# Confusing Large Models by Confusing Small Models

Vítor Albiero[1], Raghav Mehta[2], Ivan Evtimov[1], Samuel Bell[1], Levent Sagun[1], Aram Markosyan[1]
[1]Meta AI, [2]McGill University

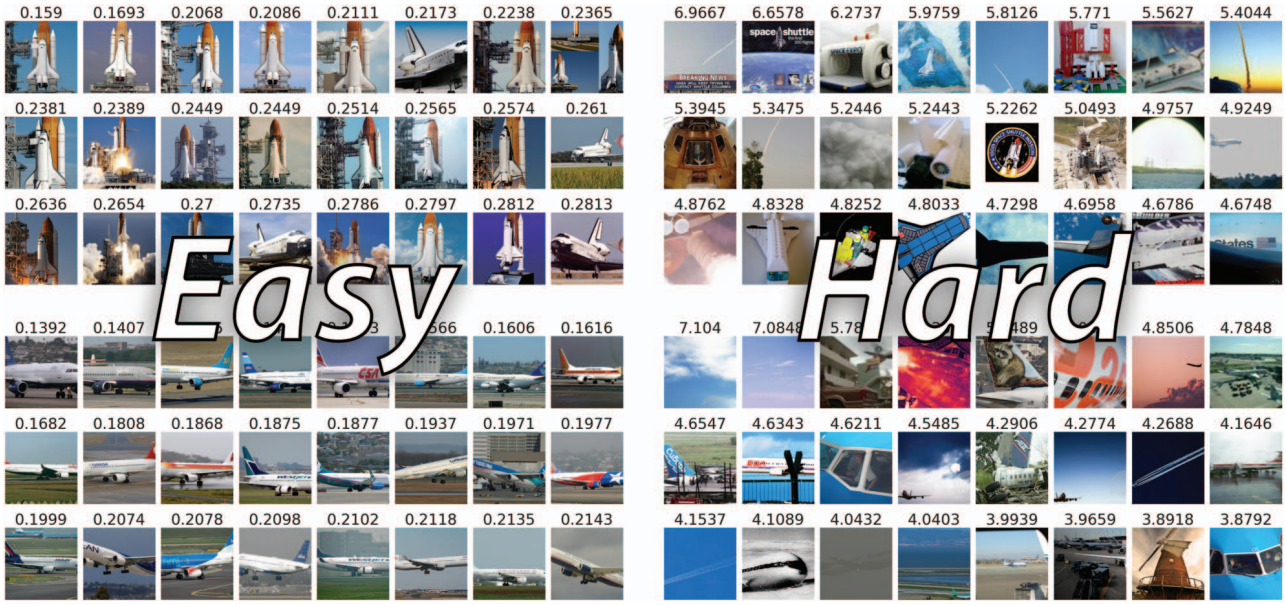



Figure 1: A *label-free* measure of data difficulty, the confusion score, neatly separates what the model finds easy from what it finds hard. **Top:** Samples from the ImageNet "space shuttle" class, captioned with confusion score from an ensemble of ResNet-50 models. **Bottom:** Samples from the "plane" class. **Left:** Easy samples are reasonably prototypical objects in common situations. **Right:** Visual inspection reveals that hard samples comprise a mix of unusual presentation, viewing angle, crop, background, or spurious correlations. **These hard samples are likely to confuse other related models, or their scaled-up counterparts, in a similar way.**

## Abstract

*Despite a steady growth in average accuracy, computer vision models continue to fail on many robustness benchmarks. In this paper, we take a step back from standard benchmarks and focus on how models perceive data, and which aspects of the data they find confusing. Using an ensemble-based confusion score we examine how the training and test samples appear simple or confusing to a given model. Based on these heuristics, we demonstrate an application of the confusion score in identifying images that appear confusing to the trained model, and show that these images are highly likely to be misclassified by the model. We further demonstrate how confusion carries over to models of various sizes and architectures, which gives rise to the possibility of identifying challenging images via ensembles of small networks to produce a* custom *benchmark of challenging data, that remains appropriate for large models where ensembling is costly to implement. Finally, we demonstrate how training via upsampling on confusing images can improve accuracy on the hard subset.*

## 1. Introduction

Computer vision models have continuously demonstrated steady improvements in overall accuracy, as measured by standard benchmarks. This progress has taken place across a range of computer vision tasks along a gradient of increasing complexity, from image classification [6, 9, 12], through object detection [24, 38], to image segmentation [22, 39] and beyond. However, beyond promising top-level *averaged* figures, models performance remains far

from robust, with failures likely to occur on dataset samples with, for example, unexpected backgrounds, spurious correlations, or trivial corruptions. The ability for researchers and practicioners to identify precisely *which* samples are challenging is an essential component of any robustness evaluation and any subsequent mitigation work.

