Knowledge Distillation with Training Dynamics

Timo Hromádka
Computer Science Laboratory
University of Cambridge
th716@cam.ac.uk

Evan Rex
Computer Science Laboratory
University of Cambridge
er647@cam.ac.uk

Abstract

In this study, we examine the efficacy of Knowledge Distillation in conjunction with Dataset Cartography within Computer Vision. Our aim was to ascertain whether training on subsets of data, selected based on training dynamics, can enhance the performance of student models while being cost-effective. Contrary to expectations set by similar NLP studies, our experiments across various architectures and datasets (CIFAR10, CIFAR100, MNIST) revealed that randomly selected data subsets consistently outperformed those chosen based on training dynamics. This unexpected outcome suggests for the task of dataset selection in Knowledge Distillation, particularly in the Computer Vision domain, a training set that is unconditionally representative of the larger dataset is better than a training dynamic-informed curriculum. Our findings highlight the complex interplay between data selection and model performance, opening new avenues for research in effective training strategies. We publish our code for future research.

1 Introduction

Large amounts of data have been pivotal to the development of high-performing AI models across numerous domains, with no exception in Computer Vision [16, 1]. A common consensus in modern deep learning is that the larger the model and more data there is to train on, the better the performance will be. While this is generally true, overparameterized models are prone to overfitting of minority labels [14] and are becoming more and more expensive to train and use for inference.

In light of growing costs of training, studies have investigated the importance of individual data points for a model's learning process, and it has been shown that not all data points contribute equally [8], and data can be selectively chosen rather than always training on the full dataset. In light of this, data selection techniques use such as active learning [21, 17, 24] or curriculum learning [18] have been extensively studied. One effective approach, fitting most closely under curriculum learning, is the concept of *Dataset Cartography* [20], a particularly attractive approach owing to its simplicity and consistency. Dataset Cartography 'maps' data points according to their training dynamic statistics, which are then used to select subsets of data. Correctly selecting the right subset of data can match or even exceed performance of training on the entire dataset. To the best of our knowledge, we are the first to apply this framework in Computer Vision.

To mitigate high costs of model inference while retaining performance, *Knowledge Distillation* [6] is an effective technique to train a smaller 'student' model to achieve high performance guided by the predictions of a stronger 'teacher' model [4]. We propose using dataset cartography with knowledge Ddstillation for teacher model training. By analyzing training dynamics statistics of the data, we aim to identify a data subset that enhances the student model's performance and lowers computational costs.

MPhil ACS L46 Principles of Machine Learning Systems Project (2023/24).

1.1 Aims

We aim to see if training dynamic-informed subsets are capable of producing similar, if not better, performance of a student model trained with knowledge distillation. If not, we aim to understand the properties of data points which are most beneficial to be used in a knowledge distillation framework. Our aims are to develop a framework for knowledge distillation with dataset cartography. Then, we aim to evaluate performance of different subset selection approaches whilst identifying the most effective subset selection approaches. Lastly, we aim to explore the limits/nature of the best subset selection approach.

1.2 Contributions

- We demonstrate the ability of faithfully calculating training dynamics for computer vision datasets w.r.t. a model.
- Selecting training dynamic-informed subsets did not typically yield competitive performance, and we discuss this result.
- We identify random subset selection as a better subset selection approach than training dynamic-informed subset selection.
- We demonstrate that we can achieve competitive student model performance using only a fraction of the full dataset.

2 Background

2.1 Knowledge Distillation

2.1.1 Overview

Knowledge Distillation, as introduced in Hinton et al.'s work [6], is a process whereby a smaller "student" model assimilates knowledge from a larger "teacher" model. This approach enables the student to achieve performance close to that of the teacher while using significantly less compute. The key advantage to knowledge distillation lies in leveraging the teacher's probabilistic output over labels, which offers a richer learning source than one-hot-encoded labels. This method captures the teacher's nuanced understanding of data including the teacher's uncertainties, allowing the student to generalize better. This technique often results in more accurate predictions by the student model.

2.1.2 Modified Training Objective

In the context of neural network training, conventional models are typically trained using a cross-entropy loss function, which can be represented as:

$$L_{CE}(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i) \tag{1}$$

where y denotes the true label, and $\hat{y}$ represents the model's prediction.

However, the training objective in knowledge distillation diverges from this traditional approach by incorporating the soft outputs of a teacher model. This modified loss function combines the standard cross-entropy loss for the true labels with the cross-entropy loss computed from the teacher model's predictions:

$$L_{KD}(y, \hat{y}, y^T) = \alpha L_{CE}(y, \hat{y}) + (1 - \alpha) L_{CE}(y^T, \hat{y}) \tag{2}$$

In this equation, y^T denotes the teacher model's prediction, and α is a hyperparameter that balances the contribution of the two loss components.

To integrate the temperature parameter in knowledge distillation, the softmax function used in the teacher's predictions is adapted with a temperature parameter T. The softmax function with temperature is defined as:

$$\text{softmax}(z_i, T) = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}) \tag{3}$$

This temperature-modified softmax function produces a softer probability distribution over classes when a higher temperature value is used, providing more nuanced information for the student model. The knowledge distillation objective with temperature adjustment is then given by:

$$L_{KD-T}(y, \hat{y}, y^T, T) = \alpha L_{CE}(y, \hat{y}) + (1 - \alpha)T^2 L_{CE}(\text{softmax}(y^T/T), \text{softmax}(\hat{y}/T))) \quad (4)$$

Note that the factor T^2 is recommended to be employed in practice to appropriately scale the gradients [6].

2.2 Dataset Cartography

Swabha Swayamdipta et. al [20] proposed a novel data selection approach based on calculating **training dynamics** for each data point. This approach involves choosing subsets of the training set according to the characteristics of each data point during training. The method has demonstrated that selecting data points with specific properties can achieve or even surpass the performance of using the entire dataset.

Similar studies have realized the varying importance of different data points. Curriculum learning [18], a method inspired by the way humans learn new skills or knowledge, involves structuring the learning process to start with easier examples or tasks and gradually increase the difficulty level, as often seen in self-paced learning [13, 10]. Analogously, temperature curriculum learning has been explored for computer vision, where the temperature (From equations 3 and 4) is adjusted during training for optimal performance. Active learning [21, 17, 24], on the other hand, is a technique in machine learning where the algorithm/model selects which data points to be labelled by a user. Thus, the model can focus on the most informative or challenging examples, with recent work suggesting 'forgotten' data points to be pivotal for active learning [11].

