TEACHER'S PET: UNDERSTANDING AND MITIGATING BIASES IN DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation is widely used as a means of improving the performance of a relatively simple "student" model using the predictions from a complex "teacher" model. Several works have shown that distillation significantly boosts the student's *overall* performance; however, are these gains uniform across all data subgroups? In this paper, we show that distillation can *harm* performance on certain subgroups, e.g., classes with few associated samples, compared to the vanilla student trained using the one-hot labels. We trace this behaviour to errors made by the teacher distribution being transferred to and *amplified* by the student model. To mitigate this problem, we present techniques which soften the teacher influence for subgroups where it is less reliable. Experiments on several image classification benchmarks show that these modifications of distillation maintain boost in overall accuracy, while additionally ensuring improvement in subgroup performance.

1 INTRODUCTION

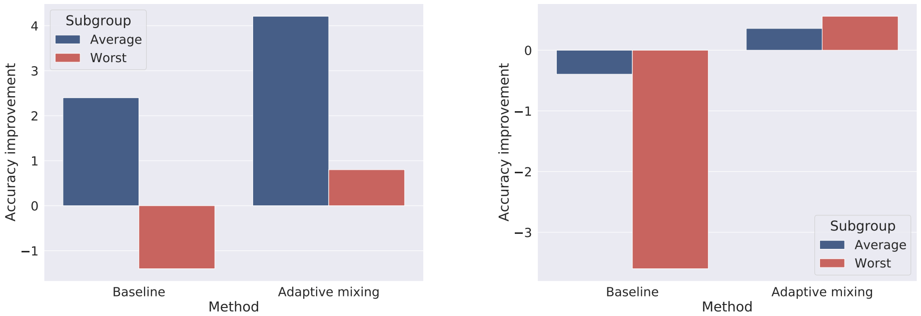
Knowledge distillation is a technique for improving the performance of a "student" model using the predictions from a "teacher" model. At its core, distillation involves replacing the one-hot training labels with the teacher's predicted label distribution. Empirically, distillation has proven successful for model compression (Bucilă et al., 2006; Hinton et al., 2015), improving the performance of a fixed model architecture (Anil et al., 2018; Furlanello et al., 2018), and semi-supervised learning (Radosavovic et al., 2018). Theoretically, several works (Lopez-Paz et al., 2016; Mobahi et al., 2020; Tang et al., 2020; Menon et al., 2020; Zhang & Sabuncu, 2020; Ji & Zhu, 2020; Allen-Zhu & Li, 2020; Zhou et al., 2021; Dao et al., 2021) have studied how distillation affects learning. Put together, both strands of work further the understanding of when and why distillation helps.

In this paper, we are similarly motivated to better understand the mechanics of distillation, but pose a slightly different question: does distillation help *all* data subgroups uniformly? Or, do its overall gains come at the expense of *degradation* of performance on certain subgroups? To our knowledge, there has been no systematic study (empirical or otherwise) of this question. This consideration is topical given the study of *fairness* of machine learning algorithms on under-represented subgroups (Hardt et al., 2016; Buolamwini & Gebru, 2018; Chzhen et al., 2019; Sagawa et al., 2020a).

Our first finding is that even in standard settings — e.g., on image classification benchmarks such as CIFAR — distillation can *disproportionately harm* performance on subgroups defined by the individual classes (see Figure 1). To discern the source of this behaviour, we ablate the teacher and student architectures (§3.2), dataset complexity (§3.3), label frequencies (§3.4). These point to the potential harms of distillation when the teacher makes *confident mispredictions* on a subgroup. We emphasise that this phenomenon is not simply the teacher's worst class accuracy translating to student; instead, distillation *amplifies* worst class imbalance much more than the teacher.

Having identified a potential limitation of distillation, we present two simple techniques to remedy it. These apply per-subgroup mixing weights between the teacher and one-hot labels, and per-subgroup margins respectively (§4). Intuitively, these limit the influence of teacher predictions on subgroups it models poorly. Experiments on image classification benchmarks show that these methods typically maintain a boost in overall accuracy, while ensuring a more equitable improvement across subgroups.

In sum, this work provides novel insights into distillation performance, with three main contributions:



(a) CIFAR-100-LT.

(b) ImageNet.

Figure 1: Illustration of the potential deleterious effects of distillation on data subgroups. We train a ResNet-56 teacher on **CIFAR-100-LT**, a long-tailed version of CIFAR-100 (Cui et al., 2019; Cao et al., 2019) where some labels have only a few associated samples, and a ResNet-50 teacher on **ImageNet**. For each dataset, we self-distill to a student ResNet of the same depth. On CIFAR-100-LT, as is often observed, distillation helps the *overall* accuracy over one-hot student training ($\sim$2% absolute). However, such gains come at significant cost on subgroups defined by the individual classes: on the ten rarest classes, distillation *harms* performance by $\sim$1%. Similarly, on ImageNet, distillation harms the average accuracy of the worst-10 classes by $\sim$3%. Our proposed techniques (§4) can roughly preserve the overall accuracy, while boosting subgroup performance.

(i) we identify a hitherto unexplored issue with distillation, namely, that its improvements in overall accuracy may come at the expense of harming accuracy on certain subgroups (§3.1). Such a finding is topical given the widespread practical use of generic learning paradigms such as distillation, and the increasing societal applications of learning systems more broadly.

(ii) we ablate potential sources for the above phenomenon (§3.2, §3.3, §3.4), and in the process identify certain characteristics of data (e.g., skewed label distributions) where it can manifest. This systematic empirical analysis aims at understanding a non-trivial phenomenon, rather than empirically comparing different learning methods.

(iii) we propose two simple modifications of distillation that mitigate the above problem, based on applying per-subgroup mixing weights and margins (§4); these perform well empirically (§5).

While the paper's analysis is empirical rather than theoretical, we note that our focus is on systematic *analysis* aiming at understanding a non-trivial phenomenon, rather than merely empirical *comparison*: Section §3 is devoted to carefully understanding the extent of, and causes for, the non-uniform gains of distillation. This is in line with works which empirically analyse neural network phenomena, e.g. Zhang et al. (2017); Müller et al. (2019); Nakkiran et al. (2020); Neyshabur et al. (2020).

2 Background and Related Work

Knowledge distillation. Consider a multi-class classification problem over instances $\mathcal{X}$ and labels $\mathcal{Y} = [L] \doteq \{1, \ldots, L\}$. Given a training set $S = \{(x_n, y_n)\}_{n=1}^N$ drawn from some distribution $\mathbb{P}$, we seek a classifier $h \colon \mathcal{X} \to \mathcal{Y}$ that minimises the *misclassification error* $E_{\mathrm{avg}}(h) \doteq \mathbb{P}(h(x) \neq y)$. In practice, one may learn *logits* $f \colon \mathcal{X} \to \mathbb{R}^L$ to minimise $\hat{R}(f) = \frac{1}{N} \sum_{n=1}^N \ell(y_n, f(x_n))$, where ℓ is a loss function such as the softmax cross-entropy, which for softmax probabilities $p_y(x) \propto \exp(f_y(x))$ is $\ell(y, f(x)) \doteq -\log p_y(x)$. One may then classify the sample via $h(x) = \arg\max_{y \in [L]} f_y(x)$.

Knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015) employs the logits $f^{\mathrm{t}} \colon \mathcal{X} \to \mathbb{R}^L$ of a "teacher" model to train a "student" model. The latter learns logits $f^{\mathrm{s}} \colon \mathcal{X} \to \mathbb{R}^L$ to minimise:

$$\hat{R}_{\mathrm{dist}}(f) = \frac{1}{N} \sum_{n=1}^{N} \Big[(1 - \alpha) \cdot \ell(y_n, f(x_n)) + \alpha \cdot \sum_{y' \in [L]} p^{\mathrm{t}}_{y'}(x_n) \cdot \ell(y', f(x_n))\Big], \tag{1}$$

where $\alpha \in [0, 1]$. Here, one converts the teacher logits to probabilities $p^{\mathrm{t}} \colon \mathcal{X} \to \Delta_L$ for simplex Δ, e.g. via a softmax transformation $p^{\mathrm{t}}_{y'}(x) \propto \exp(f^{\mathrm{t}}_{y'}(x))$. The second term *smooths* the student labels based on the teacher's confidence that they explain the sample. The first term includes the original label, so as to prevent incorrect teacher predictions from overwhelming the student. One further important trick is *temperature scaling* of the teacher logits, so that $p^{\mathrm{t}}_{y'}(x) \propto \exp(T^{-1} \cdot f^{\mathrm{t}}_{y'}(x))$. Setting $T \gg 0$ makes p^{t} more uniform, thus preventing overconfident predictions (Guo et al., 2017).

Average versus subgroup performance. The above exposition treats the misclassification error $E_{\mathrm{avg}}(h)$ as the fundamental performance measure of interest. However, suppose the data contains *subgroups* $\mathcal{G} = \{1, \ldots, G\}$. Defining the *per-subgroup errors* $\mathrm{err}_g(h) \doteq \mathbb{P}(h(x) \neq y \mid g)$, we have $E_{\mathrm{avg}}(h) = \sum_{g \in \mathcal{G}} \mathbb{P}(g) \cdot \mathrm{err}_g(h)$, which may mask errors on samples with $\mathbb{P}(g) \sim 0$ (Sagawa et al., 2020a;b; Sohoni et al., 2020). To this end, one may instead measure the *balanced* error (Menon et al., 2013) $E_{\mathrm{bal}}(h) \doteq \sum_{g \in \mathcal{G}} \frac{1}{|\mathcal{G}|} \cdot \mathrm{err}_g(h)$ which treats the subgroup distribution as uniform, or the *worst-subgroup* error (Sagawa et al., 2020a;b; Sohoni et al., 2020) $E_{\max}(h) \doteq \max_{g \in \mathcal{G}} \mathrm{err}_g(h)$, which focusses on the worst-performing subgroup. An intermediary is the average of the k worst-performing subgroups (Williamson & Menon, 2019): for ith largest per-subgroup error $\mathrm{err}^{[i]}(h)$, $E_{\mathrm{top-k}}(h) \doteq \frac{1}{k} \sum_{i=1}^{k} \mathrm{err}^{[i]}(h)$.

The definition of subgroups is a domain-specific consideration. One special case is where each label defines a subgroup (i.e., $\mathcal{G} = \mathcal{Y}$), and $\mathbb{P}(y)$ is skewed. In such *long-tail* settings (Buda et al., 2017), classifiers with good average performance can perform poorly on "tail" labels where $\mathbb{P}(y) \sim 0$.

Related work. There is limited prior study that dissects distillation's overall gains per subgroup. Zhao et al. (2020) showed that in incremental learning settings, distillation can be biased towards recently observed classes. We show that even in offline settings, distillation can harm certain classes. Recently, Zhou et al. (2021) studied the standard *aggregate* performance E_{avg} of distillation, which was tied to a certain subset of "regularisation samples". By contrast, our primary concern is to understand the *subgroup* performance of distillation.

Study of the *fairness* of machine learning algorithms on under-represented data subgroups has received recent attention (Dwork et al., 2012; Hardt et al., 2016; Buolamwini & Gebru, 2018; Chzhen et al., 2019). This has prompted dissection of the performance of techniquess such as dimensionality reduction (Samadi et al., 2018), increasing model capacity (Sagawa et al., 2020a), and selective classification (Jones et al., 2021). We follow the general spirit of such works, studying a more delicate setting involving *two* separate models (the student and teacher), each with their own inductive biases. We present more discussion of related directions in §6. In a related effort, Hooker et al. (2019) explore how model compression may harm subgroup performance compared to the original model.

3 Are distillation's gains uniform?

We demonstrate that the gains of distillation are *not uniform across subgroups*: specifically, considering subgroups defined by classes, distillation can *harm* the student's performance on the "hardest" few classes (§3.1). To understand the genesis of this problem, we perform ablations (cf. Table 1) that establish its existence in settings where there are insufficient samples to model certain classes, either due to the number of classes being large (§3.3), or the class distribution being skewed (§3.4). We then identify that the student may amplify the teacher's errors (§3.6). Finally, we corroborate these results for a more general notion of subgroup in a fairness dataset (§3.5).

3.1 Distillation hurts subgroup performance

To begin, we consider the effect of distillation on a standard image classification benchmark, namely, ImageNet. We employ a self-distillation (Furlanello et al., 2018) setup, with ResNet-34 teacher and student models, trained with standard hyperparameter choices (see Appendix B). Following Cho & Hariharan (2019), we use early stopping on the teacher model. We now ask: what is the impact of distillation on the student's *overall* and *per-class* performance? The first question has an expected answer: distillation improves the student's average accuracy by $+0.4\%$ (see the *Baseline* setting in Table 1). Judged by this conventional metric, distillation is thus a success.

A more nuanced picture emerges when we break down the source of the above improvement. We compute the per-class accuracies for the one-hot and distillation models, to understand how the overall gains of distillation are distributed. Figure 2 shows that these gains are non-uniform: distillation

Setting	Dataset	Avg acc	Worst subgroup acc
Baseline (§3.1)	ImageNet	+0.39	-0.43
Stronger teacher (§3.2)	ImageNet	+0.17	-1.19
Long tail (§3.4)	CIFAR-100 LT	+2.17	-1.46
	ImageNet LT	+0.21	-0.32
Fairness (§3.5)	UCI Adult	+3.10	-5.94
Reduce #classes (§3.3)	CIFAR-100	+1.93	+3.33
	ImageNet-100	+0.09	+0.06

Table 1: Summary of findings in the ablation analysis of distillation's subgroup performance (§3). In a range of different settings, distillation is seen to hurt the hardest subgroup accuracy (worst-k class accuracy or worst subgroup according to an attribute), despite improving the average accuracy (upper rows). Decreasing number of labels helps improve the hardest classes (bottom row).

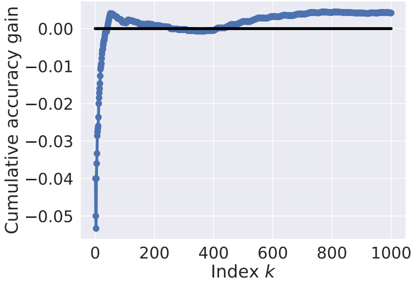


Figure 2: Cumulative gain of ResNet-34 self-distillation on ImageNet. For index k, we compute the gain in average accuracy over the k worst classes. While average accuracy ($k = 1000$) improves by $+0.4\%$, for $k \leq 40$, distillation *harms* over the one-hot model (evidenced by the negative gain).

Teacher	Student	Average accuracy	Worst-10 accuracy
EffnNet	Res-50	+0.17	-1.19
EffnNet	Res-34	0.00	-0.80
EffnNet	Res-18	+0.05	-1.60

Teacher	Student	Average accuracy	Worst-10 accuracy
Res-50	Res-50	+0.39	-0.43
Res-50	Res-34	+0.39	-2.05
Res-50	Res-18	-0.09	-1.00

Teacher	Student	Average accuracy	Worst-10 accuracy
Res-34	Res-34	+0.42	-2.60
Res-34	Res-18	+0.15	-3.60
Res-18	Res-18	-0.13	-3.80

Table 2: Summary of effect of distillation on different teacher and student architectures considered for the ImageNet dataset. The comparison is with respect to the one-hot (i.e., non-distilled) student. We find that distillation consistently hurts accuracy of the worst 10 classes.

in fact *hurts* the worst-k class performance for $k \leq 40$. (See Appendix C for a detailed per-class breakdown.) Thus, distillation may harm the student on classes that it already finds difficult. At the same time we note that distillation does improve many classes, such as the classes with relatively high accuracies from the one-hot student.