In practical settings where machine learning systems are intended for real world deployment, it is increasingly common to evaluate model performance against various robustness and fairness benchmarks. Rather than a focus on *overall* performance such as average accuracy, these benchmarks rely upon disaggregation to surface the various categories of failures that models can exhibit, and can provide a way of quantifying the safety and reliability—or conversely the harmfulness—of a system along various dimensions. For example, a fairness audit might evaluate the performance of a given model on different demographic groups, or a robustness evaluation might consider performance degradation in light of plausible distribution shifts or adversarial attacks. However, in many realistic scenarios, such datasets may not be available, demographic information may be challenging or undesirable to obtain, or the dataset may only provide partial coverage of what the model finds difficult. Identifying precisely *what* a model finds difficult or upon which samples it will fail is a major open challenge in computer vision, and while numerous methods have proposed, each has their own shortcomings [21].

In this work, we build upon a model-dependent and label-free measure of data difficulty, the confusion score [32]. The confusion score quantifies the amount of disagreement on a given sample based on the class conditional probabilities estimated by an ensemble of trained models. Our empirical evaluations reveal that different scales of model find the many of the same examples confusing, which is of significant practical importance. Using the confusion score and a small, cheap and quick-to-train model, one can construct benchmarking datasets of challenging examples that will *also* be challenging to much larger, scaled-up models. As such, we can quickly and easily create a model-specific benchmarking dataset, and then use it to evaluate model improvements in a production or near-production setting. Crucially, because a dataset constructed from high confusion score examples does not rely on labels, it is particularly useful choice when working with unlabeled data, such as in an unsupervised or self-supervised setting. We also show that, when aggregated over all samples in a dataset, the amount of confusion a model exhibits strongly correlates with its overall performance (as measured by average accuracy). To summarize, we make three primary contributions. Our experiments reveal:

1. that both large and small models within an architecture class (e.g. ResNet variants) find the same examples confusing (Section 3.2). Moreover, the statement holds true for cross-architecture classes.

2. that the overall amount of confusion exhibited by a model is strongly correlated with its performance, i.e. confused models fare worse (Section 3.1); and

3. that confusion scores can serve as a helpful error predictor for unlabeled data (Section 3.1).

## 2. Related work

### 2.1. Benchmarking failure modes

Previous work has developed multiple benchmarks for specific failure modes. For example, ImageNet-C [17] and the AugLy library [27] contain images that have different kinds of blur applied to them or have their brightness and contrast amplified as well as other "corruptions." In ImageNet-R [16], images were re-rendered to have artistic style and ImageNet-A and -O [18] contain natural images that are "adversarial" or considered out-of-distribution. Another effort collected a new test set in the same distribution to test for generalization [29] and ImageNet-X [20] relabeled the existing validation set with natural factors of variation (such as rotation of the object). Other variants of ImageNet have also been proposed for different concrete notions of robustness, such as ImageNet-P, -Sketch, -ML, -Real, -ReLabel, -Stylized, and -Hard [4, 13, 26, 31, 37, 40]. In addition, "robustness" is often defined as accuracy on a test set of adversarial examples, test instances within an $L_p$ distance of the originals generated via optimization algorithms to mislead the model [35, 36]. In this work, we focus on a model and label free notion of hardness for models.

### 2.2. Evaluating data difficulty

We touch upon a number of proposed approaches for evaluating data difficulty below, though for a review see [33]. Meding et al. find that in broad strokes, *most* models make similar errors, show that various design decisions (such as regularization strategies, optimization methods, etc.) can play a role in decreasing agreement [25]. In contrast with our work, Meding et al.'s analysis of *error* consistency relies upon ground-truth labels, whereas our use of the confusion score circumvents this requirement. Bell and Sagun explore the model-specific nature of data difficulty, highlighting the challenges inherent in identifying what a model will find challenging without actually training said model [3]. In this work, we take a step towards countering this problem by showing that it is sufficient to find challenging datapoints on a smaller, simpler model, and that these will remain challenging for a scaled-up counterpart. Hacohen et al. show that models learn training samples in a mostly similar order, where easier examples are typically