However, both active learning and curriculum learning are highly variable, and often have difficulty adapting to the subtleties of each task and model [25, 15]. Instead, training dynamics have proven to be a dataset-agnostic approach to accurately select subsets of data while retaining most, if not all, performance. Selecting data instances according to their training dynamics has been shown to be highly effective and a great way of reducing computational costs while maintaining equal performance on both in and out of distribution training [20]. Recently, additional works have shown that these training dynamic are transferable across model sizes [3], and can be effectively utilized as part of curriculum learning for training transformers [26].

2.3 Data Selection with Knowledge Distillation

Data selection using training dynamics has not been explored in computer vision, rather only in Natural Language Processing (NLP) tasks utilizing large language models. To this end, we are the first to explore training dynamic calculation with computer vision datasets and models.

Furthermore, we are the first to investigate the selection of data subsets for knowledge distillation guided by training dynamics. Our unique contribution is a framework to knowledge distillation by employing training dynamics for data selection. These dynamics have the potential to be instrumental in identifying the most appropriate subset of data for training, ensuring that we make the most of the compute already in use. In particular, [20] show that training dynamics-informed subset selection can yield improved models. As such we see potential for this technique to improve the quality of student models in a knowledge distillation framework. We further identify that this approach is highly efficient in the context of a knowledge distillation framework, as we are able to generate training dynamics during the teacher model training. Thus there is less overhead to the calculation of training dynamics in the knowledge distillation framework than in typical model training approaches.

2.3.1 Training Dynamics

Mapping Datamaps The premise of Datamaps is to visualize a dataset, with respect to a model, with a particular focus on understanding the role individual data points play towards a model's learning of a task/dataset. To achieve this, *training dynamics* are calculated for each data point during the course of a model's training, and subsequently each datapoint can be 'mapped' according to the training dynamics. We provide example datamaps in Figure 1 and 2.

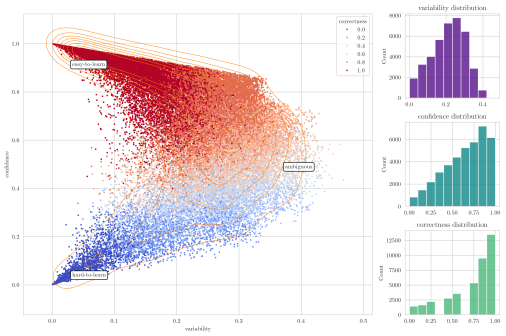


Figure 1: Datamap of Resnet Medium model on the Cifar10 dataset.

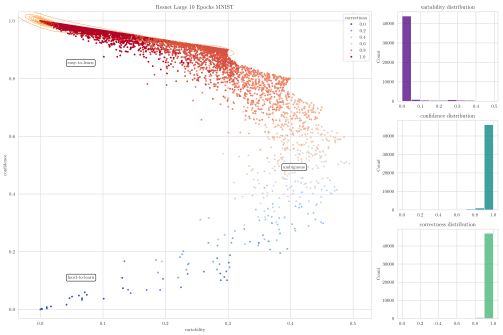


Figure 2: Datamap of Resnet Large model on the MNIST dataset.

Requirements We assume a dataset D of size N, $\mathcal{D} = \{(x_i, y_i^*)\}_{i=1}^N$ with x_i as the ith datapoint and y_i^* its true label. Model M, with parameters θ, is trained via Empirical Risk Minimization (ERM) to minimize Cross-Entropy loss, assuming M defines a probability distribution over labels given an observation. Parameters θ are learned through stochastic gradient descent, with training instances randomly ordered at each of the E epochs.

Following the procedure from the Cartography paper [20], we calculate three training dynamics across the experiments: **confidence**, **variability**, and **correctness**. In addition, we introduce a fourth training dynamic, **forgetfulness**, inspired by research highlighting the prevalence of 'forgotten' data points [23]. Each of the four training dynamics are calculated for each datapoint. In summary, the constructed datamaps reveal three distinct regions where datapoints typically lie in.

- **hard-to-learn** section (low confidence, low variability). This region comprises data points which the model consistently guesses very wrongly.
- **ambiguous** section (medium confidence, high variability). This region comprises data points which the model is unsure of, and highly oscillates in its prediction probabilities for the gold label.
- **easy-to-learn** section (high confidence, low variability). This region comprises data points which the model consistently guesses correctly, with high confidence as well.

For the full formulae, the reader may refer to the Appendix A.

3 Methodology

3.1 Implementation

We use PyTorch Lightning[1] with WandB[2] logging to conduct our experiments. Instructions for experiment reproduction are provided in the GitHub repository, where our code implementation of this research is freely available: `https://github.com/timohromadka/L46_datamaps/tree/main`. We implement our experimentation pipeline from scratch following the procedure outlined in the original Dataset Cartography paper [20]. For each experimental phase outlined in Section3.3, we construct a corresponding WandB sweep which iterates over the variable combinations for that experiment.

Each experiment is repeated multiple times with a different random seed, and the results are presented in aggregate. The random seeds are stored in the GitHub repository to enable repeatability.

3.2 Summary of Architectures and Datasets

Architectures So as to ensure the generalize-ability of our results across multiple architectures, we conduct experiments using 3 distinct computer vision architectures: the ResNet, EfficientNet

[1] `https://lightning.ai/`
[2] `https://wandb.ai/`

and MobileNet. In the realm of computer vision, these three architectures have gained widespread use due to their efficacy. The Residual Network (ResNet), introduced by He et al., revolutionized deep learning by enabling the training of substantially deeper networks through skip connections [5]. EfficientNet, proposed by Tan and Le, offered a systematic approach to scaling network width, depth, and resolution, setting new records for accuracy and efficiency [22]. MobileNet, developed by Howard et al., is tailored for mobile and resource-constrained environments, emphasizing efficient convolutional neural networks [7]. In particular, we explore larger and smaller versions of these models as teachers and students respectively. These models and their associated parameter counts are presented in Table 1.

