Investigating the Robustness of Subtask Distillation under Spurious Correlation

Pattarawat Chormai[1,2], Klaus-Robert Müller[1,3,4,5], and Grégoire Montavon[6,3,*]
[1]Machine Learning Group, Technische Universität Berlin (TU Berlin), 10587 Berlin, Germany
[2]Zuse School ELIZA, 64289 Darmstadt, Germany
[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
[4]Department of Artificial Intelligence, Korea University, Seoul 136-713, Korea
[5]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany
[6]Institute for AI in Medicine, Charité – Universitätsmedizin Berlin, 10115 Berlin, Germany

***Abstract*—Subtask distillation is an emerging paradigm in which compact, specialized models are extracted from large, general-purpose 'foundation models' for deployment in environments with limited resources or in standalone computer systems. Although distillation uses a teacher model, it still relies on a dataset that is often limited in size and may lack representativeness or exhibit spurious correlations. In this paper, we evaluate established distillation methods, as well as the recent SubDistill method, when using data with spurious correlations for distillation. As the strength of the correlations increases, we observe a widening gap between advanced methods, such as SubDistill, which remain fairly robust, and some baseline methods, which degrade to near-random performance. Overall, our study underscores the challenges of knowledge distillation when applied to imperfect, real-world datasets, particularly those with spurious correlations.**

I. Introduction

Recent years have seen a surge in the availability of large, general-purpose, 'open-weight' ML models [1], [2]. These models can be applied off-the-shelf for tasks such as object detection and question answering, and also provide abstract representations that serve as a starting point for learning specialized tasks. While these models have shown high predictive performance, they are often too large to run on standard computers. Instead, they require expensive 'inference servers', which limit their scope of application. Distillation, specifically 'subtask distillation' [3], [4], [5], offers a solution to transform these large models into more compact ones that can run on smaller machines, such as laptops or mobile phones. Recently, SubDistill [5], a method that identifies relevant structures in the teacher's model, was proposed as an effective technique for subtask distillation.

In addition to developing new distillation mechanisms, it is crucial to test the performance and robustness of the proposed approaches in a wide range of realistic settings. Thus far, benchmark evaluations have primarily focused on scenarios in which the data available for distillation is a finite, yet representative, subset of the underlying data distribution. In this paper, we focus on evaluating distillation performance when the available data is affected by spurious correlations. Spurious correlations are common in many real-world datasets [6], including clinical datasets, where they arise, e.g. from incomplete population coverage or technical confounders [7], [8], [9]. These correlations have been shown to give rise to 'Clever Hans' classifiers [10], [11], which exhibit unreliable predictive behavior and pose significant practical risks.

We conduct a benchmark evaluation of distillation techniques on popular convolutional and transformer vision models while varying the level of spurious correlation in the data. Our results demonstrate that spurious correlations significantly impact the performance of different distillation methods, albeit to varying degrees. While simple distillation baselines degrade to near-random predictions in some cases, SubDistill [5] remains fairly stable with increased levels of spurious correlations. We attribute SubDistill's stability to its superior student-teacher alignment capabilities, which prevent Clever Hans strategies from emerging. Our study also concurs with the earlier findings of [12], [13], which underscore the importance of aligning the student's and the teacher's prediction strategies. We extend their work by focusing on subtask distillation and providing a benchmark evaluation covering both recent and established distillation techniques and a variety of teacher/student architectures. Our code is available at github.com/p16i/subdistill.

II. Background

In this section, we give a concise tutorial-focused discussion of the relevant family of distillation methods. While we emphasize simple mathematical formulations for clarity, we also provide references to corresponding methods from the literature.