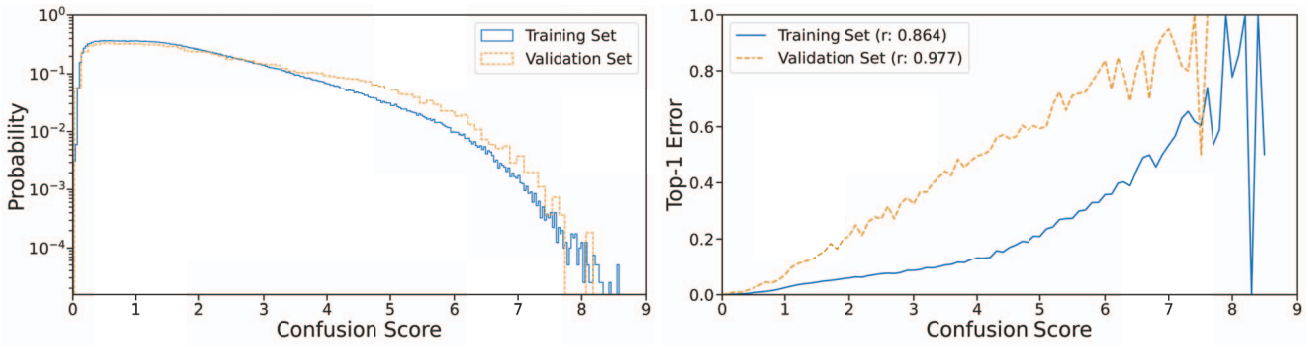



Figure 2: **Left:** Distribution (shown in log scale) of confusion scores from ensemble of ResNet-50 models on ImageNet training and validation sets. Low-confusion ("easy") samples dominate the dataset. **Right:** Confusion scores are strongly correlated (Pearson's $\rho$) with top-1 error. **The confusion score is a simple and effective way to identify error-prone samples without the need for labels.**

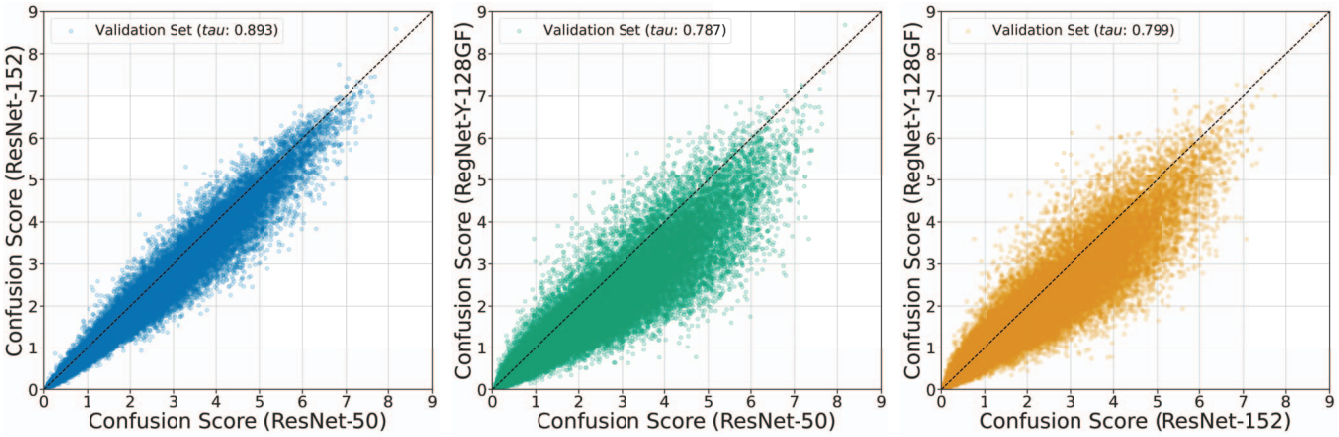
learned before more complex [14]. Hacohen et al.'s empirical evaluations show that this ordering of data difficulty is mostly shared between models of the same architecture, and shared to a lesser extent between models of different architectures, indicating the presence of both architecture-specific and architecture-independent components of data difficulty. Feldman et al. use memorization likelihood as a proxy for difficulty, and suggest that the long-tail of difficult play only a minimal role in overall training performance [11]. Baldock et al. approximate difficulty via *prediction depth*, i.e. the number of layers the sample must pass through before a prediction can be made, and show that predition depth is highly consistent across initializations of the same model, and reasonable consistent across similar model architectures [2]. Agarwal et al. propose that the variance of the per-sample gradient update, which is computationally expensive, through training as a measure of data difficulty, finding that it, like a confusion score, neatly separates examples that are intuitively difficult from those that are trivial [1]. In an NLP setting, Swayamdipta et al. use the variability of model confidence throughout training as a measure of difficulty, and argue (as do we) that datasets are mainly composed of easy examples [34].