Given that average accuracy improves, it is worth asking whether the above is a cause for concern: does it matter that performance on subgroups corresponding to the "hardest" classes suffers? While ultimately a domain-specific consideration, in general exacerbating subgroup errors may lead to issues from the fairness perspective. Indeed, we shall see that distillation can also harm in settings where the subgroups correspond to sensitive variables; we discuss this further in §3.5.

At this stage, it is apposite to ask whether the above is an isolated finding, or indicative of a deeper issue. We thus study each of the following in turn: (i) does the finding hold in settings beyond self-distillation? (ii) does the finding hold for other datasets, or is it simply due to the idiosyncrasies of ImageNet? (iii) what are some general characteristics of settings where the problem is manifest?

3.2 Is distillation biased by the model size or architecture?

Having begun with a self-distillation setup, we now demonstrate that similar findings hold when the student and teacher architectures differ. Continuing with the ImageNet dataset, in the second row in Table 1 we report statistics for the overall average accuracy and average accuracy over the worst 10 classes when distilled ResNet-50 student from a stronger teacher: Efficient-Net (Tan & Le, 2021) teacher achieving 85.7% accuracy. Again, we see improved average accuracy and harmed Worst subgroup accuracy, composed of the 10 classes with the lowest accuracy according to teacher's performance. In Table 2, we report statistics when varying teacher and student architectures. The detrimental effect of distillation on hard class performance holds across all scenarios: thus, our earlier results were not specific to self-distillation.

For self-distillation settings, smaller models appear to incur greater losses on the worst-class error. When distilling between different architectures (e.g., from ResNet-50 to ResNet-18), we observe that even average accuracy may not improve, as noted in Cho & Hariharan (2019). There is however no clear trend between the difference in architectures and drop in worst class performance. Finally, we note that there is little change in the teacher's and student's worst-k classe; see Figure 6 (Appendix).

3.3 Is distillation biased by a large number of classes?

Having seen that ImageNet consistently demonstrates a performance degradation on certain classes, we now repeat the same analysis on a smaller image classification benchmark. We return to the self-distillation setup, using ResNet-56 models on CIFAR-100. On this dataset, Table 1 shows a (perhaps more expected) result: distillation boosts *both* the average and worst-1 class performance. This indicates that, at a minimum, the behaviour of distillation's performance gains are problem-specific; on CIFAR, distillation appears a complete win for both the average and subgroup accuracy.

One plausible hypothesis is that the tension between average and subgroup performance only manifests on problems with many labels, which might informally be considered "harder". To confirm this further while fixing the dataset, we randomly select 10% of classes from the ImageNet dataset, and only keep examples corresponding to those classes across the train and validation sets. Consistent with the results for CIFAR-100, we again find that the worst-10 class accuracy is not harmed under distillation (see the right side of Table 1 where we report results on ResNet-34 self-distillation).

The above indicates that for problems with a few, balanced labels, there may not be a tension between average and worst-subgroup performance under distillation. However, we now show that even for problems with relatively few labels, one may harm subgroup performance if there is *label imbalance*.

3.4 Is distillation biased by class imbalance?

We now consider a *long-tail* setting, where the training label distribution $\mathbb{P}(y)$ is highly non-uniform. Following the long-tail learning literature (Cui et al., 2019; Cao et al., 2019; Kang et al., 2020), we construct "long-tailed" (**LT**) versions of the above datasets, wherein the training set is down-sampled so as to achieve a particular label skew. For ImageNet, we use the long-tailed version from Liu et al. (2019). For other datasets, we down-sample labels to follow $\mathbb{P}(y = i) \propto \frac{1}{\mu^i}$ for constant μ and $i \in [L]$ (Cui et al., 2019). The ratio of the most to least frequent class is set to 100.

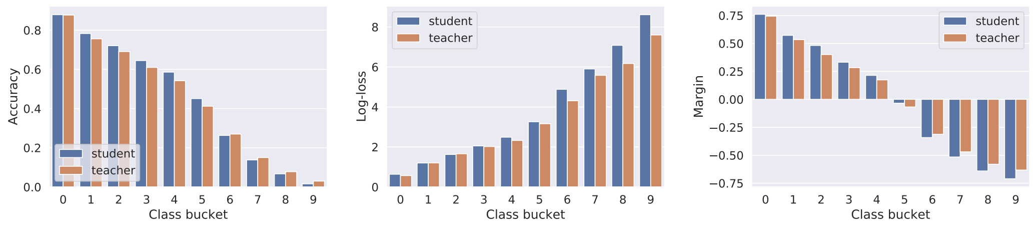
From the *Long tail* setting in Table 1, we note that on both CIFAR-100-LT and ImageNet-LT, accuracy over the hardest classes drops (we report results from self-distillation using ResNet-56 for CIFAR-100 and ResNet-50 ImageNet). The former is particularly interesting, given that the standard CIFAR-100 shows gains amongst the hardest classes. This provides evidence that for some "harder" problems — e.g., where there are insufficiently many samples from which to model a particular class — there may be a tension between the average and subgroup performance.

3.5 Beyond classes: other choices of subgroups

Our analysis thus far has focused on subgroups defined by classes. This choice is natural for long-tailed problems, where it is important to ensure good model performance on rare classes (Kang et al., 2020). In other problems, different choices of subgroups may be appropriate. For example, in problems arising in fairness, one may define subgroups based on certain sensitive attributes (e.g., sex, race). In such settings, does one similarly see varying gains from distillation across subgroups?

We confirm this can indeed hold on the UCI Adult dataset using random forest models (details in Appendix C.3). This data involves the task of predicting if an individual's income is ≥ 50K or not, and possesses subgroups defined by the individual's race and sex. Akin to the preceding results, we find that distillation can significantly improve *overall* accuracy, at the expense of *degrading* accuracy on certain rare subgroups, e.g., Black women; see Table 1, and Table 8 (Appendix). This further corroborates our basic observation on the non-uniform distribution of distillation's gains.

A distinct notion of subgroup was recently considered in Zhou et al. (2021), who identified the impact of certain "regularisation samples" on distillation. These are a subset of training samples which were seen to degrade the *overall* performance of distillation. It is of interest whether such a subgroup relates to our previously studied subgroups of "hard" classes; e.g., is there an abundance of regularisation samples in such subgroups, which might explain the poor performance of distillation? In Appendix C.4, we study the relationship between regularisation samples, and the per-label subgroups



(a) Accuracy.

(b) Log-loss.

(c) Margin.

Figure 3: Logit statistics on CIFAR-100 LT, for the teacher and distilled student under a self-distillation setup (ResNet-56 → ResNet-56). We show the statistics on 10 class buckets: these are created by sorting the 100 classes according to the teacher accuracy, and then creating 10 groups of classes. As expected, the student follows the general trend of the teacher model. Strikingly, we observe that the teacher model tends to systematically *confidently mispredict* samples in the higher buckets, thus incurring a *negative* margin; such misplaced confidence is largely transferred to the student, whose accuracy suffers on such buckets. Note that we consider statistics on the *test* set.

from our analysis; we find that, in general, these may be complementary notions. We further analyze the effect of the technique proposed in Zhou et al. (2021) on average and subgroup accuracies in §5.