Teacher Model	**# params**	**Student Model**	**# params**
`mobilenet_v2(width_mult=1.0)`	2.2M	`mobilenet_v2(width_mult=0.35)`	409K
`resnet50`	25.6M	`resnet18`	11.5M
`efficientnet-b4`	19M	`efficientnet-b0`	5.3M

Table 1: Architectures and parameter counts

Datasets In the field of machine learning, particularly in image recognition tasks, three datasets have become widely recognised benchmarks: MNIST, CIFAR-10, and CIFAR-100. The MNIST dataset, a collection of handwritten digits, is known for its simplicity and is often used as an entry point for algorithm testing [12]. CIFAR-10 and CIFAR-100, introduced by Krizhevsky et al., consist of 60,000 color images in 10 and 100 classes, respectively, providing a more complex challenge in object recognition [9]. These datasets are widely accepted and utilized in the research community, serving as standard benchmarks for evaluating the performance of image recognition algorithms. We employ all three datasets in our experiments to ensure comprehensive testing and comparability with existing research.

3.3 Experiments and Results

Across all of our experiments we adopt a very simple approach. The design of this framework is necessitated by the nature of our project aims:

- Train a large teacher model on the training set.
- After each epoch of training, calculate the four training dynamics for each data point.
- Select sub-regions of the data according to training dynamics statistics.
- Perform knowledge distillation using the trained teacher model and a smaller student model on the chosen data subset.

Due to computational restrictions, we omit a hyper-parameter search and set the temperature $T = 1$ and combination coefficient $\alpha = 0.5$ for Equation 4.

Our research process consists of 3 phases of experiments. This iterative and incremental research process follows the process outlined in [20].

- **Phase 1** Explore and compare different approaches to selecting a 33% subset of the training on the CIFAR-10 dataset. This 33% starting point is informed by the results of the [20], who show that 33% split is suitable for differentiating between various regions of the datamap, such as the hard-to-learn, ambiguous, and easy-to-learn regions.
- **Phase 2** Informed by phase 1 results, conduct a deeper exploration and comparing different approaches to selecting a 33% subset of the training on all 3 datasets.
- **Phase 3** Informed by phase 2 results, focus on a specific subset selection criteria and ablate the selected dataset size so as to potentially identify a subset size that effectively balances computation cost and classification performance.

3.3.1 Teacher Training with Training Dynamics

The first step of our experimental procedure involved the training of a teacher model for each of the explored datasets. We performed a grid search across multiple architectures and learning rates

so as to obtain the best model to use as our teacher. We have found that using models pre-trained on ImageNet [1] achieved the strongest performance. Across all of our experiments, we report the macro-F1 score[3]. To select our teacher model, we chose the model with the highest macro-F1 score on the validation set. These models are presented in Table 2.

Dataset	Model	Model Size	F1	Learning Rate
cifar10	resnet	large	82.9	0.0001
cifar100	resnet	large	57.4	0.0001
mnist	resnet	large	98.8	0.001

Table 2: Chosen Teacher Models

3.3.2 Student Training with Teacher Subset

The next step of our experimental process involved the training of student models using knowledge distillation. Calculating training dynamics of data points during the teacher's training provides us the ability to select data points according to a the training dynamics[4] outlined in Section 2.3.1.

For the following experiment phases, a variety of students were trained, each with a different subset of the training set. These subsets were chosen according to the datapoints' ranking with respect to one of the above dataset cartography metrics. For each student architecture, we additionally trained two baseline models: a model trained on the full training set, and a model trained on a randomly selected 33% of the training set.

Phase 1: 0.33 Subset Experiments In phase 1 of our experiments, for each dataset we constructed 12 subsets, each being 33% of the training set. Here, we explore whether there are subsets of the data capable of producing the same, or potentially better, performance for the student model. Due to computational constraints, we restrict this phase of the research to the CIFAR10 dataset, a staple benchmarking image classification dataset.

For each experiment we average our results over three random seed initializations. These results are presented in Table 3.

Train Data	CIFAR10		
	MobileNet	*EfficientNet*	*ResNet*
100% train	0.6164	0.817	0.774
33% random	0.469	**0.791**	**0.681**
33% low-forgetfulness	0.4363	0.709	0.604
33% high-forgetfulness	0.223	0.706	0.504
33% low-correctness	0.2999	0.767	0.603
33% high-correctness	**0.4955**	0.778	0.676
33% low-variability	0.453	0.702	0.597
33% high-variability	0.2099	0.734	0.537
33% low-confidence	0.1824	0.726	0.522
33% high-confidence	0.4349	0.701	0.593
33% ambiguous	0.4427	0.704	0.593
33% easy-to-learn	0.2084	0.709	0.522
33% hard-to-learn	0.1818	0.738	0.491

Table 3: Avg. Test-F1 scores for all architectures on all CIFAR10 in Phase 1 experiments

Our experiments indicate that selecting training subsets based on dynamic-informed criteria typically underperforms compared to random selection. Notably, subsets with high-correctness data points demonstrate comparable accuracy to random subsets. This observation aligns with the findings in [19, 2], suggesting the importance of soft labels for student training. Random selection outperforms the other subsets because of this reason, as it is able to provide the required soft labels, rather than being restricted to picking out 'edge' case datapoints of the teacher. In contrast, the 'easy-to-learn' subset, comprising the simplest data points where teacher labels closely match hard labels, limits the student's learning to superficial information and omits exposure to more challenging data. On the

[3]This metric is especially suitable given the balanced label distribution across all of our datasets.

[4]All training dynamics are calculated across all training epochs

other hand, the 'hard-to-learn' subset consists of data points where even the teacher model struggles, resembling training on potentially noisy or mislabeled data, which inhibits the student's ability to derive accurate knowledge. Similarly to the random selection approach, the 'high-correctness' subset offers a broader data spectrum compared to the 'easy-to-learn' subset, thus better facilitating knowledge distillation, while still providing correct and low-noise labels. However, the spread is still not as large as random, and thus doesn't give as generalizable of a label distribution for the student (For a visual representation of these subsets, the reader may visit the appendix section C).

Phase 2: Combinations of 0.33 In Phase 2, we conduct further experiments in accordance to the experimental protocol of the Dataset Cartography paper [20], where we group subsets of data from different region of the datamap. We aim to discover if we are able to mimic a well-informed and generalizable subset, matching the performance of the 33% random subset.

We construct multiple subsets from multiple training dynamic-informed subsets, each with 33% of the full training set size. For each experiment we average results over 3 random seed initializations, and display the results in Table 4.