### 2.3. Mitigating performance disparities

Faced with a dataset in which certain examples may be challenging, and others easy, it is likely that this will result in gaps in model performance. These performance disparities can lead to significant real-world harm, for instance in cases where the examples the model finds challenging are constrained to a certain demographic group [3], or alternatively where they represent a realistic distribution shift likely to be found in practice. While this work's primary focus is on the identification of challenging examples through the use of cheap, simple models, we explore mitigation approaches based on GroupDRO [30] in Section 3.4. Beyond collecting additional data [10], others have suggested subsampling or oversampling to equalize class sizes in order to offset performance disparities [19]. Similarly, Kirichenko et al. present results showing that fine-tuning on the final layer of a neural network on a balanced dataset can reduce disparities, though in order to balance effectively this approach relies on group annotation [23]. Alternatively, Byrd and Lipton explore the role of reweighting the importance of individual samples to the overall loss [5]. Key to each of these mitigation approaches is an understanding of the dimensions of dataset imbalance, be that along group lines or otherwise. Our experiments with identifying error-prone samples in a label-free manner, via the confusion score, are a potential step towards enabling the approaches discussed in this section but in the setting where group information may be missing or unavailable.

## 3. Empirical evaluation

We now present four empirical investigations into the confusion score and its potential applications. First, we evaluate whether the confusion score of specific samples can be used to identify which samples suffer high prediction error, to enable the creation of model-specific benchmarking datasets. While we use label information to make this point, i.e. to highlight the strong correlation between confusion score and prediction error, being able to identify challenging samples *without* label information is crucial in unsupervised settings. Second, we evaluate how transferable confusion scores are between different model scales within and between architecture classes. Here we are motivated by the desire to develop our benchmarking dataset of challenging examples using the cheapest and simplest model possible, and seek to determine whether this is possible. Third, we perform an ablation study to evaluate the robustness of confusion scores across different ensemble sizes (i.e., var-




(a) ResNet-50 vs ResNet-152 (b) ResNet-50 vs RegNet-Y-128GF (c) ResNet-152 vs RegNet-Y-128GF

Figure 3: Rank correlation (Kendall's $\tau$) of confusion scores of ImageNet validation samples between different models. **(a)** ResNet-50 and ResNet-152 exhibit very strong correlation, indicating that confusion is highly consistent as model scale increases. **(b, c)** Different architectures exhibit slightly decreased through still strong correlation, indicating the presence of both architecture-dependent and architecture-independent factors of confusion.

ious numbers of random seeds) and training times. Fourth and finally, we experiment with using confusion scores for mitigating performance disparities, by using a discretization of the confusion score as a proxy for group information (i.e., considering "hard" and "easy" as different groups). We use these proxy group labels in conjuction with Group Distributionally Robust Optimization (GroupDRO) [30], an optimization strategy that optimzes the performance of the *worst-performing* group, rather than averaging over all samples as in conventional empirical risk minimization (ERM).

Before proceeding, we briefly introduce the confusion score [32]. The confusion score of a single sample is the entropy of the of the ensemble-averaged class-conditional probabilities, then averaged over each training epoch. More precisely, following [32], let $f_t^{(i)}$ be a model that emits class-conditional probabilities for $C$ classes at epoch $t$, where the $i$ indexes the model within an ensemble of $M$. Then, for a given data point $x$ we average the probabilities over the ensemble,