3.6 Why does distillation hurt certain subgroups?

The above has established that in a range of scenarios, distillation can hurt performance on subgroups defined by individual classes. However, a firm understanding of *why* this happens remains elusive. To study this, we consider ResNet-56 self-distillation on CIFAR-100-LT — which showed a stark gap between the average and subgroup (i.e., worst-1 class) performance — and dissect the logits of the teacher and distilled student. (See the Appendix for plots where the teacher and student architectures differ.) Across classes, we seek to understand: (i) how aligned are the student and teacher *accuracies*? (ii) how reliable are the models' *probability estimates*? (iii) how do the models' *confidences* behave?

For a *test*[1] example (x, y) and predicted label distribution $p(x) \in \Delta_L$, we thus compute each models' accuracy, log-loss $\ell_{\log}(y, p(x)) = -\log p_y(x)$, and *margin* (Koltchinskii & Panchenko, 2002) $\ell_{\mathrm{marg}}(y, p(x)) = p_y(x) - \max_{y' \neq y} p'_y(x)$. Note that the latter may be negative if the model predicts the incorrect label for the example. Figure 3 shows these metrics on 10 class buckets: these are created by sorting the 100 classes according to the teacher accuracy, and then creating 10 buckets of classes. Within each bucket, we compute the average of the metric specified above.

Remarkably, for 5 out of 10 class buckets, average margins are *negative*, suggesting that the teacher is often *wrong yet confident* in predicting these classes. On these buckets, the student accuracy generally worsens compared to the teacher. Further, log-loss increases across all buckets (including those where accuracy improves), indicating reduced confidence in the true class of the distilled student. This points at a potential source of the poorer performance on the worst-1 accuracy. Recall that the distilled student's aim is to mimic the teacher's logits on the *training* samples. This is a proxy to the student's true goal, which is mimicking these logits on *test* samples, so as to attain similar generalisation performance as the teacher. When such generalisation happens, the student can thus be expected to roughly inherit the teacher's per-class performance; in settings like the above, this unfortunately implies it will perform poorly on those classes with negative teacher margin.

4 Improving subgroup performance under distillation

We now study simple means of correcting distillation to prevent the degradation of subgroup performance identified above. These leverage the insight that the behaviour is potentially a result of

[1] The choice of test, rather than train, example is crucial: an overparameterised teacher will likely correctly predict *all* training samples, thus rendering the above statistics of limited use. To leverage the insights from the above analysis in practice, we shall use a holdout set that can be carved out from the training set.

the teacher *confidently mispredicting* on some subgroups. In the following, for concreteness and simplicity, we focus on subgroups that are given by the individual classes.

4.1 Distillation with adaptive mixing weights

In §3.6, we saw that distillation can hurt on classes where the teacher is inherently inaccurate. Such inaccuracy may in fact be *amplified* by the student, which is hardly desirable. An intuitive fix is to simply rely less on the teacher for classes where it performs poorly, or is otherwise not confident; instead, the student can simply fall back onto the one-hot training labels themselves. Formally, for *per-class* mixing weights $(\alpha_1, \ldots, \alpha_L) \in [0, 1]^L$, the student can minimise

$$\bar{R}_{\text{dist}}(f) = \frac{1}{N} \sum_{n=1}^{N} \Big[(1 - \alpha_{y_n}) \cdot \ell(y_n, f(x_n)) + \alpha_{y_n} \cdot \sum_{y' \in [L]} p^{\text{t}}_{y'}(x) \cdot \ell(y', f(x_n)) \Big]. \tag{2}$$

This objective introduces a mixing weight α_y per-class, which allows us to weigh between teacher predictions and one-hot labels for each class independently. By contrast, in the standard distillation setup equation 1 we only have a single weight α that is common for all classes.

How do we choose the weights α_y? In the standard distillation objective equation 1, one only needs to tune a single scalar α, which is amenable to, e.g., cross-validation. By contrast, equation 2 involves a single scalar for each label, which makes any attempt at grid search infeasible. Following the observations in §3.6, we propose the following intuitive setting of α_y given teacher predictions p^{t}:

$$\alpha_y = \max\left(0, \mathbb{E}_{x|y}\left[\gamma_{\text{avg}}(y, p^{\text{t}}(x))\right]\right) \quad \gamma_{\text{avg}}(y, p^{\text{t}}(x)) \doteq p^{\text{t}}_y(x) - \frac{1}{L-1} \sum_{y' \neq y} p^{\text{t}}_{y'}(x). \tag{3}$$

In words, Equation 3 places greater faith in the teacher model for those classes which it predicts *correctly* with *confidence*, i.e., with large average margin γ_{avg}. When this margin is *negative* — so that the teacher is *incorrect* on average, which can occur on classes that are rare in the training set — we set $\alpha_y = 0$, and completely ignore the teacher predictions. The above requires estimating the expectation $\mathbb{E}_{x|y}[\cdot]$, which requires access to a labelled sample. This may be done using a holdout set; we shall follow this in our subsequent experiments.

4.2 Distillation with per-class margins

Our second approach for improving distillation on harder classes is to leverage recent developments in *long-tail learning*, where the goal is to improve performance on rare classes. Specifically, Khan et al. (2018); Cao et al. (2019); Tang et al. (2020); Ren et al. (2020); Menon et al. (2020); Wang et al. (2021) proposed a variant of the softmax cross-entropy with *margins* $\rho_{yy'}$ between label pairs:

$$\ell(y, f(x)) = \log \left[1 + \sum_{y' \neq y} \rho_{yy'} \cdot e^{f_{y'}(x) - f_y(x)} \right]. \tag{4}$$

Intuitively, this penalises predicting label y' instead of y when $\rho_{yy'}$ is large. For training label distribution π, Cao et al. (2019) proposed to set $\rho_{yy'} \propto \exp(\pi_y^{-1/4})$, so that rare labels receive a higher weight when misclassified. Khan et al. (2018); Ren et al. (2020); Menon et al. (2020); Wang et al. (2021) showed gains with $\rho_{yy'} \propto \frac{\pi_{y'}}{\pi_y}$, so that rare labels are not confused with common ones.

We adapt such techniques to our setting, with the intuition that we ought to increase the student penalty for misclassifying those "hard" classes that the teacher has difficulty modeling. We thus choose $\rho_{yy'} = \frac{\alpha_{y'}}{\alpha_y}$, where α_y is the adaptive per-class mixing weight from the previous section. This discourages the model from confusing "hard" labels y with "easy" labels y', when $\alpha_{y'} > \alpha_y$. To avoid a division by 0 issue, we add a small offset to α_y when it becomes 0.

We may understand the effect of Equation 4 by studying how it impacts the *Bayes-optimal* student model predictions, i.e., the optimal predictions in the infinite sample limit, and without a model capacity restriction. We have the following.

Lemma 1. *Let ℓ be the loss of Equation 4, with $\rho_{yy'} \doteq \frac{\alpha_{y'}}{\alpha_y}$. Let f^* be the Bayes-optimal student scores that minimise $R_{\text{dist}}(f) \doteq \mathbb{E}_x (p^{\text{t}}(x))^\top \ell(f(x))$. Then, $\forall x \in \mathcal{X}, y \in [L]$, $f^*_y(x) = \log \frac{p^{\text{t}}_y(x)}{\alpha_y}$.*

Table 3: Summary of student's average accuracy using one-hot and distilled labels. *Worst* k denotes accuracy over the worst k classes. Global and adaptive temperatures α_y selected using a held out dev set. The proposed AdaAlpha technique improves both mean and worst class accuracy over vanilla distillation. For AdaMargin on CIFAR-100 LT and ImageNet LT, we observed divergence during training, presumably due to this method being sensitive to the selection of hyperparameters, which in turn are estimated on very small number of examples per class.

