandler**[2]
[1]UKRI Safe and Trusted AI [2]Queen Mary University of London
israel.mason-williams@kcl.ac.uk {g.t.mason-williams,mark.sandler}@qmul.ac.uk

Abstract

Empirical findings of accuracy correlations between students and teachers in the knowledge distillation framework have served as supporting evidence for knowledge transfer. In this paper, we sought to explain and understand the knowledge transfer derived from knowledge distillation via functional similarity, hypothesising that knowledge distillation provides a functionally similar student to its teacher model. While we accept this hypothesis for two out of three architectures across a range of metrics for functional analysis against four controls, the results show that knowledge transfer is significant but it is less pronounced than expected for conditions that maximise opportunities for functional similarity. Furthermore, results from the use of Uniform and Gaussian Noise as teachers suggest that the knowledge-sharing aspects of knowledge distillation inadequately describe the accuracy benefits witnessed when using the knowledge distillation training setup itself. Moreover, in the first instance, we show that knowledge distillation is not a compression mechanism but primarily a data-dependent training regulariser with a small capacity to transfer knowledge in the best case.

1 Introduction

Although large deep neural networks can generalise well and learn strong representations in various tasks [1; 2; 3], this comes with a considerable computational cost. Thus, there is a desire to reduce the computational cost of these models whilst maintaining performance. Knowledge distillation has been introduced as a compression method to transfer the knowledge of a pre-trained model or ensemble of models into another model [4; 5], either a smaller model or a model with the same architecture via self distillation [6; 7]. Literature has sought to understand the dynamics of knowledge distillation. Some have questioned the role of knowledge transfer in the distillation process to improve student performance, suggesting that it is often not the result of transferring knowledge from the teacher to the student [8] and that in the extreme case, the teacher can be the randomly initialised version of the student [9]. We further this line of enquiry to rigorously test the functional similarity of trained models at the output layer to a teacher model against the standard self-distillation setup on CIFAR100. The four controls include independent models, Uniform Noise distillation, Gaussian Noise distillation, and standard distillation functionally compared to the teacher model used for standard distillation. Our results elucidate that two of the three architectures explored have significant functional similarity benefits of employing knowledge distillation. However, the functional similarity derived from knowledge distillation is not guaranteed and is indeed marginal, failing to explain performance gains witnessed as significant performance gains are observed for both the Uniform Noise distillation and Gaussian Noise distillation controls. Conclusively, our results show that knowledge distillation is best described as a data-dependent regulariser with marginal knowledge sharing capacity, meaning that it should not be a compression technique.

In this paper, we ask:

*Equal contribution of authors. Order decided by a fair coin flip.

Workshop on Scientific Methods for Understanding Deep Learning at Neural Information Processing Systems (NeurIPS 2024).

- Does knowledge distillation result in a significantly functionally similar model to the teacher?
- Are there computationally less expensive mechanisms for realising the benefits of knowledge distillation?
- Can knowledge distillation compress an architecture?

These questions are explored in the self-distillation regime where the student has the capacity to match the teacher's functional representation perfectly; the student is provided with the same initialisation as the teacher to maximise the potential for functional similarity further described in the Experimental Setup (Section 2) along with the Related Work (Appendix B).

Our contributions are:

- There is significant knowledge transfer in knowledge distillation, but it is inconsistent and marginal in conditions maximised for functional similarity.
- The benefits of knowledge distillation can be realised via computationally inexpensive Uniform or Gaussian Noise teachers.
- Knowledge distillation should be viewed as a ***data-dependant regulariser*** and be described as such as it is not a compression mechanism.

2 Experimental Setup

To explore the questions in the introduction, primarily whether knowledge distillation (KD) is a compression mechanism and can transfer knowledge from the functional perspective, we create a contrived and controlled environment to give the models the best chance to represent the teacher functionally.

We initialise the teacher, M_0, and then train it to convergence without any data argumentation to create the teacher M_T. Subsequently, all future models are trained using the same architecture and initialisation as the teacher M_0. This setup ensures that the models start in the same place in the loss landscape and can reach the same endpoint and function through training. Our setup attempts to isolate the effects of KD appropriately.

The independent models are then generated using the same initialisation as the teacher and training setup. The only difference in training between the teacher and independent models is the order of the training data.

The student knowledge distillation models are then initialisation with the teachers initialisation and trained with the teacher, M_T, using the conventional Knowledge Distillation setup (1),

$$\mathcal{L}(x; M_W) = (1-\alpha) * \mathcal{H}(y, \sigma(z_s; T=1)) + \alpha * \mathcal{H}(\sigma(z_t; T=t), \sigma(z_s, T=t)) \tag{1}$$

where x is the input, M_W is the student model parameters, α is the teacher weighting coefficient, $\mathcal{H}$ is the cross entropy loss function, y is the ground truth label, σ is the softmax function parameterised by the temperature T and z_s and z_t are the logits of the student and the teacher, respectively.

Two random control teachers are used to better disambiguate and understand the role of KD: Uniform Noise between 0 and 1 and Gaussian Noise with a mean of zero and standard deviation of 1. The Noise for the whole training setup is generated once to simulate a trained teacher model with deterministic output. The subsequent random student teachers are trained with the same setup as the standard KD, see (1). However, the teacher outputs are replaced with the appropriate Noise; see Appendix D Figure 3 for a diagrammatic representation.