The simplest form of model distillation consists of incentivizing the student model to match the teacher's outputs through the application of a loss function penalizing the difference between the two. Denoting $f_S, f_T \colon \mathbb{R}^d \to \mathbb{R}^C$ the functions implemented by the student and teacher models respectively, the distillation loss can be expressed as:

$$\mathcal{E}_{\text{distill}}(x) = \ell(f_S(x), f_T(x)), \tag{1}$$

where ℓ can be the KL divergence between the student and the teacher for classification tasks [14], or the squared norm

*Corresponding author: gregoire.montavon@charite.de

[15] for multivariate regression. In practice, not all C output dimensions of the function f_T may be relevant for distillation. For example, among all image categories a foundation model can predict, only classifying between subtypes of e.g. 'wading birds' might be relevant for the downstream application. In such a case, we can design a function $g\colon \mathbb{R}^C \to \mathbb{R}^{C'}$ that only retains the $C' \ll C$ task-relevant logits. In that case, after adapting the student model output accordingly, we can reformulate the distillation loss as:

$$\mathcal{E}_{\text{subtask}}(x) = \ell(f_S(x), g \circ f_T(x)). \tag{2}$$

To foster tighter teacher-student alignment, it is beneficial to also constrain the student's internal representations. The approach is known as layer-wise distillation and formulates a loss function of the type:

$$\mathcal{E}_{\text{layerwise}}(x) = \mathcal{E}_{\text{distill}}(x) + \sum_{l=1}^{L} \lambda_l \underbrace{\|a_S^{(l)} - a_T^{(l)}\|^2}_{\mathcal{E}_l(x)} \tag{3}$$

where $\sum_l$ iterates over a list of layer pairings between the teacher and the student, and where $a_S^{(l)} \in \mathbb{R}^{K_l}$ and $a_T^{(l)} \in \mathbb{R}^{d_l}$ are the student's and teacher's lth-layer activations associated to the data point x. Because student's activation vectors typically differ in size and structure [16], [17], the rudimentary formulation of Eq. (3) is typically enhanced with subnetworks (e.g. a linear map [18], [16] or a small MLP [19]) at each layer that adapt the teacher and student representations, in particular, addressing their dimensionality mismatch.

A. SubDistill Method

SubDistill is a recently proposed approach [5] that combines some of the design choices above and embeds them in a practical, numerically efficient algorithm for subtask distillation. SubDistill has two components: first, it identifies the part of the representation at each layer that is most relevant for the subtask. If we consider an activation vector a_T at some layer, we can construct a corresponding vector c_T of the same dimensions that measures the sensitivity of those activations to the subtask of interest (details in [5]). While the product $a_T \odot c_T$ readily provides a feature-wise indicator of subtask relevance, SubDistill proceeds further by identifying a maximally relevant subspace U in the teacher's activation space:

$$\max_U \ \big\{\mathbb{E}\big[\langle a_T, c_T\rangle_U\big] + \Omega(a_T, c_T)\big\}, \tag{4}$$

where we also enforce the orthogonality constraint $U^\top U = I_K$ and where $\Omega(a_T, c_T)$ stands for additional terms that ensure the objective is positive definite (cf. [5] for the exact objective formulation). Once teacher subspaces are identified at each layer, SubDistill constructs the following loss function for aligning teacher and student at the given layer:

$$\mathcal{E}_l(x) = \mathbb{E}\big[\big\|V(a_S - \mu_S) - U^\top(a_T - \mu_T)\big\|_2^2\big] \tag{5}$$

where μ_S and μ_T are the student's and teacher's mean activations, and where V is constrained to be orthogonal (similar to [20]) and represents the student's subspace rotation. The mean corrections and orthogonality constraint ensure numerical efficiency and representation alignment both before and after subspace projection. To handle potential magnitude differences, each $\mathcal{E}_l$ is normalized by $\|U^\top(a_T - \mu_T)\|_2^2$.

Overall, SubDistill operationalizes the conceptual formulations in Eqs. (1)–(3) into a numerically stable and subtask-informed distillation technique.

III. Empirical Evaluation

The main objective of this paper is to evaluate the performance and stability of different subtask distillation methods when trained on data exhibiting spurious correlation. For this, we consider the task of distilling from a fully trained ImageNet model a student, which we require to specialize on a specific subtask. We set as a subtask the classification of different 'wading birds' ('spoonbill', 'flamingo', 'crane', 'limpkin', and 'bustard'), which forms a subset of 5 classes from the original 1000 ImageNet classes. We synthetically add to the top left corner of some ImageNet images (of size 224×224) an MNIST digit (rescaled to 56×56). We generate spurious correlations in the training data by choosing MNIST digits according to the image class (i.e., spoonbill $\leftrightarrow$ MNIST-0, flamingo $\leftrightarrow$ MNIST-1, etc.), while we pick the digits randomly for the test set. This model for spurious correlation is illustrated in Fig. 1. We denote by ρ the proportion of training and test ImageNet images that we contaminate and experiment with contamination rates $\rho \in \{0\%, 50\%, 100\%\}$, which can be interpreted as adding varying levels of spurious correlation to the training data.

