Combating Label Noise in Deep Learning Using Abstention

Sunil Thulasidasan [1 2] **Tanmoy Bhattacharya** [1] **Jeffrey Bilmes** [2] **Gopinath Chennupati** [1] **Jamaludin Mohd-Yusof** [1]

Abstract

We introduce a novel method to combat label noise when training deep neural networks for classification. We propose a loss function that permits abstention during training thereby allowing the DNN to abstain on confusing samples while continuing to learn and improve classification performance on the non-abstained samples. We show how such a deep abstaining classifier (DAC) can be used for robust learning in the presence of different types of label noise. In the case of structured or systematic label noise – where noisy training labels or confusing examples are correlated with underlying features of the data– training with abstention enables representation learning for features that are associated with unreliable labels. In the case of unstructured (arbitrary) label noise, abstention during training enables the DAC to be used as an effective data cleaner by identifying samples that are likely to have label noise. We provide analytical results on the loss function behavior that enable dynamic adaption of abstention rates based on learning progress during training. We demonstrate the utility of the deep abstaining classifier for various image classification tasks under different types of label noise; in the case of arbitrary label noise, we show significant improvements over previously published results on multiple image benchmarks.

1 Introduction

The impressive performance of deep neural networks in recent years in various tasks such as image classification and speech recognition (LeCun et al., 2015) have been made possible by the availability of large quantities of human-annotated (i.e., labeled) training data. For example, the ImageNet database (Deng et al., 2009) used for training vision classifiers has now over 14 million hand annotated images(ImageNet). Though enabling deep models to match or surpass human performance, the requirement of vast quantities of human-annotated datasets can often be a bottleneck in training deep learning systems. There are generally two approaches to tackle this problem: use a scalable and affordable annotation platform like Amazon Mechanical Turk (Turk) or alternatively, automatically collect large amounts of web-based data that have associated meta-information (tags, for instance) which are used for labeling. In both cases, a certain amount of erroneously labeled data is bound to occur, often in significant fractions for the latter situation (Li et al., 2017a). While there is empirical evidence that deep networks are robust to some amount of label noise (Rolnick et al., 2017), significant label noise can degrade generalization performance due to the ability of deep models to fit random labels (Zhang et al., 2016). In such situations, it is often better to eliminate the noisy data and train with just the cleaner subset (Frénay & Verleysen, 2014), or use a semi-supervised approach (Chapelle et al., 2009) which uses all the samples, but only retains the labels in the cleaner set for training. In both cases, one needs to identify the subset of training data whose labels are likely to be unreliable.

While label noise is a well studied problem in machine learning (Nettleton et al., 2010; Zhu & Wu, 2004; Sukhbaatar et al., 2014; Reed et al., 2014; Patrini et al., 2017), there has not been much work in identifying and ignoring it during the process of training itself. In this paper we propose a novel *abstention* (or rejection) based mechanism to combat label noise while training deep models. Most of the theoretical and empirical investigations into rejection classification systems have been studied in a post-processing setting – i.e., a classifier is first trained as usual, and an abstention threshold is determined based on post-training performance on a calibration set; the DNN then abstains on uncertain predictions during inference. Abstention classifiers have been proposed for shallow learners (Chow, 1970; Cortes et al., 2016; Fumera & Roli, 2002) and for multilayer perceptrons (De Stefano et al., 2000). In the context of deep networks, this has been an under-explored area with (Geifman & El-Yaniv, 2017) recently proposing an effective technique of selective classification for optimizing

[1]Los Alamos National Laboratory, Los Alamos, NM, USA [2]Department of Electrical & Computer Engineering, University of Washington, Seattle, WA, USA. Correspondence to: Sunil Thulasidasan <sunil@lanl.gov>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

risk-vs-coverage profiles based on the output of a trained model.

In contrast to the above, the approach described in this paper employs abstention during *training* as well as inference. This gives the DNN an option to abstain on a confusing training sample thereby mitigating the misclassification loss but incurring an abstention penalty. We empirically show that our formulation ensures that the DNN continues to learn the true class even while abstaining, progressively reducing its abstention rate as it learns on the true classes and finally abstaining on only the most confusing samples. We demonstrate the advantages of such a formulation under two different label-noise scenarios: first, when labeling errors are correlated with some underlying feature of the data (systematic or structured label noise), abstention training allows the DNN to learn features that are indicative of unreliable training signals which are thus likely to lead to uncertain predictions. This kind of representation learning for abstention is useful both for effectively eliminating structured noise and also for interpreting the reasons for abstention. Second, we show how an abstention-based approach can be used as an effective *data cleaner* when training data contains arbitrary (or unstructured) label noise. A DNN trained with an abstention option can be used to identify and filter out noisy training data leading to significant performance benefits for downstream training using a cleaner set. To summarize, the contributions of this paper are:

- The introduction of abstention-based training as an effective approach to combat label noise.We show that such a deep abstaining classifier (DAC) enables robust learning in the presence of label noise.
- The demonstration of the ability of the DAC to learn features associated with systematic label noise. Through numerous experiments, we show how the DAC is able to pick up (and then abstain on) such features with high precision.
- Demonstration of the utility of the DAC as an effective *data cleaner* in the presence of arbitrary label noise. We provide results on learning with noisy labels on multiple image benchmarks (CIFAR-10, CIFAR-100 and Fashion-MNIST) that improve upon existing methods. Our method is also considerably simpler to implement and can be used with any existing DNN architecture as only the loss function is changed.

While ideally such an abstaining classifier should also learn to reliably abstain when presented with adversarially perturbed samples (Nguyen et al., 2015; Szegedy et al., 2013; Moosavi-Dezfooli et al., 2017), in this work we do not consider adversarial settings and leave that for future exploration. The rest of the paper is organized as follows: Section 2 describes the loss function formulation and an algorithm for automatically tuning abstention behavior. Section 3 discusses learning in the presence of structured noise, including experimental results and visual interpretations of abstention. Section 4 presents the utility of the DAC for data cleaning in the presence of unstructured (arbitrary) noise. Section 5 has further discussions on abstention behavior in the context of memorization. We conclude in Section 6.

2 Loss Function for the Deep Abstaining Classifier

We assume we are interested in training a k-class multi-class classifier with a deep neural network (DNN) where x is the input and y is the output. For a given x, we define $p_i = p_w(y = i|x)$ (the probability of the ith class given x) as the i^{th} output of the DNN that implements the probability model $p_w(y = i|x)$ where w is the set of weight matrices of the DNN. For notational brevity, we use p_i in place of $p_w(y = i|x)$ when the input context x is clear.

The standard cross-entropy training loss for DNNs then takes the form $\mathcal{L}_{\text{standard}} = -\sum_{i=1}^{k} t_i \log p_i$ where t_i is the target for the current sample. The DAC has an additional $k{+}1^{\text{st}}$ output p_{k+1} which is meant to indicate the probability of abstention. We train the DAC with following modified version of the k-class cross-entropy per-sample loss:

$$\mathcal{L}(x_j) = (1-p_{k+1}) \left(-\sum_{i=1}^{k} t_i \log \frac{p_i}{1-p_{k+1}} \right) + \alpha \log \frac{1}{1-p_{k+1}} \tag{1}$$

The first term is a modified cross-entropy loss over the k non-abstaining classes. Absence of the abstaining output (i.e., $p_{k+1} = 0$) recovers exactly the usual cross-entropy; otherwise, the abstention mass has been normalized out of the k class probabilities. The second term penalizes abstention and is weighted by $\alpha \geq 0$, a hyperparameter expressing the degree of penalty. If α is very large, there is a high penalty for abstention driving p_{k+1} to zero and recovering the standard unmodified cross-entropy loss; in such case, the model learns to never abstain. With α very small, the classifier may abstain on everything with impunity since the adjusted cross-entropy loss becomes zero and it does not matter what the classifier does on the k class probabilities. When α is between these extremes, things become more interesting: whether the DNN chooses to abstain or not depends on how much cross-entropy error it is making while learning on the true classes; it is this error that drives mass into the abstention class subject to the abstention penalty hyperparameter α.

Lemma 1. *For the loss function $\mathcal{L}$ given in Equation 1, if j is the true class for sample x, then as long as $\alpha \geq 0$, $\frac{\partial \mathcal{L}}{\partial a_j} \leq 0$ (where a_j is the pre-activation into the softmax unit of class j).*

The proof is given in Section A of the supplementary mate-

Algorithm 1 α auto-tuning

Input: total iter. (T), current iter. (t), total epochs (E), abstention-free epochs (L), current epoch (e), α init factor (ρ), final α (α_{final}), mini-batch cross-entropy over true classes ($\mathcal{H}_c(P^M_{1\ldots K})$