To better understand the effect of KD and the role of the teacher, the following alpha values were explored: 0.1, 0.5 and 0.9. At 0.1, there is little dependence on the student updates based on the teacher's outputs; at 0.5, there is equal reliance on student and teacher outputs for the student updates; and at 0.9, the student updates are majorly reliant on the teacher's output. If the student-teacher dynamic facilitates knowledge transfer, we expect that, as this alpha value increases, the model will become significantly more functionally representative of the teacher due to the increased reliance on the teacher's output. For all KD setups, a temperature, T, value of 1 is used.

One teacher is trained using a seed of 0, and then 385 models are trained for each setup using seeds 1-385 to ensure the results are significant and representative; this resulted in training 3851 models for each architecture explored.

The functional similarity of the models to the teacher is measured against the last layer outputs, using Activation Distance, JS-Divergence, Predictive Disagreement and the Cosine Similarity of the output logits.

In the body of this paper, we explore the ResNet18 [10] and ViT [11] architecture on CIFAR100 [12]. The training details for the ResNet and ViT are detailed in Appendix E and G, respectively. The VGG-16-BN [13] architecture is explored with this dataset; these results are referenced in the main body but are presented within Appendix F as they do not deviate significantly from the results presented for the ViT.

3 Hypothesis Tests

We give the teacher and the student the same initialisation, meaning theoretically it can represent the same function, see Section 2. We test the hypothesis that KD is a form of model compression from the functional perspective with an expected threshold of 95% or above similarity with the teacher across all similarity metrics. If the 95% or above threshold is achieved, we can consider KD a form of model compression. Otherwise, it cannot.

We hypothesise that the KD does transfer knowledge to the student model. To test this, we state that, on average, KD results in models have a lower mean for (Activation Distance, JS Divergence and Prediction dissimilarity) and a higher mean for (Cosine Similarity) than independently trained or randomly trained KD models. Therefore,

H_0: KD models are, on average, not more similar to the teacher than the baseline models.

H_a: KD models are, on average, more similar to the teacher than baseline models.

The baseline models are the independent, Uniform and Gaussian Noise KD models. These hypotheses are tested with a one-tailed z-test using a significance of 0.025.

4 Results

The ResNet-18 architecture results presented in Table 1 and Figure 1, display the mean and standard deviation from 385 runs. The results show that the independently trained models bear the most functional resemblance to that of the trained teacher across the similarity metrics explored, namely Activation Distance, JS Divergence and Cosine Similarity - with Uniform Noise knowledge distillation with an alpha value of 0.1 presenting with the lowest prediction dissimilarity.

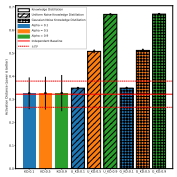
These results suggest that even when increasing the reliance on the teacher knowledge transfer via the alpha value in a traditional KD setup from 0.1 to 0.9 there is only a slight fluctuation in similarity to the teacher model. Not only does traditional KD make a less functionally similar model, but the model can be less functionally similar than a model trained with a Uniform Noise knowledge distillation setup.

The null hypothesis failed to be rejected for all functional similarity hypothesis tests on the ResNet-18. Interestingly, the Uniform and Gaussian Noise knowledge Distillation with an alpha of 0.1 achieves a lower mean than the independent baseline, suggesting that the KD process without "knowledge" can result in more similar models. Uniform Noise KD with alpha values 0.1 and 0.9 resulted in significantly more accurate models compared to the independent baseline, along with Gaussian Noise KD with alpha values 0.1. The results suggest that the KD process can lead to better performance without "knowledge" transfer, supporting the idea that KD is a data-dependant regularizer.

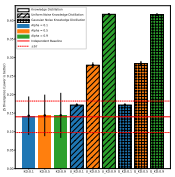
Metrics		Knowledge Distillation (KD)			Uniform Noise KD			Gaussian Noise KD		
	Independent	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Activation Distance (↓)	**0.3224±0.0284,**	0.3263±0.0344,	0.3269±0.0355,	0.3273±0.0392,	0.3487±0.0027,	0.5079±0.0043,	0.6682±0.0004,	0.3489±0.0028,	0.5121±0.0039,	0.6694±0.0014,
JS Divergence (↓)	**0.1400±0.0213,**	0.1429±0.0260,	0.1437±0.0280,	0.1439±0.0304,	0.1717±0.0023,	0.2794±0.0039,	0.4167±0.0006,	0.1718±0.0023,	0.2835±0.0037,	0.4155±0.0018,
Prediction Dissimilarity (↓)	0.3538±0.0210,	0.3568±0.0270,	0.3571±0.0286,	0.3580±0.0320,	**0.3437±0.0034,**	0.3780±0.0048,	0.3529±0.0029,	0.3438±0.0033,	0.3806±0.0060,	0.4434±0.0132,
Logits Cosine Similarity (↑)	**0.8090±0.0132,**	0.8072±0.0174,	0.8075±0.0193,	0.8078±0.0203,	0.7299±0.0026,	0.5889±0.0048,	0.5774±0.0016,	0.7295±0.0025,	0.5811±0.0054,	0.5363±0.0064
Test Accuracy (↑)	0.6269±0.0204,	0.6237±0.0264,	0.6232±0.0274,	0.6221±0.0306,	0.6463±0.0034,	0.6194±0.0053,	**0.6679±0.0033,**	0.6463±0.0034,	0.6168±0.0060,	0.5632±0.0143,
Test Loss (↓)	**1.5290±0.1047,**	1.5420±0.1310,	1.5402±0.1396,	1.5387±0.1520,	1.7775±0.0161,	2.4132±0.0236,	3.2113±0.0080,	1.7785±0.0164,	2.4454±0.0199,	3.2974±0.0290,