Dataset	**Method**	**Per-class accuracy statistics**		
		Mean	**Worst-1**	**Worst-10**
CIFAR-100	One-hot	73.31	45.67	52.12
	Distillation	75.24	49.00	54.65
	AdaAlpha	75.43	49.33	56.42
	AdaMargin	75.15	51.33	56.62
ImageNet	One-hot	76.38	14.00	23.64
	Distillation	76.35	13.00	22.02
	AdaAlpha	76.57	13.00	23.22
	AdaMargin	76.36	14.00	23.20

Dataset	**Method**	**Per-class accuracy statistics**			
		Mean	**Worst-1**	**Worst-10**	**Worst-100**
CIFAR-100 LT	One-hot	43.22	0.00	2.33	N/A
	Distillation	45.39	0.00	0.87	N/A
	AdaAlpha	48.57	0.67	4.20	N/A
	AdaMargin*	Training diverges			
ImageNet LT	One-hot	45.41	0.00	0.00	1.39
	Distillation	45.98	0.00	0.00	1.10
	AdaAlpha	46.15	0.00	0.00	1.08
	AdaMargin*	Training diverges			

Lemma 1 illustrates that using per-class margins encourages the student to mimic the teacher predictions $p^{\mathrm{t}}(x)$, but with an important modification: we up-weight the probabilities for classes that the teacher does poorly on ($\alpha_y \sim 0$). Intuitively, this makes it easier for the student to improve performance on classes with small teacher margin.

Relation to existing work. Previous works varied distillation supervision across examples towards improving average accuracy. Proposals included weighting samples based on the ratio (Tang et al., 2019; Zhou et al., 2021), and difference (Zhang et al., 2020) between student and teacher score. Similarly, Zhou et al. (2020) proposed to only apply distillation on samples the teacher gets correct.

5 Results for adaptive distillation methods

We now present results that further corroborate the potential non-uniform gains of distillation, and the ability to mitigate this with the techniques of the previous section. We emphasise here that our goal is expressly *not* to improve over the state-of-the-art in distillation techniques; rather, we wish to verify the key principles identified in the preceding study, which considers distillation from a novel angle (i.e., in terms of subgroup rather than average performance).

Setup. We report results on the datasets used in §3: CIFAR-100, ImageNet; and long-tailed (**LT**) versions of the same. For brevity, we report results under self-distillation. (For results with varying architectures, see the Appendix.) Thus, for each dataset, we train a one-hot teacher ResNet model, which is distilled to a student ResNet of the same depth. We use ResNet-56 models for CIFAR, and ResNet-50 models for all other datasets. We employ the same hyper-parameters as used in §3, except we use non-early stopped teachers for consistency across datasets; see the Appendix for details.

We compare: (i) one-hot training of the student (ii) standard distillation, i.e., minimising Equation 1 (iii) **AdaAlpha**, our proposed distillation objective with adaptive mixing between one-hot and teacher labels Equation 2, and α as per Equation 3 (iv) **AdaMargin**, our proposed distillation objective with adaptive margins (Equation 4), and $\rho_{yy'} = \frac{\alpha_{y'}}{\alpha_y}$. We summarise each method by reporting: (i) the standard *mean* accuracy over all classes; (ii) the accuracy over the *worst*-1 class; and (iii) the mean accuracy over *worst*-10 (and *worst*-100 for the LT datasets) classes.

For the Ada-* methods, per §4, creating the label-dependent α_y requires estimating the generalisation performance of the teacher. To do this, we create a random holdout split of the training set. For non-LT datasets, we randomly split into 80% (new train) – 20% (dev). For LT datasets, for each class we hold out k examples into the dev set ($k = 50$ for Imagenet-LT, $k = 20$ for CIFAR-100-LT), or half of examples for a class if the total number of per class examples is at most $2k$. We train an initial teacher on the new train slice of data, and estimate its per-class performance on the holdout dev slice. These are used to estimate α_y as per, e.g., Equation 3.

Table 3 summarises the results for all methods. We make the following observations.

Table 4: Ablations of design choices in the proposed methods: 1) remove distillation signal from the bottom 10% of classes, according to confidence; 2) randomly shuffle per-class α values; 3) weight distillation based on student and teacher confidence Zhou et al. (2021).

Dataset	Method	Per-class accuracy statistics		
		Mean	Worst-1	Worst-10
CIFAR-100	AdaAlpha	75.52 ± 0.10	49.33 ± 3.09	56.59 ± 0.44
	remove hardest 10%	75.40 ± 0.04	48.33 ± 1.89	55.79 ± 1.19
	shuffle temperatures	74.56 ± 0.93	46.00 ± 1.91	53.10 ± 1.22
	Zhou et al. (2021)	75.14 ± 0.29	45.11 ± 1.66	53.89 ± 0.66

AdaAlpha improves mean accuracy over vanilla distillation. The proposed AdaAlpha method consistently and significantly improves standard mean accuracy over vanilla distillation. Thus, AdaAlpha does not sacrifice the gains offered by distillation on average class performance, which is desirable. Other techniques sometimes perform slightly worse than standard distillation on this metric; however, as we now see, this is compensated by gains on other important dimensions.

AdaAlpha improves worst-accuracy over distillation. The proposed method consistently improves the worst-class accuracy compared to standard distillation: Thus, the technique largely fulfil their design goal of improving performance on "hard" classes, while not overly sacrificing average-case performance. In most cases, these improve *both* the average and worst-class accuracy, indicating that softening the teacher influence can be broadly beneficial.

Comparison of AdaAlpha and AdaMargin. In the Appendix, we report per class statistics for CIFAR-100 LT. AdaMargin flattens both the margin and log-loss distributions, reducing confidence on the poorly classified, tail classes. AdaAlpha consistently increases log-loss across classes, and improves margins on few buckets, leading to a positive margin on one bucket where all other methods give negative margins. Intuitively, AdaMargin tries to more aggressively control the worst-class accuracy; when this succeeds, there is a large payoff, but there is also greater risk of overfitting.

Additional ablations. We confirm that the success of AdaAlpha is not immediately replicated by simpler baselines: (i) *remove hardest 10%*, which removes the distillation loss component on bottom 10% labels according to the per class margins found using Equation 3. It helps analyze whether there is any additional gain beyond simply removing teacher's supervision where it is arguably wrong. (ii) *shuffle temperatures*, which randomly shuffles the per-class α_y values used in AdaAlpha. This determines whether the precise choice of which labels to up- or down-weight is important; (iii) the adaptive distillation scheme of Zhou et al. (2021), where distillation is weighted differently across examples depending on the teacher and student scores.

In Table 4, we find that the first two methods work worse than the proposed AdaAlpha method, indicating that the precise choice of which labels to up- or down-weight is important, and that it does not suffice to merely ignore the teacher on entire subgroups. The adaptive distillation scheme Zhou et al. (2021) is also not as effective as AdaAlpha; see Appendix for more such results.

6 Discussion and other approaches

Our goal of ensuring equitable performance across classes can be seen as encouraging *fairness* across subgroups defined by the classes. This is subtly different to the classical fairness literature (Calders & Verwer, 2010; Dwork et al., 2012; Hardt et al., 2016), wherein the subgroups are defined by certain *sensitive attributes*. Broadly, fairness techniques attempt to learn models that predict the target label *accurately*, but the subgroup label *poorly*; these are inadmissible for our setting, wherein the two labels exactly coincide. Ensuring fairness across subgroups defined by the classes has been studied in Mohri et al. (2019); Williamson & Menon (2019); Sagawa et al. (2020a), who proposed algorithms to explicitly minimise the *worst-class* (as opposed to the average) loss. Adapting such algorithms to the distillation setting is of interest for future work. More broadly, the intent of the analysis in this paper is to better understand settings where distillation can implicitly hurt certain under-represented subgroups.