Train Data			**CIFAR10**			**CIFAR100**			**MNIST**	
			MobileNet	*EfficientNet*	*ResNet*	*MobileNet*	*EfficientNet*	*ResNet*	*MobileNet*	*ResNet*
100% train			0.616	0.817	0.774	0.188	0.547	0.441	0.984	0.987
33% Random			**0.469**	**0.791**	**0.681**	0.081	**0.447**	**0.447**	0.966	0.984
hard	amb.	easy								
16.5%	16.5%	0%	0.202	0.749	0.522	0.025	0.269	0.140	**0.977**	0.988
16.5%	0%	16.5%	0.432	0.720	0.626	**0.106**	0.383	0.280	0.972	0.990
0%	16.5%	16.5%	0.417	0.728	0.630	0.084	0.405	0.275	0.974	0.989
14%	14%	5%	0.333	0.739	0.589	0.056	0.344	0.22	0.979	0.987
14%	5%	14%	0.391	0.721	0.618	0.101	0.388	0.268	0.971	0.986
5%	14%	14%	0.409	0.729	0.616	0.082	0.383	0.267	0.970	0.984
11%	11%	11%	0.418	0.750	0.611	0.086	0.371	0.256	0.976	**0.989**

Table 4: Avg. Test-F1 scores for all architectures on all datasets in Phase 2 experiments

Once again, we find that the randomly selected subset yields the best performance across the majority of dataset-student architecture combinations. Deviations occur for the MNIST dataset, and the MobileNet model on the CIFAR100 dataset. For the MNIST dataset, we can observe in Figure 2 that almost all datapoints fall into the easy-to-learn category. Thus, when we select a percentage of datapoints according to different criteria, in reality we are selecting easy-to-learn datapoints. As such, all selection methods are equivalent to random sampling. It is unsurprising then that the random sampling approach is the median performing approach for this dataset. For the CIFAR100-MobileNet combination, we observe that the students' performance is extremely poor across all subset selection approaches when compared to other architectures. This indicates the architecture is too simple for this classification task. As such, the subset selection approach is likely not a major contributor to the model performance.

This phase again identifies the random sampling approach to be a superior subset selection method in the context of knowledge distillation.

Phase 3: Random Ablation Study In phase 3, we explore the effect of reducing the randomly selected training subset size on model performance. As we have already shown this to be the best subset selection method, we are now interested in further exploring the quality of this subset selection method across different subset sizes.

These results are presented in Table 5, and visualised in Figure 3. They illustrate a consistent positive correlation with the training dataset size, and the performance of the model. However, there are diminishing returns from increasing subset size, with minimal accuracy improvement above 50% of training set utilization.

4 Conclusion

The primary finding of our study is that random selection of a subset of the training data, accounting for approximately 33% of the full dataset, consistently outperforms subsets selected based on specific training dynamics. This observation was consistent across various architectures and datasets,

Train Data	CIFAR10			CIFAR100			MNIST	
	MobileNet	*EfficientNet*	*ResNet*	*MobileNet*	*EfficientNet*	*ResNet*	*MobileNet*	*ResNet*
1% Random	0.100017	0.423433	0.515250	0.0002846	0.09796	0.0382	0.1018	0.7807
5% Random	0.202000	0.595333	0.603650	0.008573	0.2411	0.1136	0.8205	0.9529
10% Random	0.337417	0.706683	0.603283				0.9038	0.9558
25% Random	0.438033	0.769783	0.646850	0.06882	0.4211	0.2767	0.9677	0.9809
33% Random	0.469183	0.790833	0.681117	0.08103	0.4468	0.3106	0.9655	0.9842
55% Random	0.544117	0.804733	0.730817	0.1229	0.493	0.3667	0.9748	0.987
100% Train	0.619600	0.819133	0.769900	0.1884	0.5467	0.4412	0.984	0.9874

Table 5: Avg. Test-F1 scores for all architectures on all datasets in Phase 3 experiments[5]

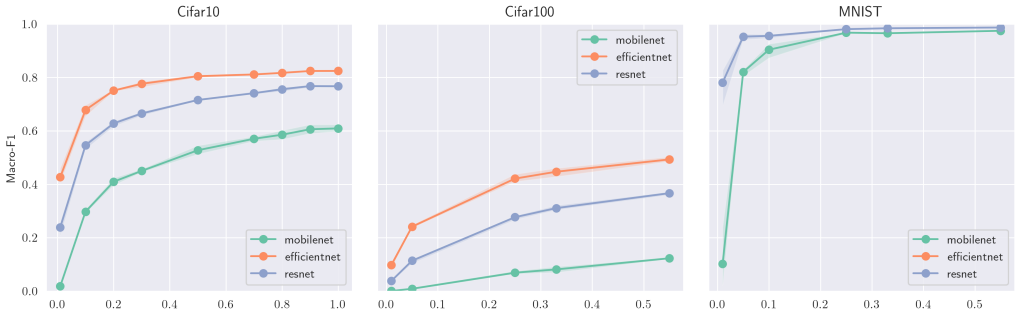


Figure 3: Ablation Experiments across three different models and three different datasets. Using even just a fraction of the original dataset can yield comparable performance to using the full training set.

including CIFAR10, CIFAR100, and MNIST. Drawing on insights from Stanton et al.'s [19] study on knowledge distillation, we deduce that the decline in performance is due to the lack of a generalizable subset offering information-rich labels to student models. Our phase 3 experiment shows that achieving student model generalizability requires only a fraction of the complete training dataset.

Our findings challenge the prevailing assumption that training on a carefully selected subset, based on detailed training dynamics, always leads to superior or equivalent model performance [20]. In the context of Knowledge Distillation in Computer Vision, this study highlights the importance of dataset selection, but also underscores the complexity and potential unpredictability of the relationship between data selection and model performance. It is argued that the strength of Knowledge Distillation lies not in forcing the student to emulate the teacher, but in enabling the student to better generalise [19]. It is possible that the less general subset obtained from selection conditioned by the teacher's training dynamics inhibits the students' generalisation ability, undermining one of the benefits of knowledge distillation.

This research opens avenues for further investigation into the role of training dynamics and dataset cartography in Computer Vision and other domains. Future research could explore more nuanced or hybrid approaches to dataset selection, combining random sampling with training dynamics insights. Such research will uncover properties under which knowledge distillation provides the highest degree of generalizability for the student model.