Table 1: ResNet18 Results for CIFAR100

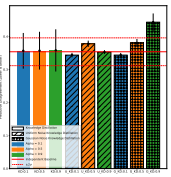
The ViT architecture results presented in Table 2 and Figure 2, display the mean and standard deviation from 385 runs. The results show that the knowledge distillation models bear the most



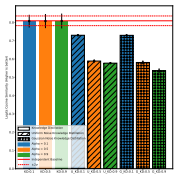
(a) Activation Distance (↓)



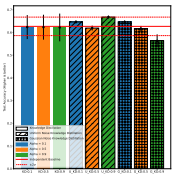
(b) JS Divergence (↓)



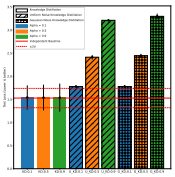
(c) Prediction Dissimilarity (↓)



(d) Logits Cosine Similarity (↑)



(e) Test Accuracy (↑)



(f) Test Loss (↓)

Figure 1: Functional similarity across each metric (a-d) and test accuracy and loss (e-f) for the alpha values, 0.1, 0.5, and 0.9 compared to the independent baseline for the ResNet-18 architecture.

functional resemblance to that of the trained teacher across the similarity metrics explored, with an alpha value of 0.9, resulting in the most similar models when compared to the baselines, which is in direct contrast the ResNet18 architecture results presented in Table 1.

For the ViT, the null hypothesis was suitably rejected for the JS Divergence and Logits Cosine Similarity hypothesis tests across all alpha values when compared to the baselines, concluding that for the ViT, the KD does result in more functionally similar models regarding JS Divergence and Logits Cosine Similarity between the student and the teacher. The null hypothesis was suitably rejected for predictive dissimilarity when using alpha values 0.5 and 0.9 and for Activation Distance when using an alpha value of 0.9. Concluding that, for the ViT, KD can result in models that are more functionally similar to the teachers, indicating there is a transfer of knowledge between the teacher and student. Although the increase in similarity is statistically significant, the increase between the student and the teacher could be considered marginal, as the maximum increase achieved across metrics is 0.0266 when using an alpha of 0.9; as the experimental setup 2 attempts to maximise functional similarity, this could be viewed as an upper bound of knowledge transfer.

Interestingly, although KD resulted in models more functionally similar to their teacher, this increase in similarity does not adequately explain the performance increase achieved through using KD. This is because the model with the best test accuracy was Uniform Noise KD, with an alpha of 0.5. In conjunction with the results for ResNet18, this result provides strong evidence to suggest that KD is predominately a regulariser first and a weak knowledge transfer mechanism second.

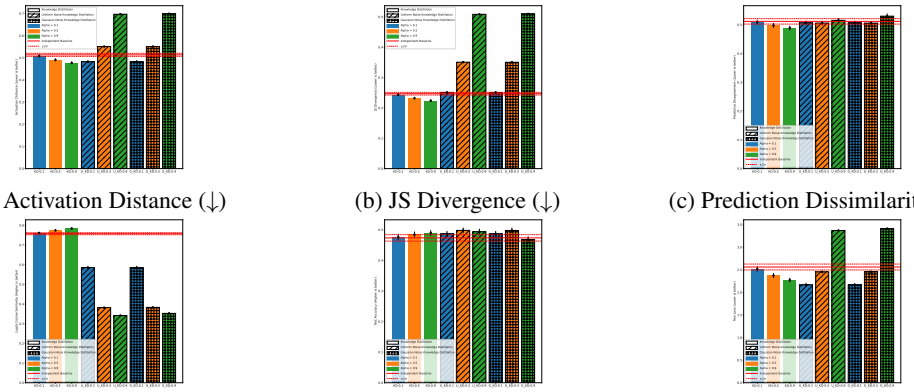
Metrics		Knowledge Distillation (KD)			Uniform Noise KD			Gaussian Noise KD		
	Independent	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Activation Distance (↓)	0.5135+-0.0034,	0.5084+-0.0033,	0.4905+-0.0033,	**0.4773+-0.0030,**	0.4826+-0.0029,	0.5512+-0.0016,	0.6969+-0.0003,	0.4826+-0.0028,	0.5511+-0.0016,	0.6994+-0.0005
JS Divergence (↓)	0.2469+-0.0022,	0.2436+-0.0021,	0.2324+-0.0021,	**0.2239+-0.0020,**	0.2511+-0.0019,	0.3508+-0.0013,	0.5086+-0.0006,	0.2511+-0.0018,	0.3505+-0.0012,	0.5100+-0.0007
Prediction Dissimilarity (↓)	0.5119+-0.0047,	0.5086+-0.0044,	0.4976+-0.0043,	**0.4889+-0.0039,**	0.5065+-0.0039,	0.5062+-0.0039,	0.5154+-0.0037,	0.5066+-0.0038,	0.5059+-0.0037,	0.5300+-0.0047
Logits Cosine Similarity (↑)	0.7576+-0.0025,	0.7614+-0.0026,	0.7747+-0.0024,	**0.7842+-0.0023,**	0.5851+-0.0044,	0.3818+-0.0019,	0.3418+-0.0014,	0.5851+-0.0044,	0.3830+-0.0018,	0.3526+-0.0016
Test Accuracy (↑)	0.4736+-0.0053	0.4762+-0.0052,	0.4843+-0.0050,	0.4897+-0.0045,	0.4867+-0.0047,	**0.4981+-0.0046,**	0.4936+-0.0048,	0.4867+-0.0048,	0.4979+-0.0046,	0.4679+-0.0056
Test Loss (↓)	2.5623+-0.0326,	2.5181+-0.0299,	2.3728+-0.0272,	2.2736+-0.0245,	**2.1656+-0.0205,**	2.4548+-0.0160,	3.3705+-0.0100,	2.1657+-0.0207,	2.4339+-0.0167,	3.4102+-0.0126