Ethics statement

This work studies settings wherein distillation can harm performance on subgroups. The aim of such a study is to help identify and mitigate potential adverse effects from distillation in practical applications where fairness considerations are important. Further, our proposed techniques were shown to improve the subgroup performance of distillation, and thus could be useful for preventing such adverse behaviour. As a limitation, we considered natural definitions of such subgroups, but an important practical consideration is settings where subgroup membership is unknown.

Reproducibility statement

For all experiments, we used the default hyperparameters from previous works (see Appendix B for details).

References

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *CoRR*, abs/2012.09816, 2020. URL https://arxiv.org/abs/2012.09816.

Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E. Dahl, and Geoffrey E. Hinton. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*, 2018.

Cristian Bucilă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, New York, NY, USA, 2006. ACM.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *arXiv:1710.05381 [cs, stat]*, October 2017.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Conference on Fairness, Accountability, and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

Toon Calders and Sicco Verwer. Three Naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32*, pp. 1565–1576, 2019.

J. H. Cho and B. Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801, 2019.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 12760–12770. Curran Associates, Inc., 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=m4UCf24r0Y.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 214–226, 2012.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 1602–1611, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, December 2016.

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*, pp. 770–778, 2016.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Sara Hooker, Aaron C. Courville, Yann N. Dauphin, and Andrea Frome. Selective brain damage: Measuring the disparate impact of model pruning. *CoRR*, abs/1911.05248, 2019. URL http://arxiv.org/abs/1911.05248.

Guangda Ji and Zhanxing Zhu. Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems*, 2020.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations*, 2021.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020.

Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2018. doi: 10.1109/TNNLS.2017.2732482.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1–50, 02 2002.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2537–2546. Computer Vision Foundation / IEEE, 2019.

D. Lopez-Paz, B. Schölkopf, L. Bottou, and V. Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*, November 2016.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

Aditya Krishna Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 603–611, 2013.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment, 2020.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in hilbert space, 2020.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems 32*, pp. 4696–4705, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0607f4c705595b911a4f3e7a127b44e0-Paper.pdf.

Ilija Radosavovic, Piotr Dollár, Ross B. Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4119–4128, 2018.

Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4175–4186. Curran Associates, Inc., 2020.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020a.

S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning (ICML)*, 2020b.

Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 10976–10987. Curran Associates, Inc., 2018.

N. Sohoni, J. Dunnmon, G. Angus, A. Gu, and C. Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *To appear in Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021.

Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H. Chi, and Sagar Jain. Understanding and improving knowledge distillation. *CoRR*, abs/2002.03532, 2020.

Shitao Tang, Litong Feng, Wenqi Shao, Zhanghui Kuang, Wayne Zhang, and Zheng Lu. Learning efficient detector with semi-supervised adaptive distillation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, pp. 215. BMVA Press, 2019. URL https://bmvc2019.org/wp-content/uploads/papers/0145-paper.pdf.

Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, United States, December 2018. Institute of Electrical and Electronics Engineers (IEEE). ISBN 978-1-5386-6421-6. doi: 10.1109/CVPR.2018.00914. URL http://cvpr2018.thecvf.com/. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018 ; Conference date: 18-06-2018 Through 22-06-2018.

Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Robert C. Williamson and Aditya Krishna Menon. Fairness risk measures. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 6786–6797, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei. Prime-aware adaptive distillation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 658–674. Springer International Publishing, 2020. ISBN 978-3-030-58529-7.

Zhilu Zhang and Mert R. Sabuncu. Self-distillation as instance-specific label smoothing. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13205–13214. IEEE, 2020.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias–variance tradeoff perspective. In *International Conference on Learning Representations*, 2021.

Zaida Zhou, Chaoran Zhuge, Xinwei Guan, and Wen Liu. Channel distillation: Channel-wise attention for knowledge distillation. *CoRR*, abs/2006.01683, 2020. URL https://arxiv.org/abs/2006.01683.

A Proofs

Proof of Lemma 1. We may write

$$\begin{aligned}
\ell(y, f(x)) &= \log \left[1 + \sum_{y' \neq y} \frac{\alpha_{y'}}{\alpha_y} \cdot e^{f_{y'}(x) - f_y(x)} \right] \\
&= \log \left[1 + \sum_{y' \neq y} e^{\ln \alpha_{y'} - \ln \alpha_y} \cdot e^{f_{y'}(x) - f_y(x)} \right] \\
&= \log \left[1 + \sum_{y' \neq y} e^{\bar{f}_{y'}(x) - \bar{f}_y(x)} \right] \\
&= -\log \frac{e^{\bar{f}_y(x)}}{\sum_{y' \in [L]} e^{\bar{f}_{y'}(x)}},
\end{aligned}$$

where $\bar{f}_y(x) \doteq f_y(x) + \ln \alpha_y$. The population loss is

$$\begin{aligned}
R_{\text{dist}}(f) &= \mathbb{E}_x \left[(p^{\text{t}}(x))^\top \ell(f(x)) \right] \\
&= \mathbb{E}_x \left[\text{KL}(p^{\text{t}}(x) \,\|\, \bar{p}^{\text{s}}(x)) \right],
\end{aligned}$$

where $\bar{p}^{\text{s}}_y(x) \propto \exp(\bar{f}_y(x))$. Thus, at optimality we must have $\bar{p}^{\text{s}}(x) = p^{\text{t}}(x)$, or $\bar{f}_y(x) = \log p^{\text{t}}_y(x)$. By definition of $\bar{f}$, we thus see that $f_y(x) = \log p^{\text{t}}_y(x) - \log \alpha_y = \log \frac{p^{\text{t}}_y(x)}{\alpha_y}$. □

B Details of experiments

B.1 Architecture

We use ResNet with batch norm (He et al., 2016) for all our experiments with the following configurations. For CIFAR, we experiment with ResNet-56 and ResNet-32. For ImageNet, we use ResNet-50. We list the architecture configurations in terms of (n_{layer}, n_{filter}, stride) corresponding to each ResNet block in Table 5.

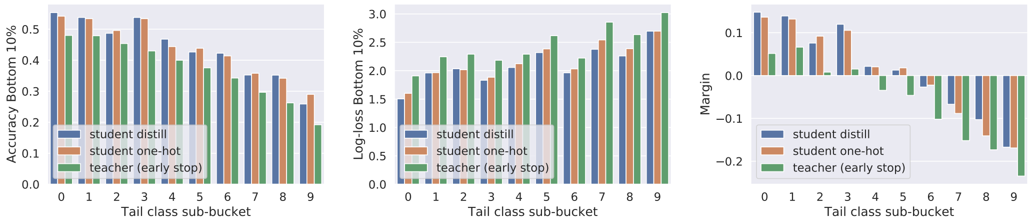
Architecture	**Configuration: [($\mathbf{n}_{\text{layer}}$, $\mathbf{n}_{\text{filter}}$, stride)]**
CIFAR ResNet-32	[(5, 16, 1), (5, 32, 2), (5, 64, 2)]
CIFAR ResNet-56	[(9, 16, 1), (9, 32, 2), (9, 64, 2)]
ImageNet ResNet-18	[(2, 64, 1), (2, 128, 2), (2, 256, 2), (2, 512, 2)]
ImageNet ResNet-34	[(3, 64, 1), (4, 128, 2), (6, 256, 2), (3, 512, 2)]
ImageNet ResNet-50	[(3, 64, 1), (4, 128, 2), (6, 256, 2), (3, 512, 2)]*

Table 5: ResNet Architecture configurations used in our experiments (He et al., 2016). [*] Note that ImageNet ResNet-50 uses larger blocks with 3 convolutional layers per residual block compared to ResNet-18 and 34. We refer to He et al. (2016) for more details.