In conclusion, our study contributes to the ongoing dialogue in the field of machine learning, particularly in the areas of Knowledge Distillation and dataset optimization. While our results deviate from expectations set by previous works in NLP, they offer valuable insights and direction for future research, emphasizing the need for a nuanced and context-sensitive approach to dataset selection, and uncover the importance of rich labels in Knowledge Distillation within Computer Vision.

References

[1] ImageNet: A large-scale hierarchical image database | IEEE Conference Publication | IEEE Xplore.

[2] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[3] Yupei Du, Albert Gatt, and Dong Nguyen. FTFT: efficient and robust Fine-Tuning by transFerring Training dynamics, October 2023. arXiv:2310.06588 [cs].

[4] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6):1789–1819, June 2021. arXiv:2006.05525 [cs, stat].

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. arXiv:1503.02531 [cs, stat].

[7] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[8] Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher D. Manning. Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering, July 2021. arXiv:2107.02331 [cs].

[9] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[10] M. Kumar, Benjamin Packer, and Daphne Koller. Self-Paced Learning for Latent Variable Models. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[11] Beong-woo Kwak, Youngwook Kim, Yu Jin Kim, Seung-won Hwang, and Jinyoung Yeo. TrustAL: Trustworthy Active Learning using Knowledge Distillation, January 2022. arXiv:2201.11661 [cs].

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[13] Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1721–1728. IEEE Computer Society, 2011.

[14] Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. OVERPARAMETERISATION AND WORST-CASE GENER- ALISATION: FRIEND OR FOE? 2021.

[15] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A Survey of Deep Active Learning. *ACM Computing Surveys*, 54(9):180:1–180:40, October 2021.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, January 2015. arXiv:1409.0575 [cs].

[17] Burr Settles. Active Learning Literature Survey.

[18] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum Learning: A Survey, April 2022. arXiv:2101.10382 [cs].

[19] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in Neural Information Processing Systems*, 34:6906–6919, 2021.

[20] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, November 2020. Association for Computational Linguistics.

[21] Rinyoichi Takezoe, Xu Liu, Shunan Mao, Marco Tianyu Chen, Zhanpeng Feng, Shiliang Zhang, and Xiaoyu Wang. Deep Active Learning for Computer Vision: Past and Future. *APSIPA Transactions on Signal and Information Processing*, 12(1), 2023. arXiv:2211.14819 [cs].

[22] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[23] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An Empirical Study of Example Forgetting during Deep Neural Network Learning, November 2019. arXiv:1812.05159 [cs, stat].

[24] Mingfei Wu, Chen Li, and Zehuan Yao. Deep Active Learning for Computer Vision Tasks: Methodologies, Applications, and Challenges. *Applied Sciences*, 12(16):8103, January 2022. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.

[25] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. WHEN DO CURRICULA WORK? 2021.

[26] Osman Batur İnce, Tanin Zeraati, Semih Yagcioglu, Yadollah Yaghoobzadeh, Erkut Erdem, and Aykut Erdem. Harnessing Dataset Cartography for Improved Compositional Generalization in Transformers. 2023. Publisher: arXiv Version Number: 1.

A Calculating Training Dynamics

Confidence is defined as the mean probability the model M assigns to x_i's true label, across E epochs.

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^*|x_i)$$

where $p_{\theta^{(e)}}$ denotes model M's probability with parameters $\theta^{(e)}$ at the end of the e^{th} epoch.

Variability defines the spread of $p_{\theta^{(e)}}(y_i^*|x_i)$ across epochs as the standard deviation of the model's probability prediction of the true label.

$$\hat{\sigma}_i = \sqrt{\frac{1}{E}\sum_{e=1}^{E}\left(P(\theta_e)(y_i^* \mid x_i) - \hat{\mu}_i\right)^2}$$

Correctness defines the ratio of how often a model correctly predicts a data point's true label across *e* epochs.

Forgetfulness defines the model's tendency to 'forget' a data point. Formally, for each *i*th data point x_i, forgetfulness f is the slope of the least-squares regression line across the model's gold label probabilities of each epoch e for N epochs.

$$f_i = \frac{N\sum(e \cdot p_\theta^{(e)}(y_i^*)) - \sum e \sum p_\theta^{(e)}(y_i^*)}{N\sum e^2 - (\sum e)^2}$$

B Example Training Dynamics Calculation

To intuitively show how training dynamics are calculated, and what they are measuring, we demonstrate the calculations on one particular data point, a Cifar10 sample of a frog (Figure 4). Using a model capable of defining a probability distribution over labels, we use the softmaxed probability distribution (summing to 1) for calculation, as shown in Table **??**.

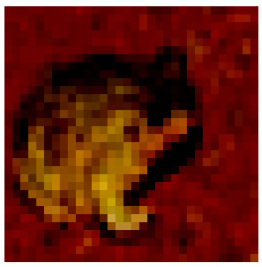


Figure 4: Index 4000, label 6 (**frog**).

Epoch	Labels									
	plane	*car*	*bird*	*cat*	*deer*	*dog*	***frog***	*horse*	*ship*	*truck*
1	0.001	0.001	0.004	0.002	0.027	0.043	**0.893**	0.025	0.003	0.001
2	0.001	0.001	0.004	0.002	0.019	0.033	**0.911**	0.025	0.003	0.001
3	0.080	0.131	0.057	0.056	0.095	0.103	**0.762**	0.074	0.093	0.092
4	0.003	0.003	0.004	0.009	0.036	**0.477**	0.438	0.011	0.013	0.011
5	0.005	0.003	0.005	0.011	0.030	**0.727**	0.188	0.011	0.013	0.012

Table 6: Probability Distribution of model's label probabilities over epochs for a Cifar-10 datapoint. The true label is **frog**. This particular model handled the sample quite well at the beginning but started to predict the image as 'dog' in later epochs.