Table 2: ViT Results for CIFAR100

5 Discussion

Knowledge distillation is not a compression mechanism. From our results across the range of architectures explored, it is evident that a traditional KD setup cannot be considered a compression mechanism from a functional perspective. Given that in the contrived experiments, where the model is given every chance to ***identically*** replicate the teacher, through having the same ***initialisation***, and only differing the training data order, the trained models using KD, were never able to reach within 95% similarity of the teacher model. As a result, we argue that KD is not a compression mechanism.

Knowledge distillation can transfer knowledge, but this does not adequately explain the performance increase. Our results show that knowledge distillation can sometimes induce a transfer of "knowledge"; however, the functional increase, although significant above an independent baseline, is marginal. Albeit, knowledge transfer does not directly play a role in the performance benefits obtained



(a) Activation Distance (↓) (b) JS Divergence (↓) (c) Prediction Dissimilarity (↓)

(d) Logits Cosine Similarity (↑) (e) Test Accuracy (↑) (f) Test Loss (↓)

Figure 2: Functional similarity across each metric (a-d) and test accuracy and loss (e-f) for the alpha values, 0.1, 0.5, and 0.9 compared to the independent baseline for the ViT architecture.

through KD. When using Uniform or Gaussian Noise as the teacher, similar or better performance than the baseline and conventional (trained teacher) KD setups can be achieved. Given that this is a contrived setup to encourage functional similarity as much as possible through sharing the same initialisation as the teacher, KD should instead be perceived and understood as a ***data-dependant regulariser*** due to the limited functional similarity achieved and the unclear relationship between the improved similarity and performance increases.

Safety implications of findings. As a result of our findings, a KD training set-up should be used carefully when using unverified teachers with potential backdoors, as KD has been shown in some cases to transfer knowledge to the student model. However, it requires a high alpha value. We would also argue that backdoors could be more easily transferred due to the data-dependant regularisation offered by KD, which would artificially trigger larger weight updates due to a higher loss for adversarial examples within batches. Therefore, in scenarios where the teacher model cannot be effectively verified, this work suggests that Uniform or Gaussian noise KD should be used instead as it exhibits a similar regularisation effect and can realize accuracy benefits.

Computational expense of knowledge distillation. Provided the training setup of KD is complex and requires inference of the teacher model for each update to create the combined loss, it incurs significant computational costs. Given that we show in our results that the KD setup does not always create a significantly similar model and that this is not important for performance improvements, we argue that this computational expense is excessive. Similar or significantly improved accuracy gains can be realised using the Uniform or Gaussian Noise as the teacher, which does not require inference. With correct hyperparameter optimisation for the alpha value for Uniform or Gaussian Noise, the expensive training setup of KD can be subverted, and accuracy improvements can be recorded, meaning that the training setup of KD is still of use to the community.

6 Conclusion

In this body of work, we concretely reject the hypothesis that knowledge distillation is a compression mechanism. In a setup where data order is the only barrier to full functional similarity, we show that there is *significant but marginal knowledge transfer*. Moreover, through the use of Uniform and Gaussian Noise KD control setups and their significant performance gains, we show that the performance benefits provided by KD are not adequately explained by such *significant but marginal knowledge transfer*. In most cases, using Uniform or Gaussian Noise as the teacher would suffice to reduce the computational cost of inference on a trained teacher. Furthermore, our work contributes to demystifying the nature of knowledge distillation and the notion of knowledge transfer, showing that knowledge distillation should be considered as a ***data-dependant regulariser with weak knowledge sharing*** making it an inadequate compression technique.

Acknowledgments and Disclosure of Funding

This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT [14].

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. [Online]. Available: https://arxiv.org/pdf/2304.02643.pdf

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[4] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2006, p. 535–541. [Online]. Available: https://doi.org/10.1145/1150402.1150464

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. [Online]. Available: https://arxiv.org/pdf/1503.02531.pdf

[6] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3713–3722. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/papers/Zhang_Be_Your_Own_Teacher_Improve_the_Performance_of_Convolutional_Neural_ICCV_2019_paper.pdf

[7] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=Uuf2q9TfXGA

[8] S. Stanton, P. Izmailov, P. Kirichenko, A. A. Alemi, and A. G. Wilson, "Does knowledge distillation really work?" in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 6906–6919. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/376c6b9ff3bedbbea56751a84fffc10c-Paper.pdf

[9] F. Sarnthein, G. Bachmann, S. Anagnostidis, and T. Hofmann, "Random teachers are good teachers," in *International Conference on Machine Learning*, 2023. [Online]. Available: https://arxiv.org/abs/2302.12091

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[12] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[14] T. King, S. Butcher, and L. Zalewski, *Apocrita - High Performance Computing Cluster for Queen Mary University of London*, Mar. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.438045

[15] C. Buciluundefined, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 535–541. [Online]. Available: https://doi.org/10.1145/1150402.1150464

[16] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 925–10 934. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Beyer_Knowledge_Distillation_A_Good_Teacher_Is_Patient_and_Consistent_CVPR_2022_paper.pdf