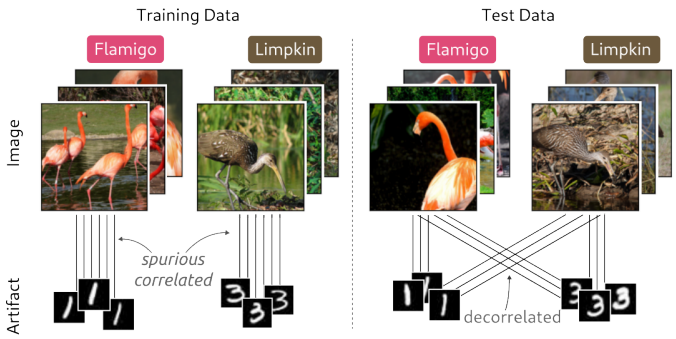

Fig. 1. Spurious correlation model considered in this work, where MNIST digits are inserted in the top-left corner of ImageNet images. On the training data, the digit's class is spuriously correlated to the true image class (bird type), whereas on the test data the digits are assigned randomly.

To cover a broad range of practical scenarios, we perform the evaluation on three teacher-student pairs. The first setup transfers from a standard ResNet18 [22] to 'ResNet18-S', a smaller ResNet18 student (for which we adjust its first three blocks and the last block to be 32 and 24, respectively). We then experiment with distilling from WideResNet101 [23] to MBNetv4 [24], a setup in which the teacher is substantially larger, and the student has a specialized architecture for deployment on mobile devices. Finally, we experiment with distilling from ViTB16 [25] to EffFormerv2 [26], a setup that

TABLE I

TEST SET ACCURACY OF STUDENTS TRAINED WITH DIFFERENT DISTILLATION APPROACHES UNDER DIFFERENT MNIST-DIGIT CONTAMINATION RATES (I.E. DIFFERENT LEVELS OF SPURIOUS CORRELATION) ON THE IMAGENET 'WADING BIRD' SUBTASK. THE 'REF' COLUMN SHOWS THE TEACHER'S ACCURACY. THE ACCURACY OF EACH CONFIGURATION IS AVERAGED OVER THREE RANDOM INITIALIZATIONS, AND THE ENTRY IS HIGHLIGHTED IN RED, ORANGE, AND YELLOW IF THE ACCURACY FALLS IN THE RANGES $[0, 50]$, $[50, 75]$, AND $[75, 90]$, RESPECTIVELY. THE LAST COLUMN IS THE LARGEST STANDARD ERROR OF EACH ROW.

Teacher → Student	Contamination Rate (%)	*Ref.*	Output Only [14]	AT [21]	VID [19]	VKD [20]	SubDistill [5]	*max. err.*
ResNet18 → ResNet18-S	0	*98.0*	90.8	91.3	92.3	92.0	93.6	± 1.1
	50	*97.6*	83.9	86.9	91.6	89.1	91.7	± 0.9
	100	*97.2*	45.6	52.0	61.6	57.2	86.4	± 2.1
WideResNet101 → MBNetv4	0	*98.4*	90.0	91.3	90.5	91.1	96.8	± 1.9
	50	*98.0*	62.7	63.3	67.1	75.1	92.7	± 1.2
	100	*98.4*	32.0	29.3	28.4	30.7	66.9	± 3.9
ViTB16 → EffFormerv2	0	*99.6*	91.9	93.3	94.0	95.7	96.4	± 0.8
	50	*99.6*	72.7	73.3	76.8	73.7	92.1	± 1.5
	100	*99.6*	24.0	24.3	28.1	28.5	58.3	± 2.3

distinguishes itself by the transformer-based architecture of these models. For each distillation experiment, we use as a teacher the corresponding ImageNet pretrained models from TorchVision [27].