$$p_t(x) = \frac{1}{M} \sum_{i=1}^{M} f_t^{(i)}(x) \,, \tag{1}$$

before computing the entropy of the averaged distribution,

$$s_t(x) = -\sum_{j=1}^{C} p_t^j(x) \log p_t^j(x) \,. \tag{2}$$

Finally, the entropy is averaged over epochs to reduce noise,

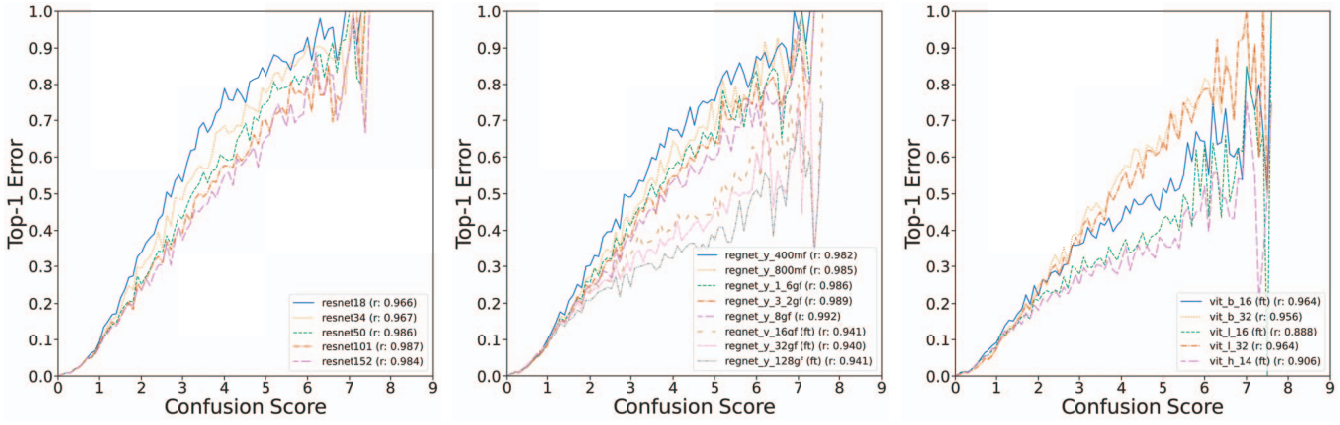
$$s(x) = \frac{1}{T} \sum_{t=1}^{T} s_t(x) \,, \tag{3}$$

such that $s(x)$ is the sample's confusion score for the given model ensemble. Note that to calculate the confusion score we rely on passing it through a model that outputs label *predictions*, but we do not need a ground truth label.

### 3.1. Error identification without labels

We trained a ResNet-50 [15] model with three random seeds, for a total of 100 epochs on ImageNet [7]. Then, using Eq. 3, we calculate the confusion scores across all epochs and 3 random seeds. The distribution of the obtained scores for training and validation sets are shown in Figure 2. We observe that most of the dataset is concentrated around lower scores, while only a small fraction of it is labeled with high scores. As shown in the Figure 2, we observed a correlation between the confusion scores and prediction errors in both training and validation sets. This result suggests that confusion scores can be used as error predictions for unlabeled data, which is crucial when deploying and testing real-world models. When comparing the sets, we can see less errors in the training set, since the models were trained on this data, but, surprisingly, we see better correlation with confusion scores in the validation sets, as showed by the higher Pearson's $\rho$ value. Samples from the ImageNet "n04266014 - space shuttle" and "n03954731 - plane" classes are shown in Figure 1, where we selected the least confusing (easiest) samples, and the most confusing (hardest) samples, using only the confusion scores, without the target labels. As we show in the the figure, the proposed confusion scores are able to separate the data into easy and hard samples, enabling the creation of "hard" sets.

The oscillatory behaviour of confusion scores for the




(a) ResNet Models (b) RegNet Models (c) ViT Models

Figure 4: Transferability across foundational models. Top-1 errors are strongly correlated (Pearson's $\rho$) with confusion scores, while varying their robustness w.r.t. to model size or model training data (ft models). Smaller models (e.g., ResNet18, RegNet-Y-400MF, etc) show higher errors across the entire confusion scores range, when compared to either larger models (e.g., ResNet-152, RegNet-Y-8GF) or models that were pre-trained with large amounts of data (e.g., RegNet-Y-128GF ft, ViT-H-14 ft).

high scores was also observed in the original work by Simsek *et al.* and is attributed to the lack of data for that score band.

### 3.2. Confusion is transferable across the scales and architectures

Calculating confusion scores on large models is expensive, so we conduct experiments to validate how the scores calculated on smaller models transfer to other (preferably larger) models. For this study we trained two ResNet [15] family models ResNet-50 and ResNet-152 following the same setup. In Figure 3, we show the correlation across their confusion scores for both training and validation sets. We can see that the confusion scores for the samples are strongly correlated between the ResNet-50 and ResNet-152. We can observe that correlation is much tighter for the low and high values of the confusion scores. One can also observe a skewness in the figures. Namely, more of the samples that have a given confusion score for the ResNet-50 correspond to lower confusion score for the ResNet-152. This can be interpreted as ResNet-50 being less robust than ResNet-152.

To further illustrate this point, we have included in our study 18 of the most used foundational models stemming from ResNet [15], RegNet [28], and ViT [8] architectures. We used publicly available trained weights[1], and confusion scores calculated with ResNet-50 across 100 epochs and 3 seeds. First, we observe that all models evaluated significantly increase their errors as the confusion scores increase. Second, we observe that larger (or fine-tuned *-ft*) models are more "robust" to confusing samples, e.g., there is a clear pattern across ResNets, where ResNet152 is robuster than ResNet101, ResNet101 is robuster than ResNet50, and so forth.