[17] J.-W. Jung, H.-S. Heo, H.-J. Shim, and H.-J. Yu, "Knowledge distillation in acoustic scene classification," *IEEE Access*, vol. 8, pp. 166 870–166 879, 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9186616

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019. [Online]. Available: https://arxiv.org/pdf/1910.01108.pdf

[19] N. Aghli and E. Ribeiro, "Combining weight pruning and knowledge distillation for cnn compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3191–3198. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021W/EVW/papers/Aghli_Combining_Weight_Pruning_and_Knowledge_Distillation_for_CNN_Compression_CVPRW_2021_paper.pdf

[20] T. Li, J. Li, Z. Liu, and C. Zhang, "Few sample knowledge distillation for efficient network compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 639–14 647. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Few_Sample_Knowledge_Distillation_for_Efficient_Network_Compression_CVPR_2020_paper.pdf

[21] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu, "Compressing visual-linguistic model via knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1428–1438. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/papers/Fang_Compressing_Visual-Linguistic_Model_via_Knowledge_Distillation_ICCV_2021_paper.pdf

[22] C. Wang, Q. Yang, R. Huang, S. Song, and G. Huang, "Efficient knowledge distillation from model checkpoints," *Advances in Neural Information Processing Systems*, vol. 35, pp. 607–619, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/03e0712bf85ebe7cec4f1a7fc53216c9-Paper-Conference.pdf

[23] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," *Advances in neural information processing systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/e1e32e235eee1f970470a3a6658dfdd5-Paper.pdf

[24] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 716–33 727, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/da669dfd3c36c93905a17ddba01eef06-Paper-Conference.pdf

[25] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701. [Online]. Available: https://www.isca-archive.org/interspeech_2017/fukuda17_interspeech.pdf

[26] R. D. Nathoo, M. Kegler, and M. Stamenovic, "Two-step knowledge distillation for tiny speech enhancement," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 141–10 145. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10446796

[27] D. Bergmann, "What is knowledge distillation," 2024. [Online]. Available: https://www.ibm.com/topics/knowledge-distillation

[28] S. Teki, "Knowledge distillation: Principles, algorithms, applications," 2023. [Online]. Available: https://neptune.ai/blog/knowledge-distillation

[29] R. Kundu, "Knowledge distillation: Principles algorithms [+applications]," 2022. [Online]. Available: https://www.v7labs.com/blog/knowledge-distillation-guide

[30] E. Library, "Introduction to knowledge distillation," 2023. [Online]. Available: https://www.edenlibrary.ai/introduction-to-knowledge-distillation/

[31] T. Team, "Knowledge distillation: a way to make a large model more efficient and accessible," 2024. [Online]. Available: https://toloka.ai/blog/knowledge-distillation/

[32] P. Potrimba, "What is knowledge distillation? a deep dive." 2023. [Online]. Available: https://blog.roboflow.com/what-is-knowledge-distillation/

[33] I. Mason-Williams, "NEURAL NETWORK COMPRESSION: THE FUNCTIONAL PERSPECTIVE," in *5th Workshop on practical ML for limited/low resource settings*, 2024. [Online]. Available: https://openreview.net/forum?id=Q7GXKjmCSB

[34] S. Stanton, P. Izmailov, P. Kirichenko, A. A. Alemi, and A. G. Wilson, "Does knowledge distillation really work?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 6906–6919, 2021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/376c6b9ff3bedbbea56751a84fffc10c-Paper.pdf

[35] U. Ojha, Y. Li, A. Sundara Rajan, Y. Liang, and Y. J. Lee, "What knowledge gets distilled in knowledge distillation?" *Advances in Neural Information Processing Systems*, vol. 36, pp. 11 037–11 048, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/2433fec2144ccf5fea1c9c5ebdbc3924-Paper-Conference.pdf

[36] A. Golatkar, A. Achille, A. Ravichandran, M. Polito, and S. Soatto, "Mixed-privacy forgetting in deep networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 792–801. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/papers/Golatkar_Mixed-Privacy_Forgetting_in_Deep_Networks_CVPR_2021_paper.pdf

[37] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. [Online]. Available: https://doi.org/10.1609/aaai.v37i6.25879

[38] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," *arXiv preprint arXiv:1912.02757*, 2019. [Online]. Available: https://arxiv.org/pdf/1912.02757

[39] G. Mason-Williams and F. Dahlqvist, "What makes a good prune? maximal unstructured pruning for maximal cosine similarity," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=jsvvPVVzwf

[40] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991. [Online]. Available: https://ieeexplore.ieee.org/document/61115

A Debunking Challenge

The introduction and nomenclature of knowledge distillation infer that it relies upon a trained teacher being able to transfer knowledge to a student model [5]. Since its introduction, the notion has been fortified in academic literature [15; 5; 16; 17; 18; 19; 20; 21; 22; 6; 23; 24; 25; 26] and been disseminated via blog posts [27; 28; 29; 30; 31; 32], that it is possible and a trivial process to transfer one model into another or to condense a larger model into a smaller model. As a result, it is baked into common wisdom that knowledge sharing is the key component of performance improvements witnessed in the student-teacher dynamic.