We consider in our benchmark five distillation methods ranging from the classical 'output-only' method to more recent methods with enhanced student-teacher alignment, such as SubDistill. The first method, '*output only*' [14], induces the student and teacher to predict similarly through minimizing a KL divergence loss placed on their output probability vectors. The second method, '*Attention Transfer (AT)*' [21], provides finer guidance by ensuring a match between the student and teacher attention maps at each layer. '*Variational Information Distillation (VID)*' [19] looks at layer-wise representation more comprehensively and defines the layer-wise loss in terms of maximizing the mutual information between the teacher's and the student's representations. *VKD* [20] is a fairly recent approach that uses task-dependent normalization and orthogonal adapters to improve alignment between the two representations. Finally, *SubDistill* [5] is a recently proposed layer-wise distillation method specifically designed for subtask distillation, which we have described in Section II-A.

For all four evaluated layer-wise distillation approaches, we employ the corresponding layer-wise loss at four different teacher-student layers, as in [5]. To make hyperparameter search feasible, we tie the layer-wise loss weighting coefficients λ_l to a single parameter λ and grid-search over $\lambda \in \{0.01, 0.1, 1.0, 10, 100\}$ (covering the range used in [23], [19], [20]). We select λ for each method using a validation set constructed with the same contamination statistics as for the test set, similar to the setups in [28], [29], [30]. Specifically, the validation set consists of 20% instances held out from the original training data, and the MNIST digits we use to contaminate images are selected randomly (see [31], [32]). This can be interpreted as an 'oracle' selection criterion, and each method in our benchmark benefits from it. We train student models for 100 epochs using AdamW [33] and set the initial learning rate equal to 0.001 with a decay of 0.5 at every 25 epochs. We employ early-stopping via the accuracy on the validation set and repeat the training for each configuration for three runs (different initializations of neural network parameters and validation-set splits).

A. Quantitative Results

The performance of each tested distillation method for each distillation setup is presented in Table I. Across all considered teacher-student pairs and levels of spurious correlation, we observe the expected trend that prediction performance degrades in the presence of increased levels of spurious correlation in the training data. Another noticeable trend is that advanced methods based on layer-wise distillation, and especially the recent SubDistill method, retain higher stability under high levels of spurious correlation compared to a basic distillation scheme that only considers the model outputs. Among layer-wise distillation methods, it is noteworthy that SubDistill consistently retains accuracies above 90% even when half of the data is contaminated with spuriously correlated MNIST digits, whereas other layer-wise distillation approaches lose up to 20 percentage points. Lastly, we observe that distillation performance degrades less for the simple ResNet-18 setup compared to more complex approaches based on WideResNet/MobileNet or transformers, where some methods degrade to accuracy slightly above random guess (20%) at the worst-case 100% contamination rate.

Overall, the results presented here strongly suggest that spurious correlations need to be actively addressed through (1) data selection or generation techniques that prevent spurious correlations from occurring, (2) design of distillation techniques that enforce a tight student-teacher alignment, and (3) selection of teacher and student models for which distillation techniques are maximally effective.