B.2 Training set

For all datasets, we train using SGD and weight decay 10^{-4} for CIFAR, and 0.5×10^{-4} for Imagenet datasets. We have the following dataset specific settings.

CIFAR-100. We train for 450 epochs with an initial learning rate of 1.0, with a linear warmup in the first 15 epochs, and an annealed learning rate schedule. We drop the learning rate by a factor of 10 at epochs number: 200, 300 and 400. We use a mini-batch size of 1024. We use SGD with Nesterov momentum of 0.9.



(a) Accuracy (tail classes).

(b) Log-loss (tail classes).

(c) Margin (tail classes).

Figure 4: Logit statistics for ResNet-50 self-distillation on ImageNet, for the (early-stopped) teacher, self-distilled student, and one-hot (non-distilled) student. Per Figure 3, we first create 10 class buckets. We zoom in on the "tail" bucket (comprising the 100 "hardest" classes), and further split them into 10 "tail sub-buckets". As in Figure 3, the teacher is seen to confidently mispredict most samples on the last few buckets, with such misplaced confidence being transferred to the student.

For our distillation experiments we train only with the cross-entropy objective against the teacher's logits. For each method we find the best temperature from the list of values: $\{1, 2, 3, 4, 5\}$.

ImageNet. We train for 90 epochs with an initial learning rate of 0.8, with a linear warmup in the first 5 epochs, and an annealed learning rate schedule. We drop the learning rate by a factor of 10 at epochs number: 30, 60 and 80. We use a mini-batch size of 1024.

For our distillation experiments we train with the distillation objective as defined in Equation 1 setting $\alpha = 0.2$. For each method we fix the temperature to 0.9.

Long-tail (LT) datasets. We follow setup as in the non-long tail version, except for the learning rate schedule, which we change to follow the cosine schedule (Loshchilov & Hutter, 2017).

C Additional Experiments

We present additional experiments to those in the body.

C.1 Further varying datasets and model architectures

On Imagenet, we summarise statistics for *three* models: the early-stopped teacher, distilled student, and the one-hot (non-distilled) student. As with CIFAR-100-LT, we sort classes by teacher accuracy, and bucket them into 10 groups. Owing to the larger number of labels, we further zoom into the "tail" bucket (comprising the 100 "hardest" classes), and split them into 10 sub-buckets. From Figure 4, the distilled student performs worse than its one-hot counterpart on the last bucket; this is in keeping with our results in Table 1.

Figure 5 shows logit statistics for additional settings to considered in the body. On ImageNet-LT, e.g., we see again that the margin of the teacher model systematically worsens and becomes negative on the hardest classes.

In Table 6 we report results from the inherently long-tailed iNaturalist 2018 dataset (Van Horn & Perona, 2017). Our observations made for other considered datasets hold: adaptive margin method improves over both one hot and plain distillation in terms of the worst class accuracy. We also observe, how the average accuracy improves.

C.2 Logit plots under Ada-* methods

Figure 7 shows the logit statistics under the proposed AdaMargin and AdaAlpha methods on CIFAR-100 LT. We see that AdaMargin can generally improve the student margin and accuracy on the hardest classes, while also reducing the log-loss. This confirms that the gains of the method come from improving behaviour of the scores on these hard classes.

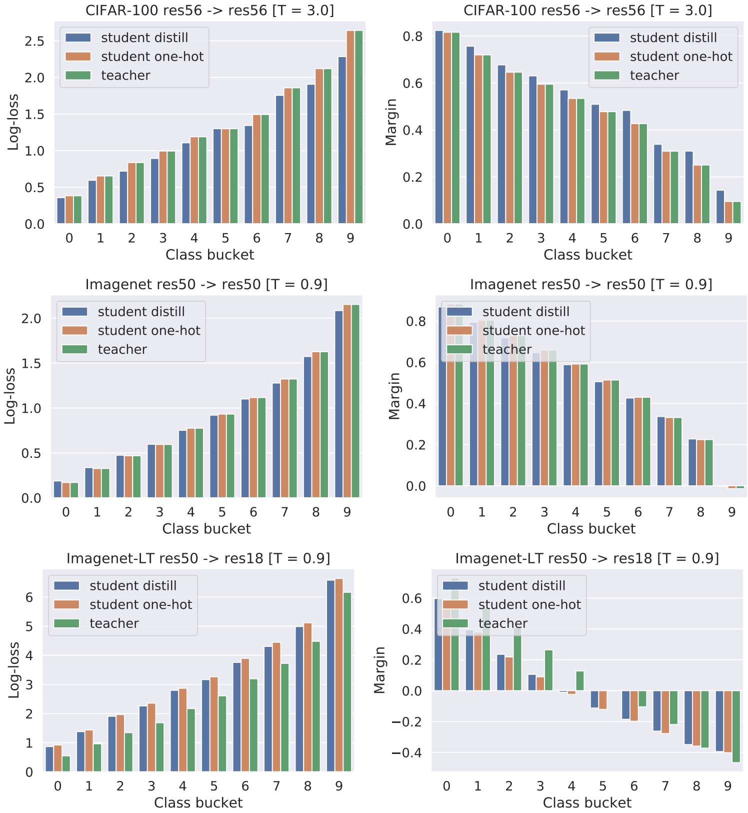


Figure 5: Logit statistics for the teacher, student with one-hot labels, and student with distilled labels across: datasets and architectures.

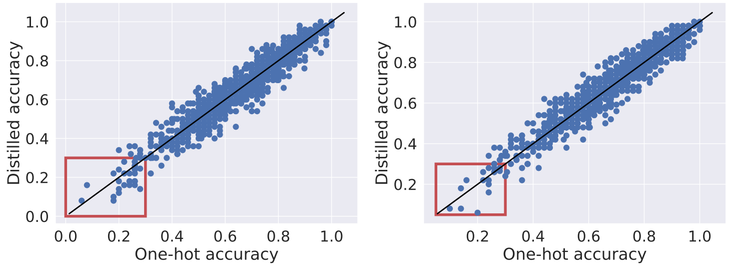


Figure 6: Per-class accuracies for one-hot and self-distilled ResNet-18 (left) and ResNet-34 (right) on ImageNet. The diagonal denotes classes where both models achieve the same accuracy. Distillation tends to worsen performance on "hard" classes for the one-hot model, i.e., those with low accuracy (red rectangle).

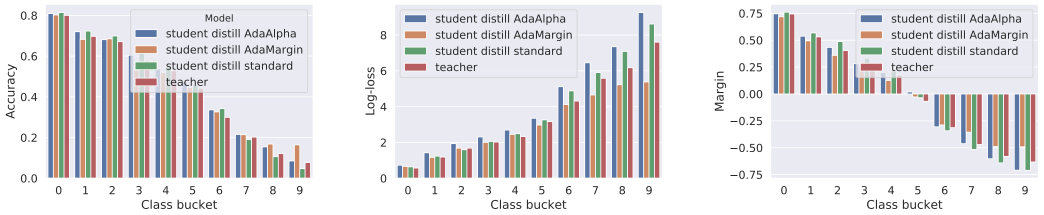
C.3 Results on Adult dataset

We report the results of an experiment on the UCI Adult dataset. This data comprises $\sim 48K$ examples, with the target being a binary label denoting whether or not an individual has income $\geq$ 50K. The data is mildly imbalanced, with 24% of samples being positive.

Inspired by Dao et al. (2021), we consider a random forest based distillation setup: we use a teacher model that is a random forest *classifier* comprising 500 trees with a maximum depth of 20, and a student model that is a random forest *regressor* comprising 1 tree with a maximum depth of 20. The teacher model achieves a test (balanced) accuracy of 81.8%.