Our body of work states that this notion remains up for debate as despite recording a significant functional similarity for the VGG and ViT students to the teacher - we consider the level of knowledge sharing marginal. Provided that we give students and teacher the same instillation and only change the order of training data, we would expect knowledge distillation with effective knowledge transfer to create a functionally indistinguishable student model. As a result, while we acknowledge knowledge sharing, we do not think it plays a major role in performance benefits witnessed in knowledge distillation setups. Subsequently, we argue that knowledge distillation can be adequately described as a data-dependant regularisation technique, this perception aligns with literature showing that employing KD can increase performance. Our results also reaffirm that knowledge transfer is not the root cause for the performance increase witnessed when using knowledge distillation, as we see performance increases when using the Uniform and Gaussian Noise in the KD training setup. For both the Uniform and Gaussian Noise, no knowledge can be transferred from teacher to student, but we still manage to observe performance increase; this leads the authors to believe that this is due to forcing weight updates by artificially increasing the loss. Moreover, the study averages results over considerable trials, meaning that conclusions drawn are robust and help to move away from championing knowledge transfer in the knowledge distillation framework. The authors believe that the results call for caution when applying anthropomorphic terms to describe the dynamics of neural networks, as it can lead to significant misunderstandings that appear plausible but do not tell the whole story when considering artificial forms of cognition.

The work represents an essential step towards decoupling the misunderstandings caused by naming conventions and providing a framework for conducting such experiments. We hope the community shares in the nuanced understanding of knowledge distillation as a data-dependant regulariser instead of a knowledge-based transferring mechanism and that, broadly speaking, it cannot be considered a reliable or effective compression mechanism.

B Related Work

B.1 Knowledge Distillation

Knowledge distillation [15; 5] was introduced as a way to compress the information of a single or ensemble of teachers into a single model. Since its ideation the fundamental idea of knowledge transfer has been employed across modalities [16; 17; 18] and is referenced as an effective compression method [19; 20; 21; 22]. Many variations of knowledge distillation setups have been developed such as the self-distillation for improving model performance where the student and teacher have the same architecture and capacity [6]. However, despite empirical findings of correlation between student and teacher performance recent work has questioned the role of knowledge transfer in this dynamic [33; 34; 35]; our work contributes similarly to this endeavour of understanding and probing the dynamics of knowledge distillation.

B.2 Functional Similarity Metrics

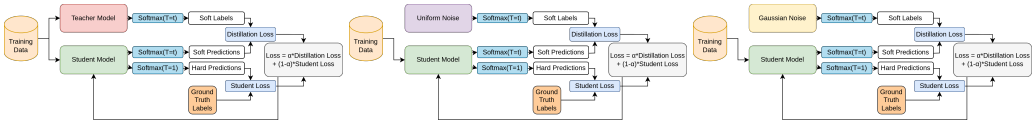
Functional similarity of model outputs has been compared in literature ranging from use cases in verifying unlearning approaches [36; 37], explaining the dynamics of ensembles [38] and as a method for understanding pruning [39]. Metrics such as Activation Distance, JS Divergence, Prediction Dissimilarity and Cosine Similarity have been used for functional analysis. Activation Distance represents the $\mathcal{L}_2$ distance on the softmax output distribution of two models, enabling functional comparison. While JS Divergence represents the Jensen-Shannon information-theoretic divergence that employs a weighted average of KL divergence of distributions, giving a directed divergence between non-continuous distributions [40]. As a result, JS Divergence can be employed to gauge the loss divergence of two models on the same data. Prediction Dissimilarity compares the disagreement of label predictions between models, allowing for an enriched perspective on the alignment of the model's functions [38]. Cosine Similarity is a metric to measure the cosine angle between two vectors; a value close to 1 suggests a similar representation and a value close to zero suggests orthogonal outputs, with -1 representing polar outputs.

C Limitations

For this experiment we acknowledge that our findings could be more robust if we explored more architectures and datasets. In doing this, we could further validate our findings to understand how much knowledge transfer traditional KD offers. Additionally, we have only looked at the vision domain in this experimental setup. Though we expect our results to transfer to any other modality, it would be worthwhile to invest computational effort to confirm this. Moreover, this experimental setup only implemented the base setup of KD. As a result, our conclusions are constrained to this KD implementation. To further confirm our findings, it is essential to look at other KD implementations and see if the same findings hold, though we expect this to be the case. Finally, our results could have benefited from a more representative range of alpha values and temperature values other than 1.

D Diagrammatic Representation of Learning Processes

Figure 3 shows a visual representation of the knowledge distillation, Uniform noise distillation and Gaussian noise distillation setups designed and employed in the body of the paper.



(a) Knowledge Distillation (b) Uniform Noise Distillation (c) Gaussian Noise Distillation

Figure 3: Diagram of Knowledge Distillation Process with a (a) Teacher Model (b) Uniform Noise and (c) Gaussian Noise.

E ResNet18

E.1 Training details

The ResNet 18 architecture is trained with a batch size of 128 using Stochastic Gradient Descent (SGD), with a learning rate of 0.1, a momentum of 0.9 and weight decay of 0.0005 for 100 epochs, along with a multi-step scheduler with a gamma of 0.2 after epoch 60. The ViT architecture is trained with a batch size of 128 using AdamW, with a learning rate of 5e-4 and a weight decay of 1e-3 for 70 epochs. A linear warm-up to the learning rate of 5e-4 is used for 10 epochs, followed by a cosine decay to a learning rate of 1e-5 for the remaining 60 epochs.

F VGG-16-BN

F.1 Training details

The VGG-16-BN architecture is trained with a batch size of 128 using Stochastic Gradient Descent (SGD), with a learning rate of 0.1, a momentum of 0.9 and a weight decay of 0.0005 for 50 epochs, along with a multi-step scheduler with a gamma of 0.2 after epoch 30.