$\alpha_{set} = False$
for $t := 0$ to T **do**
  if $e < L$ **then**
    $\beta = (1 - P^M_{k+1})\mathcal{H}_c(P^M_{1\ldots K})$
    if $t = 0$ **then**
      $\tilde{\beta} = \beta${ // initialize moving average}
    end if
    $\tilde{\beta} \leftarrow (1 - \mu)\tilde{\beta} + \mu\beta$
  end if
  if $e = L$ and **not** α_{set} **then**
    $\alpha := \tilde{\beta}/\rho${ // initialize α at start of epoch L}
    $\delta_\alpha := \frac{\alpha_{final} - \alpha}{E - L}$
    $update_{epoch} = L$
    $\alpha_{set} = True$
  end if
  if $e > update_{epoch}$ **then**
    $\alpha \leftarrow \alpha + \delta_\alpha${ //then update α once every epoch}
    $update_{epoch} = e$
  end if
end for

rial. This ensures that, during gradient descent, learning on the true classes persists even in the presence of abstention, even though the true class might not end up be the winning class. We provide additional discussion on abstention behavior in Section 5.

2.1 Auto-tuning α

Let $g = -\sum_{i=1}^{k} t_i \log p_i$ be the standard cross-entropy loss and a_{k+1} be the pre-activation into the softmax unit for the abstaining class. Then it is easy to see that:

$$\frac{\partial \mathcal{L}}{\partial a_{k+1}} = p_{k+1}\left[(1 - p_{k+1})\left[\log \frac{1}{1 - p_{k+1}} - g\right] + \alpha\right]. \quad (2)$$

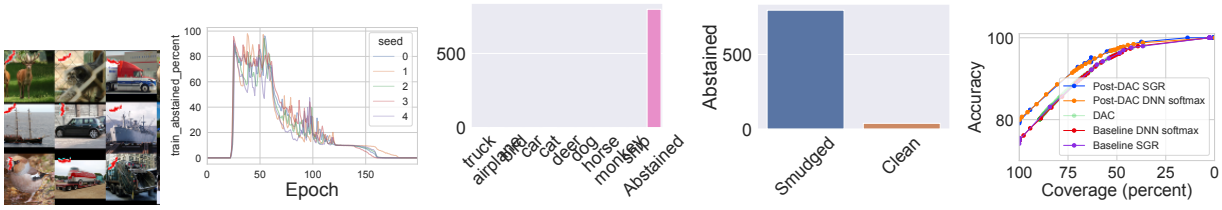
During gradient descent, abstention pre-activation is increased if $\frac{\partial \mathcal{L}}{\partial a_{k+1}} < 0$. The threshold on α for this is $\alpha < (1 - p_{k+1})\left(-\log \frac{p_j}{1-p_{k+1}}\right)$ where j is the true class for sample x. If only a small fraction of the mass over the actual classes is in the true class j, then the DAC has not learned to correctly classify that particular sample from class j, and will push mass into the abstention class provided α satisfies the above inequality. This constraint allows us to perform auto-tuning on α during training (Algorithm 1). $\tilde{\beta}$ is a smoothed moving average of the α threshold (initialized to 0), and updated at every mini-batch iteration. We perform abstention-free training for L initial epochs (a warm-up period) to accelerate learning, triggering abstention from epoch $L + 1$ onwards. At the start of abstention, α is initialized to a much smaller value than the threshold $\tilde{\beta}$ to encourage abstention on all but the easiest of examples learnt so far. As the learning progresses on the true classes, abstention is reduced. We linearly ramp up α over the remaining epochs (updating once per epoch)to a final value of α_{final} In the experiments in the subsequent sections, we illustrate how the DAC, when trained with this loss function, learns representations for abstention remarkably well.

3 The DAC as a Learner of Structured Noise

While noisy training labels are usually an unavoidable occurrence in real-world data, such noise can often exhibit a pattern attributable to training data being corrupted in some non-arbitrary or systematic manner. This kind of label noise can occur when some classes are more likely to be mislabeled than others, either because of confusing features or a lack of sufficient level of expertise or unreliability of the annotator. For example, the occurrence of non-iid, systematic label noise in brain-computer interface applications – where noisy data is correlated with the state of the participant – has been documented in (Porbadnigk et al., 2014; Görnitz et al., 2014). In image data collected for training (that might have been automatically pre-tagged by a recognition system), a subset of the images might be of degraded quality, causing such labels to be unreliable[1]. Further, systematic noise can also occur if all the data were labeled using the same mechanism (Brodley & Friedl, 1999); for a comprehensive survey of label noise see (Frénay & Verleysen, 2014).