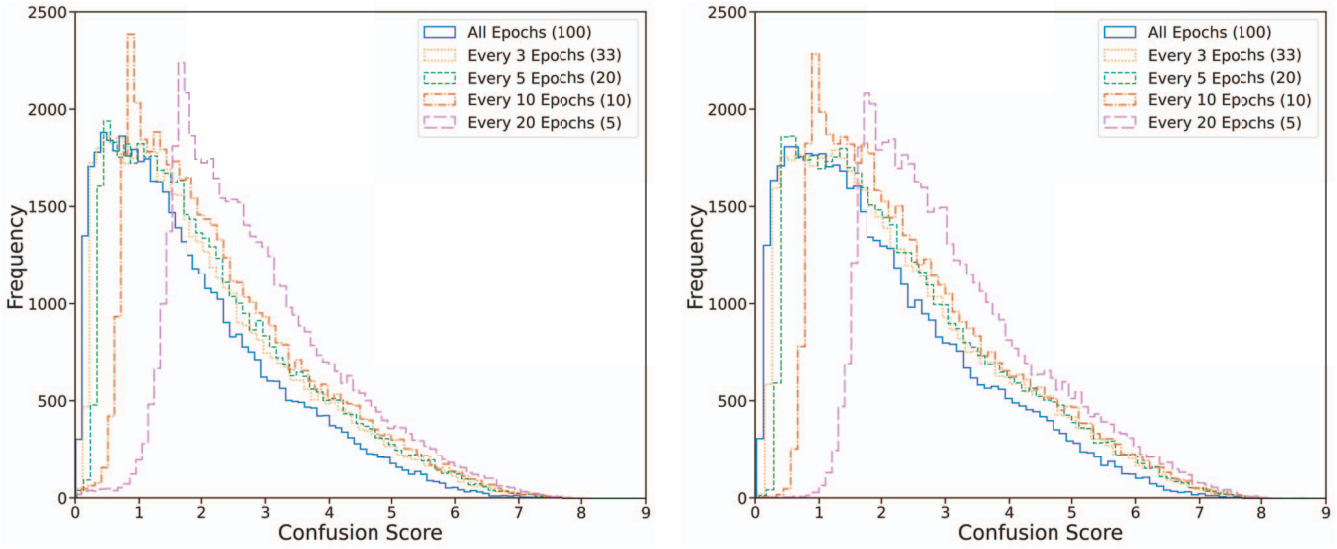
All in all, these experiments suggest that we can train a smaller model, which can be used to calculate confusion scores, to find problematic data that may affect other larger models.

### 3.3. Robustness of confusion scores

In this section we investigated the sensitivity of our results with respect to the number of seeds and the number of epochs. In Figure 5, we show the distribution of confusion scores when calculated using 5, 10, 20, 33, and 100 epochs using either 1 or 3 seeds. As we can see, for both 1 and 3 seeds, the confusion score distribution shifts towards higher values when the number of epochs used is reduced. When the number of epochs is low (10 and 5), we see a high peak in distribution around 1 for every 10 epochs, and around 2 for every 5 epochs. On the other hand, if more than 20 epochs are used, we do not see such high peaks, and we conclude that using at least 20 epochs is sufficient. We observe similar patterns across all 5 results when comparing 3 seeds to 1 seed, with the main difference that the confusion score distributions for 3 seeds are wider than for 1 seed, suggesting that by averaging across seeds, we gain more scores.