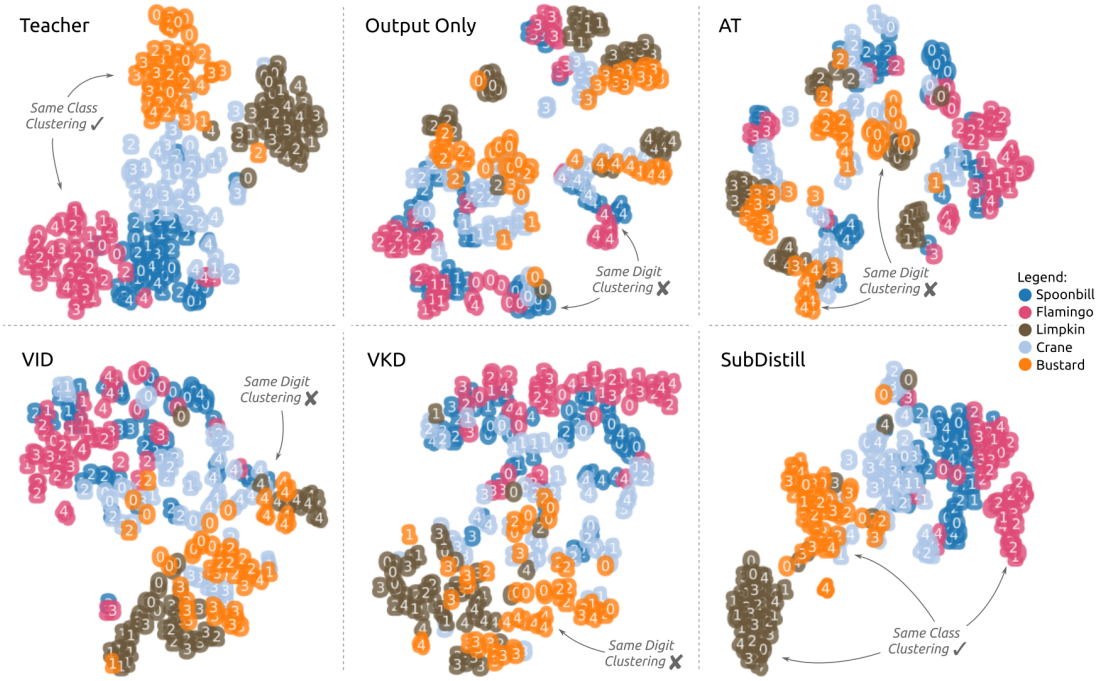

Fig. 2. Data representations produced by the ResNet18 teacher and the ResNet18-S students trained with different distillation methods under the presence of spurious correlation (100% level), and visualized using t-SNE. The color of the marker corresponds to the true class label (i.e. the type of wading bird present in the image), while the annotated digit corresponds to the type of artifact (superposed MNIST digits 0 to 4). The representations are taken from the best run of each method from the output of the global average pooling layer.

B. Comparing Student Representations with T-SNE

In this section, we aim to further corroborate our findings on the performance gap observed between distillation methods by analyzing the student models' hidden representations. Focusing on the ResNet18 pair, we investigate how semantics (of classes and spurious correlation) are encoded in the student's representation. For this purpose, we employ t-SNE [34] (an established non-linear dimensionality reduction algorithm) to visualize the representations of the student extracted after the global pooling layer.

Fig. 2 shows the t-SNE visualizations of the representations from the teacher and the students trained (trained with the 100% contamination rate, i.e., maximum spurious correlation). In these visualizations, each data point is colored according to its ImageNet class, and annotated according to the class of the MNIST digit that has been added synthetically. For the teacher, as expected, we see that its representation consists of clear clusters of colors, indicating that the class semantics, rather than the MNIST artifact, drive the similarity.

In contrast, the representations of the baseline students tend to form clusters that primarily encode the artifact (the MNIST digit) rather than the actual image class. This is clearly visible in the representation of the 'output only' student and also visible (although to a lesser degree) for the other baseline approaches. The distortion here clearly illustrates that the spurious correlation severely affects the learning of the student, driving it to learn the semantics of the MNIST digits rather than the actual image content. This is, however, not the case for the SubDistill student, which produces a representation that distinctly resembles that of the teacher, i.e. with clusters that code for image content rather than the MNIST artifact. Overall, these results highlight the ability of SubDistill to ensure tight student-teacher representational alignment even in the presence of strong spurious correlations in the training data.

C. Comparing Student's Decision Strategies with XAI

As a further qualitative comparison of the behavior of different distillation methods under spurious correlation, we use Explainable AI (XAI) [35], [36] to identify visual patterns used by each student model for predicting. Because our analysis is mainly aimed at determining whether the students focus on the signal (the image) or the top-left corner MNIST artifact, we find it sufficient and robust to apply an occlusion-based method [37], [38] with an appropriate patch size and patch replacement scheme. We view each input image as a 4×4 grid of patches of size 56×56, and for each prediction, we estimate the effect of removing each patch on the model output. Specifically, denoting by x the input image, x_p the pth patch, and $x_{\setminus p}$ the original image stripped from its pth patch, and by f_S the student model, we can measure the relevance of the pth patch via the simple formula:

$$R_p(x) = f_S(x) - f_S(x_{\setminus p}).$$

To generate the instances $x_{\setminus p}$, we make a distinction between the top-left patch containing the MNIST digit and the remaining patches forming the actual image. For the MNIST