Method	**Per-class accuracy statistics**			
	Mean	**Worst 20**	**Top 20%**	**Δ20**
One-hot	53.00	5.00	100.00	48.00
Distill	52.67	5.00	100.00	47.67
AdaMargin	53.33	8.33	98.33	45.00
AdaAlpha avg	54.67	13.33	100.00	41.33

Table 6: Self-distillation experiments (from Resnet-50 to Resnet-50) on the iNaturalist dataset Van Horn et al. (2018) with student's average accuracy using one-hot and distilled labels. Worst 20 denotes accuracy averaged over worst 20 classes Δ20 denotes the difference between the mean accuracy and the worst 20 classified classes. The proposed AdaMargin technique improves mean and worst-class accuracy over both one-hot training and standard distillation.



(a) Accuracy.

(b) Log-loss.

(c) Margin).

Figure 7: Logit statistics for ResNet-56 self-distillation on CIFAR-100 LT, for the teacher, self-distilled student, and our adaptive methods. Per Figure 3, we create 10 class buckets. AdaMargin flattens both the margin and log-loss distributions. AdaAlpha increases log loss across classes, while improving margins on few buckets, including flipping the bucket *5* to have positive margin.

Table 7: Difference between distillation and one-hot performance on Adult dataset. Here, subgroups are defined by the sex and label. Δ refers to the difference between the distilled and one-hot student's accuracy on the subgroup.

sex	label	Δ
Male	0	-2.222
Female	0	0.393
Female	1	2.373
Male	1	8.384

We perform distillation by feeding the student model the teacher's prediction scores, mixed in with the binary training labels with a weight $\alpha = 0.9$. Distillation improves the student's overall (balanced) accuracy significantly, from 76.2% to 79.3%. However, this gain is not distributed uniformly: using per-label subgroups, we find that distillation helps the positive class by $+7.4\%$, but hurts the negative class by -1.2%. While by itself suggestive of asymmetry in distillation performance, the data admits an arguably more natural subgroup creation, based on available sex and sex features. For example, we find that amongst low-income males, distillation hurts by -2.2%; further restricting to those who are Asian Pacific-Islander, the degradation is -5.9%. This confirms that in scenarios where fairness may be a consideration, a naïve application of distillation may be inadmissible.

C.4 Analysis of regularisation samples

Recently, Zhou et al. (2021) proposed the notion of *regularisation samples* to understand how distillation's performance can be improved. In brief, such samples correspond to cases where the teacher's prediction on the training label is less than the distilled student's prediction on this label; these may be shown to correspond to cases where a certain notion of "variance reduction" dominates

Table 8: Difference between distillation and one-hot performance on Adult dataset. Here, subgroups are defined by the sex, race, and label. Δ refers to the difference between the distilled and one-hot student's accuracy on the subgroup.

race	sex	label	Δ
Amer-Indian-Eskimo	Female	1	-66.667
Asian-Pac-Islander	Male	0	-5.941
Other	Male	0	-4.347
Black	Female	1	-2.381
White	Male	0	-2.248
Black	Male	0	-1.639
Black	Female	0	-0.140
Asian-Pac-Islander	Female	0	0.000
White	Female	0	0.388
Other	Female	0	2.439
White	Female	1	2.724
Amer-Indian-Eskimo	Female	0	6.349
Amer-Indian-Eskimo	Male	0	6.493
Asian-Pac-Islander	Female	1	7.692
White	Male	1	7.796
Black	Male	1	9.489
Other	Male	1	15.000
Asian-Pac-Islander	Male	1	19.626
Other	Female	1	20.000
Amer-Indian-Eskimo	Male	1	25.000

a notion of "bias reduction". Given our analysis above of the asymmetric effects of distillation on certain subgroups, it is natural to consider whether or not these relate to the presence of regularisation samples in these groups.

Figure 8(a) visualises the distribution of regularisation samples inside subgroups defined by 10 label buckets. where the labels are sorted in descending order of label frequency. Here, we compare the predicted probabilities of the teacher and final distilled student models on all *training* samples (as was done in the analysis of Zhou et al. (2021)). Interestingly, we see that the tail buckets tend to have very few regularisation samples; i.e., for rare labels, the teacher prediction on the training label is generally higher than that of the distilled student model. We confirm this in Figure 8(b).

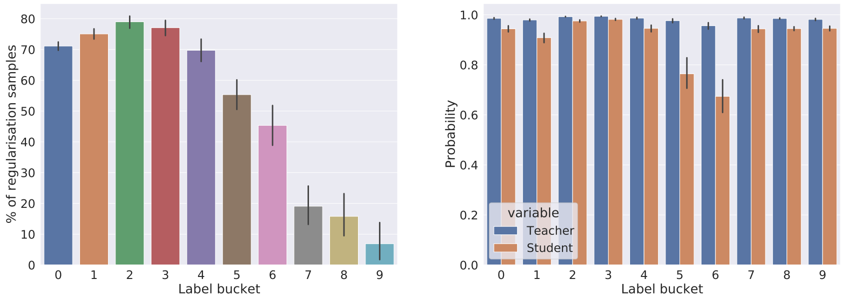
While the analysis of Zhou et al. (2021) was primarily for training samples — since the aim in identifying regularisation samples was to mitigate their influence during training — we may also identify the breakdown of such samples on test data. Figure 9(a) shows that, compared to the training set, there are in absolute terms more such samples across nearly every label bucket; however, there is again no clear correlation between the label bucket and the fraction of such samples. In particular, the tail bucket is again the one with the *fewest* regularisation samples. This is corroborated by the probability scores of the teacher and student in Figure 9(b).

Overall, this results suggest that the existing notion of regularisation samples may not, by themselves, be sufficient to predict the poor performance of distillation on certain subgroups defined by labels.

C.5 Impact of repeated distillation

In the body, we showed that performing distillation once can harm worst-class accuracy. However, what is the effect of repeating this process, and distilling using the resulting student as a new teacher? Does the worst-class accuracy get further harmed?

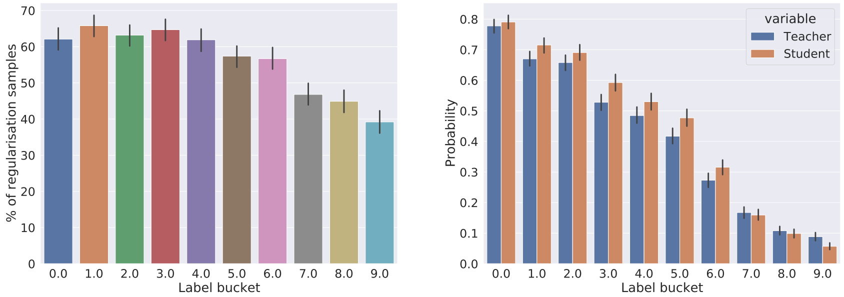
Table 9 shows that on CIFAR-100 LT, repeating distillation can indeed harm worst-class performance, even though average performance remains roughly similar. This further highlights the potential tension between average and worst-case performance under distillation.



(a) Distribution of regularisation samples per label bucket.

(b) Teacher versus distilled student probailities on training label.

Figure 8: Study of regularisation samples on training set, CIFAR-100 LT.



(a) Distribution of regularisation samples per label bucket.

(b) Teacher versus distilled student probailities on test label.

Figure 9: Study of regularisation samples on test set, CIFAR-100 LT.

Method	**Per-class accuracy statistics**			
	Mean	**Worst 10**	**Top 10%**	**Δ10**
One-hot	44.16	3.00	87.70	41.16
Distillation 1×	45.49	0.90	88.10	44.59
Distillation 2×	45.22	0.00	88.40	45.22
Distillation 3×	44.80	0.00	87.60	44.80

Table 9: Results of repeated distillation on CIFAR-100 LT. Using a distilled student as teacher for a subsequent round of distillation is seen to further hurt worst-class accuracy.