Confidence (μ) To calculate confidence we simply average the model's gold label probabilities

$$\mu = \frac{0.893 + 0.911 + 0.762 + 0.438 + 0.188}{5} = \mathbf{0.638} \tag{5}$$

Variability (σ) To calculate variability we calculate the standard deviation of the model's gold label probabilities:

$$\sigma = \sqrt{\frac{(0.893-\mu)^2 + (0.911-\mu)^2 + (0.762-\mu)^2 + (0.438-\mu)^2 + (0.188-\mu)^2}{5}} = \mathbf{0.282} \tag{6}$$

Correctness (C) To calculate correctness we count the proportion of times the gold label was correctly predicted (i.e., the highest probability of all other classes)

$$C = \frac{3}{5} = \mathbf{0.6} \tag{7}$$

Forgetfulness To calculate forgetfulness we retrieve the slope of the least-squares regression linear fit line through the gold label probabilities. Colloquially, this is an indication of whether the model generally was learning or 'forgetting' the data point over the course of training. The forgetfulness value for this sample is **-0.188**, as shown in Figure **??**.

C Data Subset Visualization

The 'easy-to-learn' subset, focusing on basic data where teacher and hard labels align, limits learning to basic information and misses complex data. The 'high-correctness' subset, akin to random selection, covers a wider data range than the 'easy-to-learn' subset, enhancing knowledge transfer while ensuring correct, low-noise labels. However, its variety still falls short of random selection, leading to a less diverse label distribution for the student.

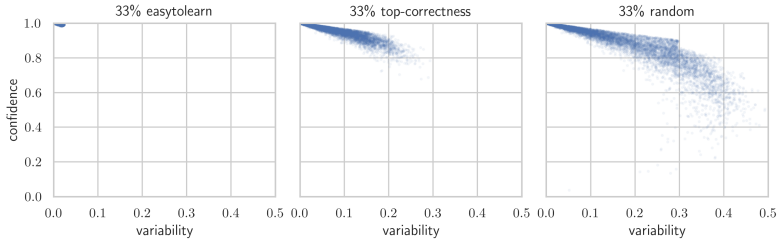


Figure 5: Enter Caption

D Datamap Regions Visualized

We present a few samples from each of the dataset's easy-to-learn, ambiguous, and hard-to-learn regions. The easy-to-learn samples are straightforward for the model to learn. The ambiguous data points typically are data points that the model struggles with at first, but learns to correctly classify it. Hard-to-learn samples the model is unable to learn during the entirety of its training. A qualitative analysis shows that these data points are often very difficult even for humans, and are potentially mislabeled (e.g. for Cifar-100, the bottom-right hard-to-learn image with two women is labelled as 'truck').