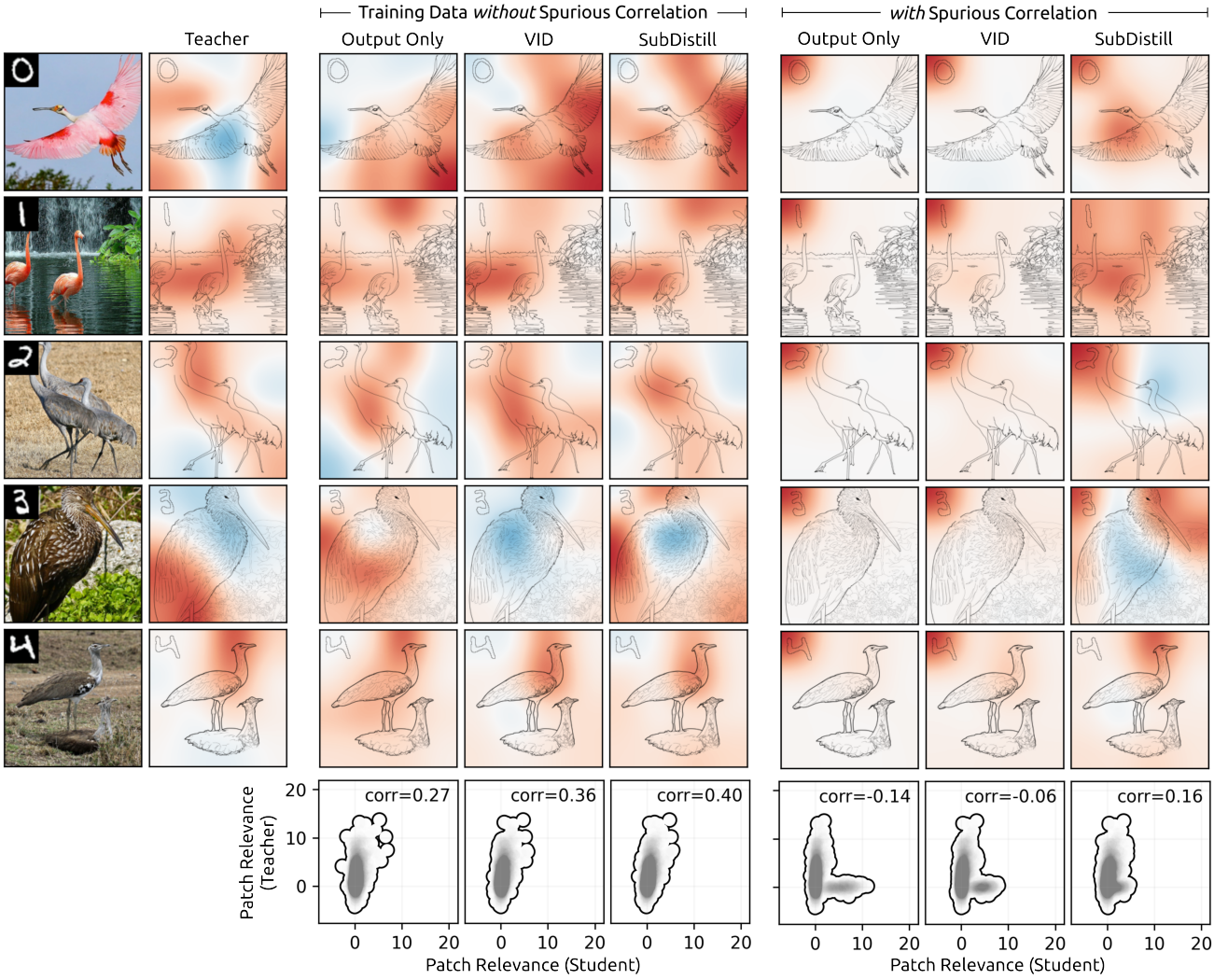

Fig. 3. Top: Explainable AI analysis of the teacher and students trained with the Output Only, VID, and SubDistill approaches using the 0%-MNIST-contaminated training data (no spurious correlation) and the 100%-MNIST-contaminated training data (spurious correlation). Red highlighting indicates visual features that contribute positively to the prediction, and blue highlighting indicates features that contribute negatively. Bottom: Scatter plots between each patch's relevance obtained from the teacher and the student at the level of 56×56 patches. The relevance scores are computed from all MNIST-contaminated test samples and averaged across the three training runs for each student. The depicted 'corr' is the Pearson correlation coefficient.

patch, we replace it with another MNIST digit of a different class. For the other patches, we simply set them to ImageNet's mean RGB. The collection of the $\{R_p(x)\}$ constitutes the explanation and can be visualized as a heatmap of the same size as the input image.