F.2 Results

The VGG-16-BN architecture results presented in Table 3 and Figure 4, displaying the mean and standard deviation from 385 runs, show that the knowledge distillation models bear the most functional resemblance to that of the trained teacher across the similarity metrics explored, with an alpha value of 0.9, resulting in the most similar models when compared to the baselines, which is in direct contrast the ResNet18 architecture results presented in Table 1.

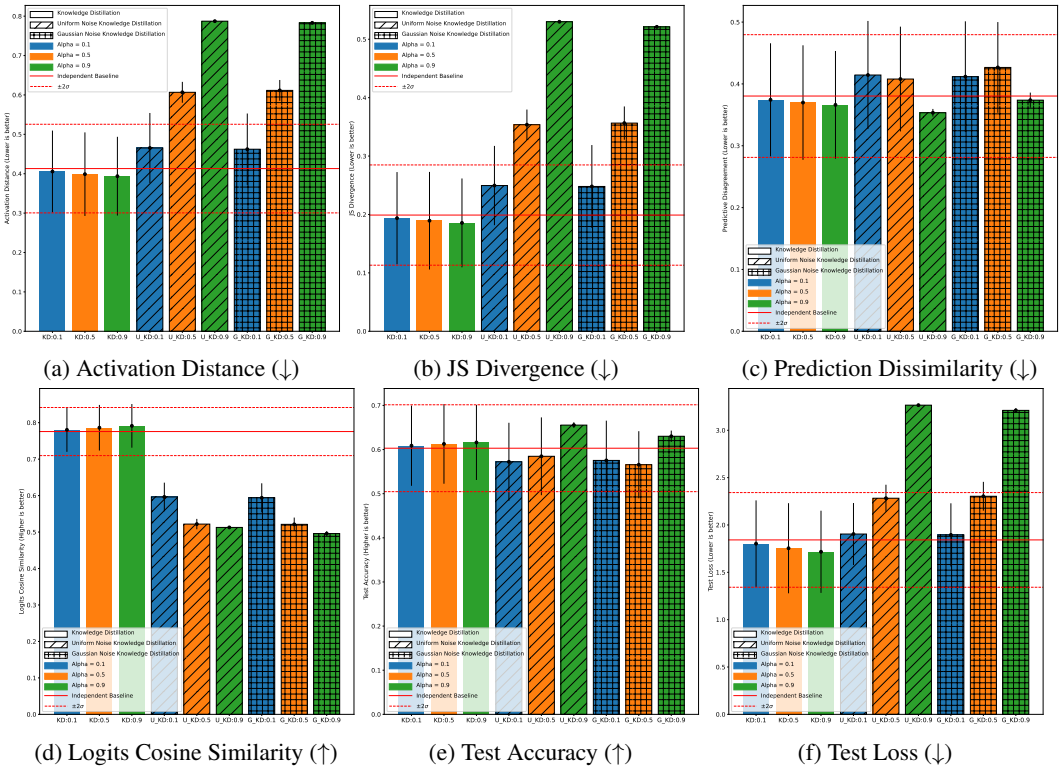
For the VGG-16-BN, when using an alpha value of 0.1, the null hypothesis failed to be rejected for the Activation Distance, JS Divergence. However, for alpha values of 0.5 and 0.9, the null hypothesis was suitably rejected for the Activation Distance, JS Divergence hypothesis tests when compared to the baselines. The null hypothesis was suitably rejected for the Logits Cosine Similarity hypothesis

tests when compared to the baselines with all alpha values. Concluding that for the VGG-16-BN, the KD can result in more functionally similar models regarding Activation Distance, JS Divergence and Logits Cosine Similarity between the student and the teacher. The null hypothesis failed to be rejected for predictive dissimilarity for all alpha values. In conclusion, for the VGG-16-BN, KD can result in models that are more functionally similar to the teachers regarding Activation Distance, JS Divergence and Logits Cosine Similarity, indicating a transfer of knowledge between the teacher and student. Although the increase in similarity is statistically significant, the increase between the student and the teacher could be considered marginal, as the maximum increase achieved across metrics is 0.0193 when using an alpha of 0.9.

Interestingly, although KD resulted in models more functionally similar to their teacher, this increase in similarity does not adequately explain the performance increase achieved through using KD. This is because the model with the best test accuracy was Uniform Noise KD, with an alpha of 0.9. In conjunction with the results for ResNet18 and the ViT, this result provides strong evidence to suggest that KD is predominately a regulariser first and a knowledge transfer mechanism second.