Figure 6: Cifar-10

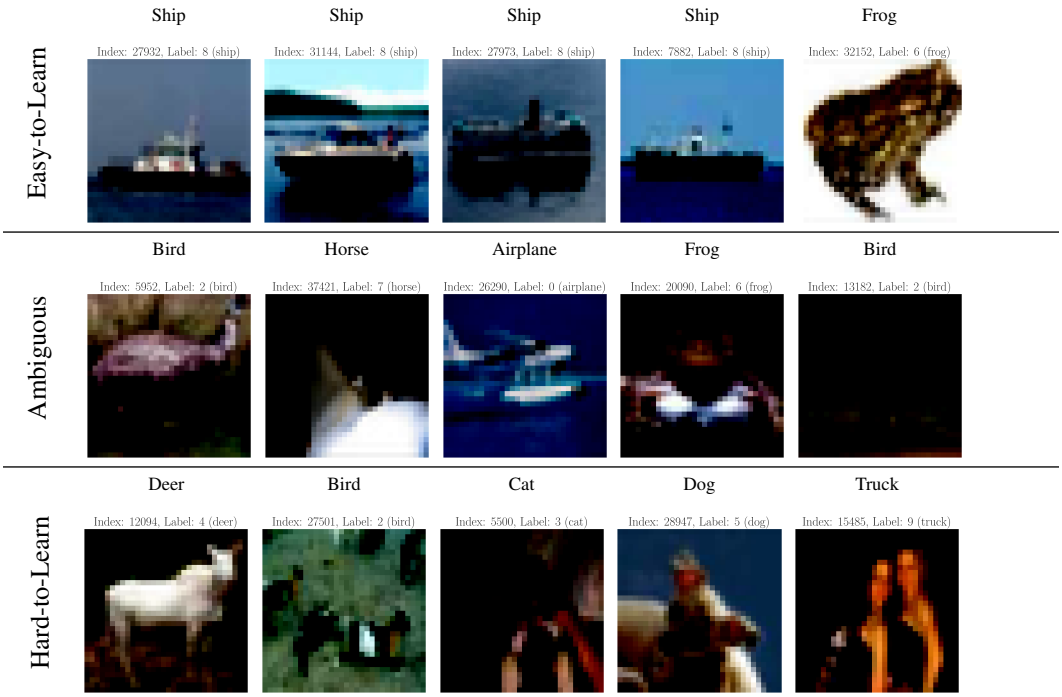


Figure 7: MNIST

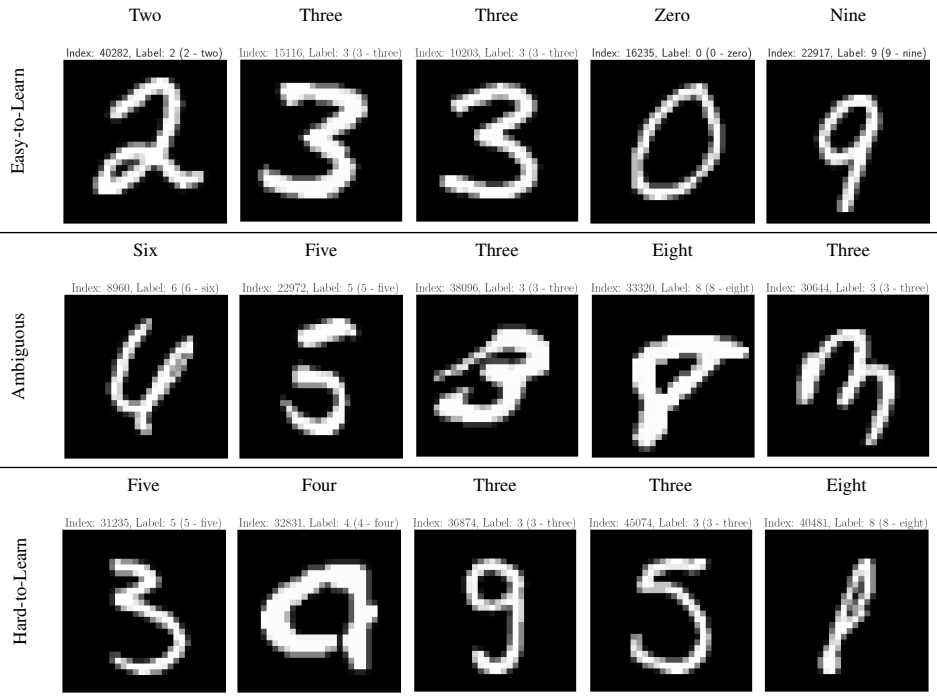
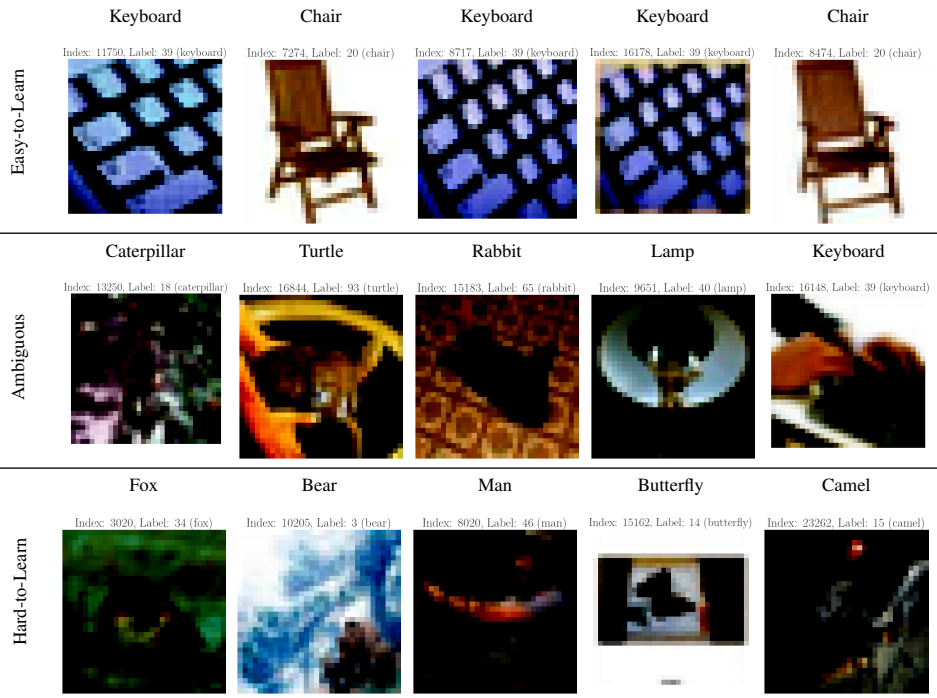


Figure 8: Cifar100



E Additional Datamaps

It is apparent that the stronger the model, the more data points are found in the 'easy-to-learn' region. Additionally, stronger models are typically more 'decisive' about data points, do not have as many samples in the mid-left region of the datamap (low variability, medium confidence).