Fig. 3 (top) presents the results of the XAI analysis for cases in which students are trained with MNIST contamination rates of 0% (clean) and 100%. For the 0% case (no spurious correlation), we see that the heatmaps of VID and SubDistill students tend to be more similar to the teacher's than to the output-only student's. In contrast, for the 100% case (maximum spurious correlation), we observe stronger qualitative differences, with only SubDistill showing significant robustness and ability to concentrate on the bird features, whereas students resulting from 'output only' and VID primarily focus on the top-left corner (i.e. the artifact). The same effect is shown quantitatively in Fig. 3 (bottom), where we compare the patch-wise relevance of the students and the teacher across a representative set of contaminated images, presented as a scatter plot. We observe that training with spurious correlation leads to a lower correlation between the teacher's and the students' relevances, which, in the worst case, becomes negative. In this analysis, SubDistill again fares better than other approaches, and we can observe for the spurious-correlation case that it is the only method that produces decision strategies that positively correlate with those of the teacher.

IV. Discussion and Conclusion

In this work, we have investigated the emerging subtask distillation paradigm, especially whether it lends itself well to settings where the data available for training exhibits spurious correlations.

Through systematic experiments across several model pairs, including convolutional and transformer architectures, we have demonstrated wide performance gaps between simple distillation baselines, which degrade to nearly random on test data, and more advanced distillation methods, specifically SubDistill, whose performance remains high. Our experiments were complemented by t-SNE and Explainable AI analyses, allowing us to link the substantial performance gap between methods to preferential encoding of signal or artifact in internal representation and to the individual models' prediction strategies.

Our work has several limitations. First, our investigation is limited in scope to the image modality and to artificially introduced spurious correlations. While such a semi-artificial setting provides fine control on the spurious correlation strength, we consider it as future work to extend the investigation to other datasets and modalities with natural forms of spurious correlations. As a further limitation, our benchmark so far focuses on existing distillation methods that are not explicitly designed to mitigate spurious correlations. We anticipate that extending distillation techniques with, e.g. user-in-the-loop mechanisms [39], [40], could lead to further robustness improvements in these challenging data settings.

Overall, our work highlights significant differences in the robustness of existing distillation methods when trained on data with spurious correlations. Ensuring a tight student-teacher alignment in terms of both prediction and overall prediction strategy for the subtask of interest appears crucial in order to limit the potential negative impact of unrepresentative, spuriously correlated training data on the learned student model.

ACKNOWLEDGEMENT

This work was in part supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under Grants BIFOLD24B, BIFOLD25B, 01IS18037A, 01IS18025A, and 01IS24087C. P.C. is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. K.R.M. was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation). We thank Ali Hashemi for helpful and fruitful discussions.

REFERENCES

[1] R. Bommasani and et al., "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.

[3] C. Liang, S. Zuo, Q. Zhang, P. He, W. Chen, and T. Zhao, "Less is more: Task-aware layer-wise distillation for language model compression," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 20 852–20 867.

[4] L. Shi, Z. Shi, and J. Yan, "SelKD: Selective knowledge distillation via optimal transport perspective," in *ICLR*. OpenReview.net, 2025.

[5] P. Chormai, A. Hashemi, K.-R. Müller, and G. Montavon, "Distilling lightweight domain experts from large ML models by identifying relevant subspaces," *CoRR*, vol. abs/2601.05913, 2026.

[6] C. S. Calude and G. Longo, "The deluge of spurious correlations in big data," *Foundations of Science*, vol. 22, no. 3, pp. 595–612, 2016.

[7] J. Zeng, M. F. Gensheimer, D. L. Rubin, S. Athey, and R. D. Shachter, "Uncovering interpretable potential confounders in electronic medical records," *Nature Communications*, vol. 13, no. 1, 2022.