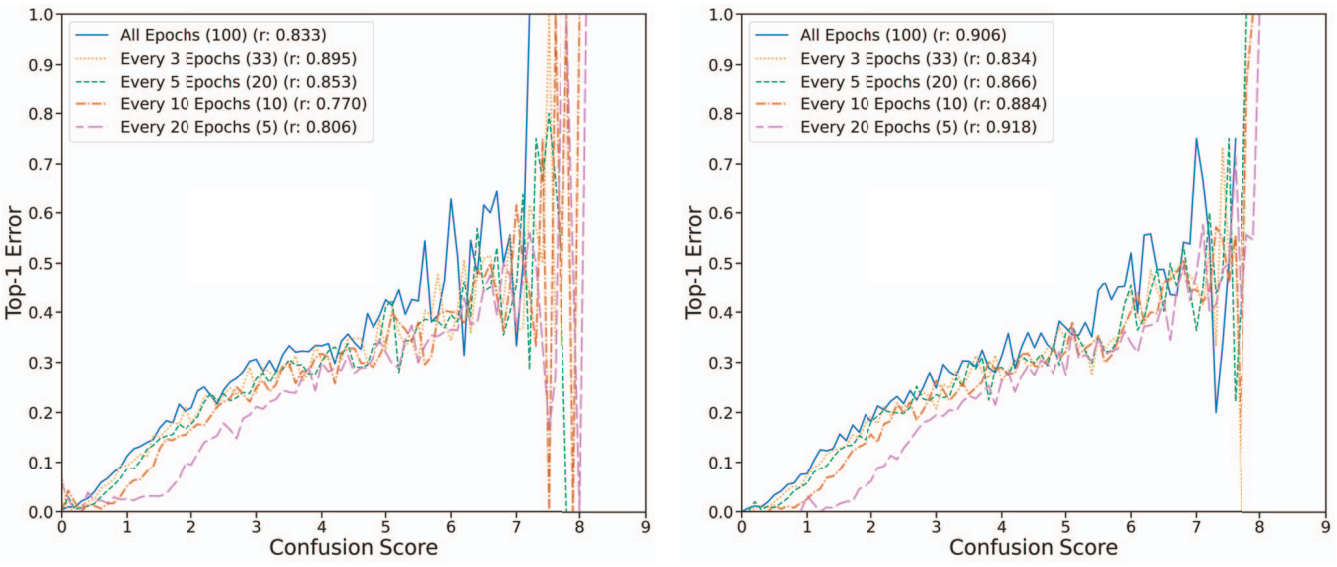
In the Figure 6 we show the comparison between the confusion scores calculated as aforementioned, and the most robust foundational model tested, ViT-H-14 (ft). For

[1] https://pytorch.org/vision/stable/models.html#table-of-all-available-classification-weights




(a) 1 Seed (b) 3 Seeds

Figure 5: The frequencies of confusion scores when calculated calculated using 5, 10, 20, and 100 epochs. The results are shown for **(a)** 1 and **(b)** 3 seeds. There are significant differences between the lower values (less than 20) of number of epochs. We also observe a widening of the distributions for larger number of seeds.




(a) 1 Seed (b) 3 Seeds

Figure 6: Confusion scores when calculated using 5, 10, 20, and 100 epochs correlation with the most robust foundational model evaluated (ViT-H-14 ft). The results are shown for **(a)** 1 and **(b)** 3 seeds. Using more seeds, leads to less noise in higher confusion score bands, due to the distribution widening. There is also a stronger correlation (Pearson's $\rho$) between confusion scores and top-1 error when more seeds are used.

both 1 and 3 seeds, we observe that top-1 errors increase w.r.t to the amount of epochs used to calculate the confusion scores, mostly due to the distribution shift previously discussed. We also observe significant amount of noise when confusion scores are higher than 7, particularly for the results using a single seed. This is attributed to to the low amount of data in high value bands, and is less significant when using 3 seeds. Results obtained using 3 seeds are also more robust, as can be observed by higher Pearson's $\rho$ values.

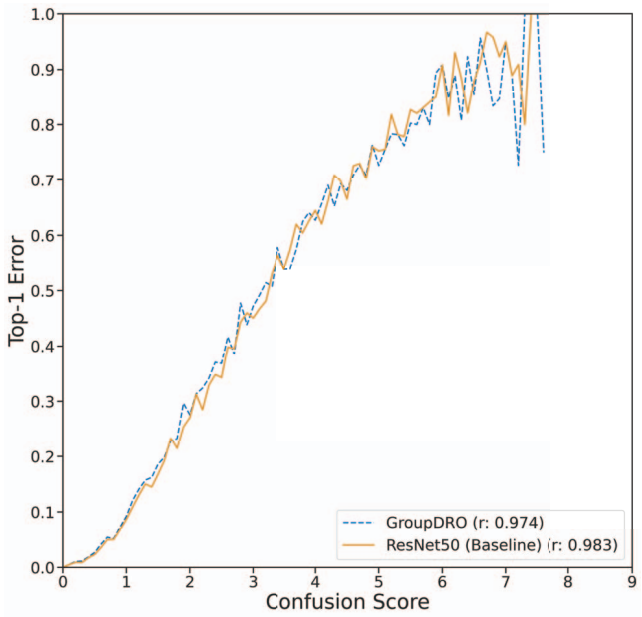



Figure 7: ResNet-50 training with GroupDRO using confusion scores as subgroup information, and traditional training. Using confusion scores as sub-group information, GroupDRO shows lower top-1 errors for most difficult samples.