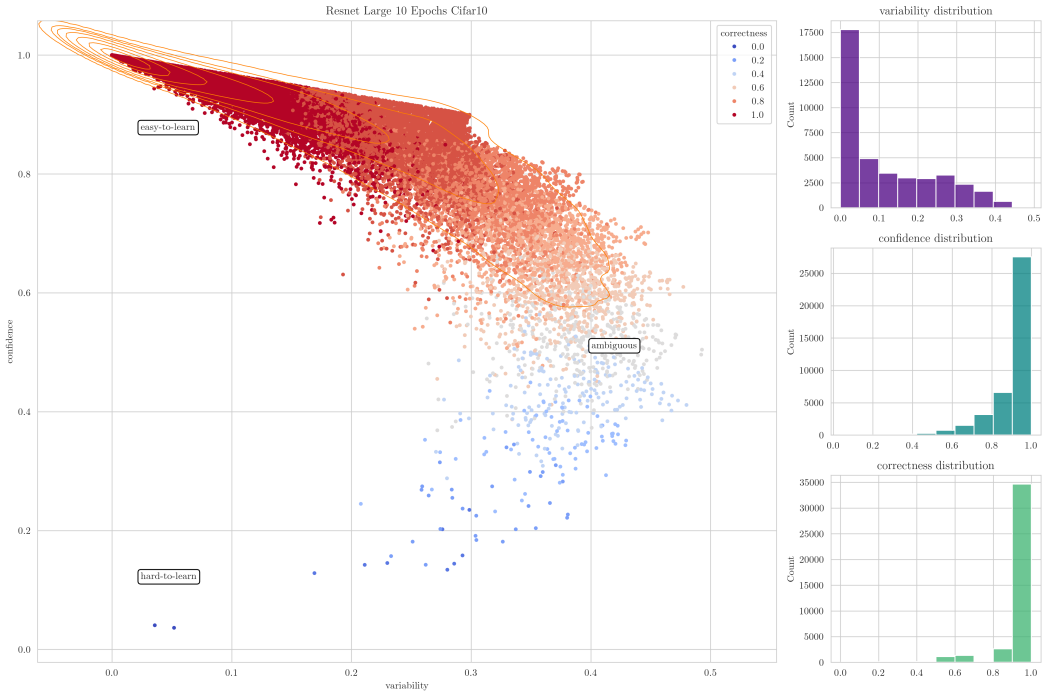


Figure 9: ResNet Large - Cifar-10

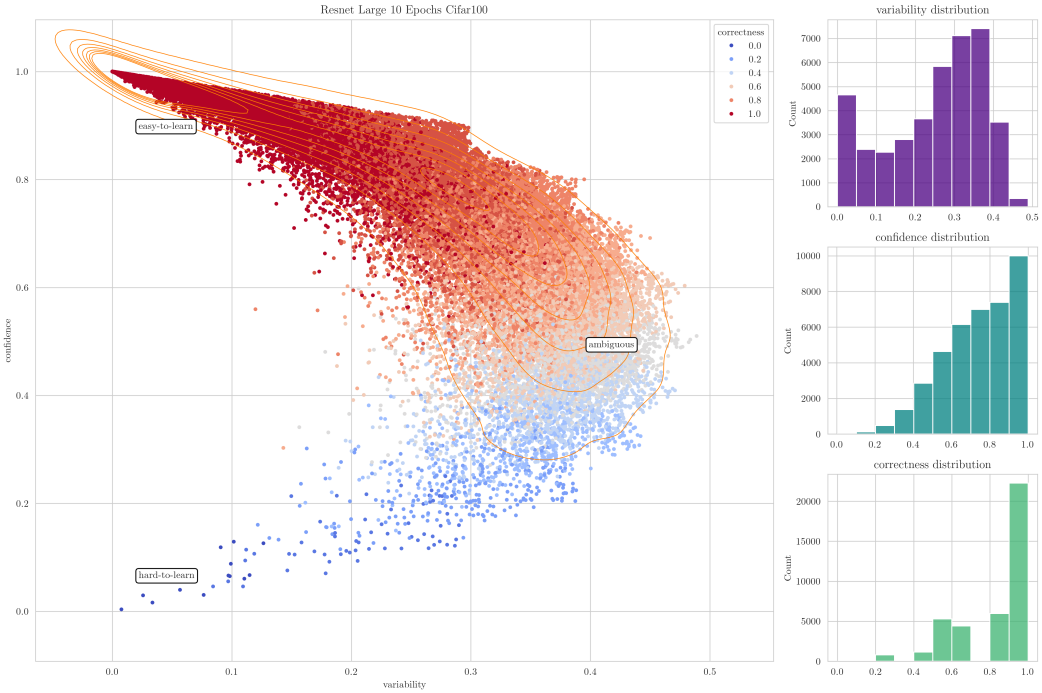


Figure 10: ResNet Large - Cifar-100

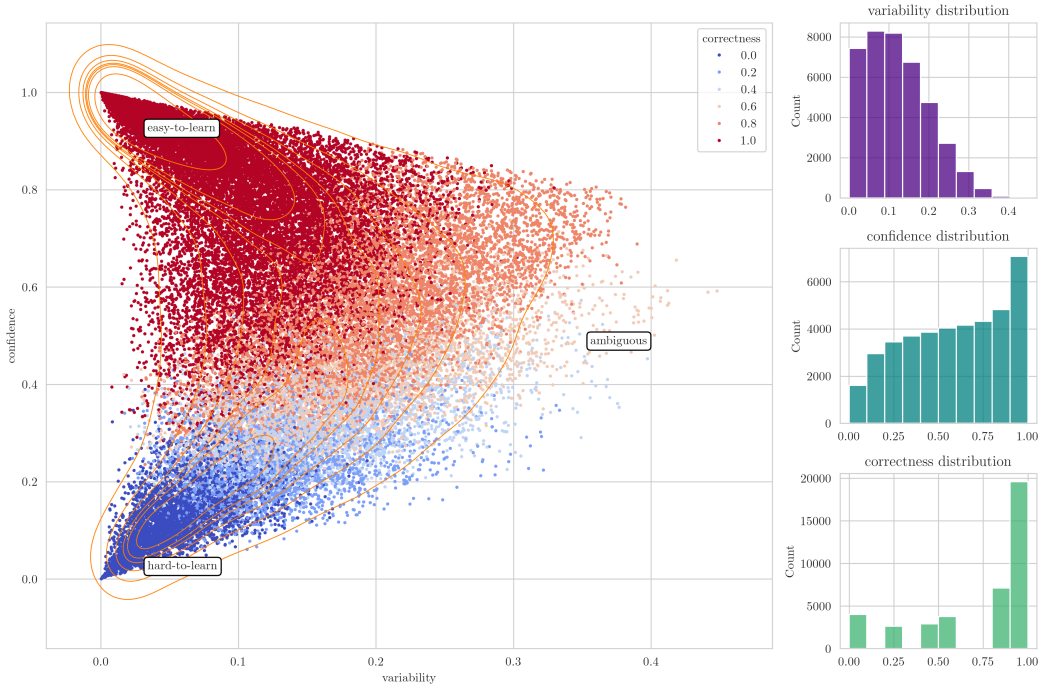


Figure 11: ResNet Small - Cifar10

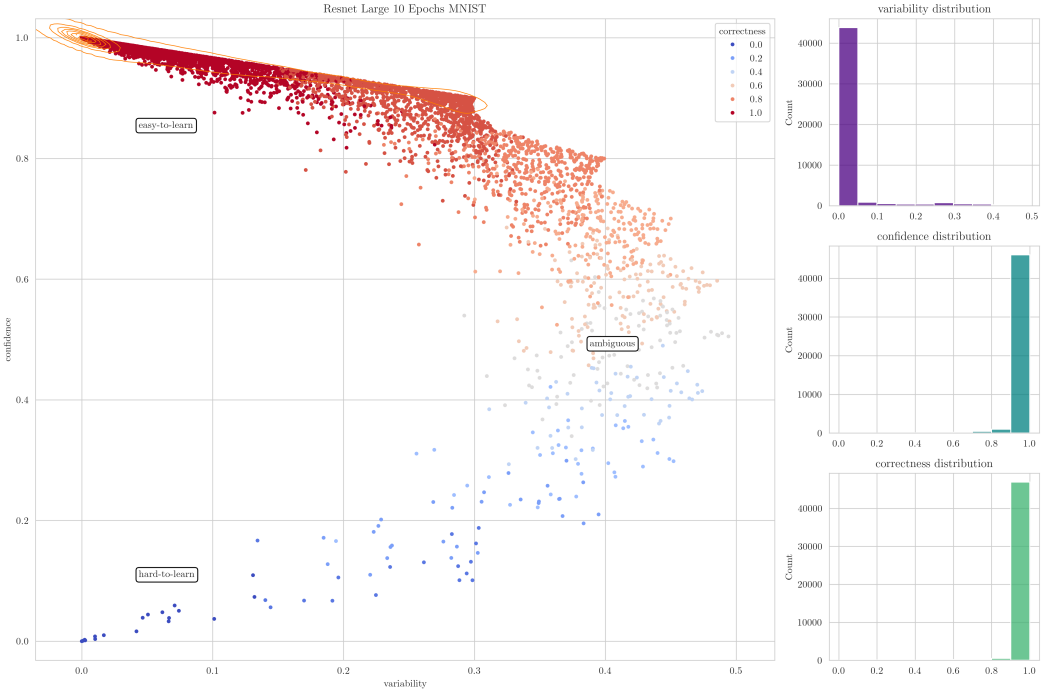


Figure 12: ResNet Large - MNIST