Metrics		Knowledge Distillation (KD)			Uniform Noise KD			Gaussian Noise KD		
	Independent	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Activation Distance (↓)	0.4131+-0.0562,	0.4057+-0.0519,	0.3989+-0.0530,	**0.3938+-0.0500,**	0.4654+-0.0444,	0.6067+-0.0133,	0.7873+-0.0003,	0.4622+-0.0453,	0.6115+-0.0132,	0.7831+-0.0005
JS Divergence (↓)	0.1990+-0.0429,	0.1937+-0.0395,	0.1893+-0.0418,	**0.1856+-0.0380,**	0.2495+-0.0339,	0.3537+-0.0130,	0.5300+-0.0007,	0.2480+-0.0354,	0.3565+-0.0142,	0.5212+-0.0012
Prediction Dissimilarity (↓)	0.3804+-0.0496,	0.3743+-0.0457,	0.3699+-0.0463,	0.3661+-0.0436,	0.4144+-0.0437,	0.4079+-0.0424,	**0.3534+-0.0031,**	0.4117+-0.0448,	0.4266+-0.0367,	0.3737+-0.0061
Logits Cosine Similarity (↑)	0.7757+-0.0330,	0.7804+-0.0302,	0.7862+-0.0313,	**0.7913+-0.0299,**	0.5963+-0.0196,	0.5219+-0.0071,	0.5127+-0.0014,	0.5940+-0.0199,	0.5212+-0.0093,	0.4958+-0.0033
Test Accuracy (↑)	0.6030+-0.0492,	0.6088+-0.0455,	0.6129+-0.0451,	0.6161+-0.0425,	0.5723+-0.0442,	0.5847+-0.0441,	**0.6554+-0.0032,**	0.5754+-0.0451,	0.5656+-0.0380,	0.6301+-0.0064
Test Loss (↓)	1.8423+-0.2500,	1.8025+-0.2286,	1.7542+-0.2379,	**1.7168+-0.2170,**	1.9044+-0.1637,	2.2825+-0.0713,	3.2653+-0.0060,	1.8953+-0.1668,	2.3033+-0.0761,	3.2103+-0.0117

Table 3: VGG-16-BN Results



(a) Activation Distance (↓) (b) JS Divergence (↓) (c) Prediction Dissimilarity (↓)

(d) Logits Cosine Similarity (↑) (e) Test Accuracy (↑) (f) Test Loss (↓)

Figure 4: Functional similarity across each metric (a-d) and test accuracy and loss (e-f) for the alpha values, 0.1, 0.5, and 0.9 compared to the independent baseline for the VGG-16-BN architecture trained on CIFAR100. Mean, and 2 standard deviations are reported from 385 runs.

G ViT

G.1 Training details

The ViT architecture is configured with an image size of 32, a patch size of 4, an embedding dimension of 128, 4 attention heads, a forward multiplier of 2, a dropout set at 0.1 and a total of 6 layers, with a classifier layer of 100. The architecture is trained with a batch size of 128 using AdamW, with a learning rate of 5e-4 and a weight decay of 1e-3 for 70 epochs. A linear warm-up to the learning rate of 5e-4 is used for 10 epochs, followed by a cosine decay to a learning rate of 1e-5 for the remaining 60 epochs.

H P-Value

H.1 ResNet18

The Z-score P-Values for ResNet18 KD against an independent model are shown in Table 4. It shows that the null hypothesis failed to be rejected for all metrics.

	Alpha Values		
Functional Similarity Metrics	**0.1**	**0.5**	**0.9**
Activation Distance	0.956	0.975	0.975
JS Divergence	0.959	0.982	0.981
Predictive Dissimilarity	0.954	0.966	0.983
Cosine Similarity	0.949	0.895	0.821

Table 4: ResNet18 P-Values for Z-score Test: Knowledge Distillation vs Independent Models

H.2 ViT

The Z-score P-Values for ViT KD against the independent models and the Uniform Noise KD with an alpha of 0.1 are shown in Table 5 and 6, respectively. These tables show that the null hypothesis was suitably rejected for the JS Divergence and Logits Cosine Similarity and that the null hypothesis was suitably rejected for Predictive Dissimilarity when using alpha values 0.5 and 0.9 and for Activation Distance when using alpha values above 0.9.

	Alpha Values		
Functional Similarity Metrics	**0.1**	**0.5**	**0.9**
Activation Distance	0.000	0.000	0.000
JS Divergence	0.000	0.000	0.000
Predictive Dissimilarity	0.000	0.000	0.000
Cosine Similarity	0.000	0.000	0.000

Table 5: ViT P-Values for Z-score Test: Knowledge Distillation vs Independent Models

	Alpha Values		
Functional Similarity Metrics	**0.1**	**0.5**	**0.9**
Activation Distance	1.000	1.000	0.000
JS Divergence	0.000	0.000	0.000
Predictive Dissimilarity	1.000	0.000	0.000
Cosine Similarity	0.000	0.000	0.000

Table 6: ViT P-Values for Z-score Test: Knowledge Distillation vs Uniform Noise Knowledge Distillation with Alpha 0.1

H.3 VGG-16-BN

The Z-score P-Values for VGG-16-BN KD against the independent models and the Uniform Noise KD with an alpha of 0.9 are shown in Table 7 and 8, respectively. These tables show that the null hypothesis failed to be rejected for JS Divergence and Activation Distance with an alpha of 0.1 and Predictive Dissimilarity for all alpha values. However, the null hypothesis was suitably rejected for the JS Divergence and Activation Distance with an alpha of 0.5 and 0.9 and Cosine Similarity for all alpha values.

	Alpha Values		
Functional Similarity Metrics	**0.1**	**0.5**	**0.9**
Activation Distance	0.029	0.000	0.000
JS Divergence	0.037	0.001	0.000
Predictive Dissimilarity	0.038	0.001	0.000
Cosine Similarity	0.019	0.000	0.000

Table 7: VGG-16-BN P-Values for Z-score Test: Knowledge Distillation vs Independent Models

	Alpha Values		
Functional Similarity Metrics	**0.1**	**0.5**	**0.9**
Activation Distance	0.000	0.000	0.000
JS Divergence	0.000	0.000	0.000
Predictive Dissimilarity	1.000	1.000	1.000
Cosine Similarity	0.000	0.000	0.000

Table 8: VGG-16-BN P-Values for Z-score Test: Knowledge Distillation vs Uniform Knowledge Distillation with Alpha 0.9