### 3.4. Confusion-based mitigation

So far, we talked about how confusion scores can be used to discover unlabeled data that have high chances of prediction errors. However, we also conducted experiments to validate if the confusion scores can be used as a mitigation. We re-trained ResNet-50 model from scratch, using the previously calculated confusion scores as sub-group information to enable GroupDRO [30]. GroupDRO is an optimization strategy designed to boost worst-group accuracy.

The model trained with GroupDRO using confusion score as sub-group information, and traditional trained are shown in the Figure 7. As we can see, compared to the baseline, the model trained with GroupDRO powered by confusion scores, achieve overall higher errors, but when analyzing the most confusing samples, the pattern is reversed. This becomes clear when we split the confusion scores in the middle, creating two bands, $< 5$ and $\geq 5$. For $< 5$ images, GroupDRO achieves a top-1 error of 25.8%, while the baseline achieves 25%. On the other hand, for $\geq 5$ images, GroupDRO achieves a top-1 error of 80.6%, while the baseline achieves 81.9%, showing an increase of 1.3% on harder samples.

## 4. Conclusions and discussion

In the present empirical study we investigated the transferability properties of the confusion score developed by Simsek *et al.* [32]. We have made the following key observations:

1. the confusing examples are universally confusing across various foundational architectures;

2. there is a strong correlation between the model's performance and the confusion exhibited by the model;

3. the confusion scores can serve as a good indicator of model performance on the unlabeled data.

In light of our findings, we suggest confusion scores can be used to support the quick, cheap and easy creation of "hard" benchmark datasets that can be used in investigations of new and ongoing mitigations. We also see this technique as a potential way of directly powering mitigations that require sub-group information, e.g. re-weighting [5] or over/undersampling [19], an idea we explored in Section 3.4. The present approach can be effortlessly extended to other modalities.

## References

[1] Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10368–10378, June 2022. 3

[2] Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10876–10889. Curran Associates, Inc., 2021. 3

[3] Samuel J. Bell and Levent Sagun. Simplicity bias leads to amplified performance disparities, 2022. 2, 3

[4] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 2

[5] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 872–881. PMLR, 09–15 Jun 2019. 3, 7

[6] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[10] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2803–2813. PMLR, 13–18 Jul 2020. 3

[11] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2881–2891. Curran Associates, Inc., 2020. 3

[12] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 1

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[14] Guy Hacohen, Leshem Choshen, and Daphna Weinshall. Let's agree to agree: Neural networks share classification order on real datasets. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3950–3960. PMLR, 13–18 Jul 2020. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 2

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 2

[19] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 336–351. PMLR, 11–13 Apr 2022. 3, 7

[20] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022. 2

[21] Paul F Jaeger, Carsten T Lüth, Lukas Klein, and Till J Bungert. A call to reflect on evaluation practices for failure detection in image classification. *arXiv preprint arXiv:2211.15259*, 2022. 2

[22] Dongjun Kim, Yeongmin Kim, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022. 1

[23] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2022. 3

[24] Chuyi Li, Lulu Li, Yifei Geng, Hongliang Jiang, Meng Cheng, Bo Zhang, Zaidan Ke, Xiaoming Xu, and Xiangxiang Chu. Yolov6 v3. 0: A full-scale reloading. *arXiv preprint arXiv:2301.05586*, 2023. 1

[25] Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. Trivial or impossible — dichotomous data difficulty masks model differences (on imagenet and beyond). In *International Conference on Learning Representations*, 2022. 2

[26] Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[27] Zoe Papakipos and Joanna Bitton. Augly: Data augmentations for robustness, 2022. 2

[28] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 5

[29] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 2

[30] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 3, 4, 7

[31] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on ImageNet. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings*

*of Machine Learning Research*, pages 8634–8644. PMLR, 13–18 Jul 2020. 2

[32] Berfin Simsek, Melissa Hall, and Levent Sagun. Understanding out-of-distribution accuracies through quantifying difficulty of test samples. *arXiv preprint arXiv:2203.15100*, 2022. 2, 4, 7

[33] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[34] 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. Association for Computational Linguistics, Nov. 2020. 3

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[36] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 2

[37] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[38] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022. 1

[39] Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. Pfgm++: Unlocking the potential of physics-inspired generative models. *arXiv preprint arXiv:2302.04265*, 2023. 1

[40] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2340–2350, 2021. 2