[8] A. Vaidya, R. J. Chen, D. F. K. Williamson, A. H. Song, G. Jaume, Y. Yang, T. Hartvigsen, E. C. Dyer, M. Y. Lu, J. Lipkova, M. Shaban, T. Y. Chen, and F. Mahmood, "Demographic bias in misdiagnosis by computational pathology models," *Nature Medicine*, vol. 30, no. 4, p. 1174–1190, 2024.

[9] J. Kömen, E. D. de Jong, J. Hense, H. Marienwald, J. Dippel, P. Naumann, E. Marcus, L. Ruff, M. Alber, J. Teuwen, F. Klauschen, and K.-R. Müller, "Towards robust foundation models for digital pathology," *CoRR*, vol. abs/2507.17845, 2025.

[10] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1096, 2019.

[11] R. Geirhos, J. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.

[12] P. R. A. S. Bassi, A. Cavalli, and S. Decherchi, "Explanation is all you need in distillation: Mitigating bias and shortcut learning," *CoRR*, vol. abs/2407.09788, 2024.

[13] A. Parchami-Araghi, M. Böhle, S. Rao, and B. Schiele, "Good teachers explain: Explanation-enhanced knowledge distillation," in *ECCV (73)*, ser. Lecture Notes in Computer Science, vol. 15131. Springer, 2024, pp. 293–310.

[14] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[15] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *NIPS*, 2014, pp. 2654–2662.

[16] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling BERT for natural language understanding," in *EMNLP (Findings)*, ser. Findings of ACL, vol. EMNLP 2020. Association for Computational Linguistics, 2020, pp. 4163–4174.

[17] Y. Gu, L. Dong, F. Wei, and M. Huang, "MiniLLM: Knowledge distillation of large language models," in *ICLR*. OpenReview.net, 2024.

[18] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *3rd International Conference on Learning Representations, ICLR, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[19] S. Ahn, S. X. Hu, A. C. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 9163–9171.

[20] R. Miles, I. Elezi, and J. Deng, "VkD: Improving knowledge distillation using orthogonal projections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 15 720–15 730.

[21] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *ICLR (Poster)*. OpenReview.net, 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE Computer Society, 2016, pp. 770–778.

[23] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *BMVC*. BMVA Press, 2016.

[24] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, F. Yang, W. Wang, C. R. Banbury, C. Ye, B. Akin, V. Aggarwal, T. Zhu, D. Moro, and A. G. Howard, "MobileNetV4: Universal models for the mobile ecosystem," in *ECCV (40)*, ser. Lecture Notes in Computer Science, vol. 15098. Springer, 2024, pp. 78–96.

[25] 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 *ICLR*. OpenReview.net, 2021.

[26] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "Efficientformer: Vision transformers at mobilenet speed," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 934–12 949, 2022.
[27] TorchVision maintainers and contributors, "Torchvision: Pytorch's computer vision library," https://github.com/pytorch/vision, 2016.
[28] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *ICLR*. OpenReview.net, 2021.
[29] M. Pezeshki, O. Kaba, Y. Bengio, A. C. Courville, D. Precup, and G. Lajoie, "Gradient starvation: A learning proclivity in neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1256–1272, 2021.
[30] N. Ye, K. Li, H. Bai, R. Yu, L. Hong, F. Zhou, Z. Li, and J. Zhu, "Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization," in *CVPR*. IEEE, 2022, pp. 7937–7948.
[31] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *CoRR*, vol. abs/1907.02893, 2019.
[32] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *ICLR*. OpenReview.net, 2020.
[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR (Poster)*. OpenReview.net, 2019.
[34] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
[35] D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019.
[36] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 2021.
[37] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV (1)*, ser. Lecture Notes in Computer Science, vol. 8689. Springer, 2014, pp. 818–833.
[38] I. Covert, S. M. Lundberg, and S. Lee, "Explaining by removing: A unified framework for model explanation," *J. Mach. Learn. Res.*, vol. 22, pp. 209:1–209:90, 2021.
[39] C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Finding and removing Clever Hans: Using explanation methods to debug and improve deep models," *Inf. Fusion*, vol. 77, pp. 261–295, 2022.
[40] S. Bender, C. J. Anders, P. Chormai, H. Marxfeld, J. Herrmann, and G. Montavon, "Towards fixing Clever-Hans predictors with counterfactual knowledge distillation," in *ICCV (Workshops)*. IEEE, 2023, pp. 2599–2607.