In these scenarios, there are usually consistent indications in the input x that tend to be correlated with noise in the labels, but such correlations are rarely initially obvious. Given the large amount of data required to train deep models, the process of curating the data down to a clean, reliable set might be prohibitively expensive. In situations involving sensitive data (patient records, for example) crowd-sourcing label annotations might not even be an option. However, given that DNNs can learn rich, hierarchical representations, one of the questions we explore in this paper is whether we can exploit the representational power of DNNs to *learn* such feature mappings that are indicative of unreliable or confusing samples. Since abstention is driven by the cross-entropy in the training loss, features that are consistently picked up by the DAC during abstention should thus have high feature weights with respect to the abstention class, suggesting that the DAC might learn to make such associations. In the following sections, we describe a series of experiments on image data that demonstrate precisely this behavior – using abstention training, the DAC learns features that are associ-

[1] We assume, in this case, one only has access to the labels, and not the confidence scores



(a) (b) (c) (d) (e)

Figure 1: (a) A sample of smudged images on which labels were randomized.(b) Abstention percentage on training set as training progresses (c)The DAC abstains on most of the smudged images on the test set (abstention recall) (d) Almost all of the abstained images were those that were smudged (abstention precision) (e) Risk-coverage curves for baseline DNNs, DAC and post-DAC DNN. For the baseline and post-DAC DNNs, we report coverage based on softmax and SGR thresholds. First-pass training with the DAC improves performance for both softmax and SGR methods.

ated with difficult or confusing samples and reliably learns to abstain based on these features.

3.1 Experiments

Setup: For the experiments in this section, we use a deep convolutional network employing the VGG-16 (Simonyan & Zisserman, 2014) architecture, implemented in the PyTorch (Paszke et al., 2017) framework. We train the network for 200 epochs using SGD accelerated with Nesterov momentum and employ a weight decay of .0005, initial learning rate of 0.1 and learning rate annealing using an annealing factor of 0.5 at epoch $60, 120$ and 160. We perform abstention-free training during the first 20 epochs which allows for faster training[2] To enable better visualization, in this section, we use the labeled version of the STL-10 dataset (Coates et al., 2011), comprising of 5000 and 8000 96x96 RGB images in the train and test set respectively, augmented with random crops and horizontal flips during training. We use this architecture and dataset combination to keep training times reasonable, but over a relatively challenging dataset with complex features. For the α auto-update algorithm we set ρ (α initialization factor) to 64 and μ to 0.05; we did not tune these parameters.

3.2 Noisy Labels Co-Occurring with an Underlying Cross-Class Feature

We first demonstrate the ability of the DAC to learn features associated with confusing labels in a toy experiment. In this experiment we simulate the situation where an underlying, generally unknown feature occurring in a subset of the data often co-occurs with inconsistent mapping between features and ground truth. In a real-world setting, when encountering data containing such a feature, it is desired that the DAC will abstain on predicting on such samples and hand over the classification to an upstream (possibly human) expert. To simulate this, we randomize the labels (over the original K classes) on 10% of the images in the training set, but add a distinguishing extraneous feature to these images. In our experiments, this feature is a *smudge* (Figure 1a) that represents the aforementioned feature co-occurring with label noise. We then train both a DAC as well as a regular DNN with the usual K-class cross-entropy loss. Performance is tested on a set where 10% of the images are also smudged. Since it is hoped that the DAC learns representations for the structured noise occurring in the dataset, and assigns the abstention class for such training points, we also report the performance of a DNN that has been trained on a set where the abstained samples were eliminated (*post-DAC*) at the best-performing epoch of the DAC[3] Performance is reported in terms of accuracy-vs-abstained (i.e., risk-vs-coverage) curves for the DAC, and the standard softmax threshold-based abstention for the baseline DNNs and post-DAC DNNs. As an additional baseline, we also compare the performance of the recently proposed selective guaranteed risk (SGR) method in (Geifman & El-Yaniv, 2017) for both the baseline and post-DAC DNNs that maximizes coverage subject to a user-specified risk bound (we use their default confidence parameter, δ, of 0.001 and report coverage for a series of risk values.)

Results When trained over a non-corrupted set, the baseline (i.e., non-abstaining) DNN had a test accuracy over 82%, but this drops to under 75% (at 100% coverage) when trained on the label-randomized smudged set (Figure 1e). For the DAC, Figure 1b shows the abstention progress on the train set. When abstention is initiated after 20 epochs, the DAC chooses to abstain on all but the easiest samples learned so far, but progressively abstains on less data till the abstention reaches steady behavior between epochs 120 and 150, abstaining on about 10% of the data, representing the smudged images. Further annealing of learning rate (at epoch 160) causes the DAC to go into memorization

[2]Training with abstention from the start just means we have to train for a longer number of epochs to reach a given abstention-vs-accuracy point.

[3]we describe in Section 4 how noisy data is eliminated

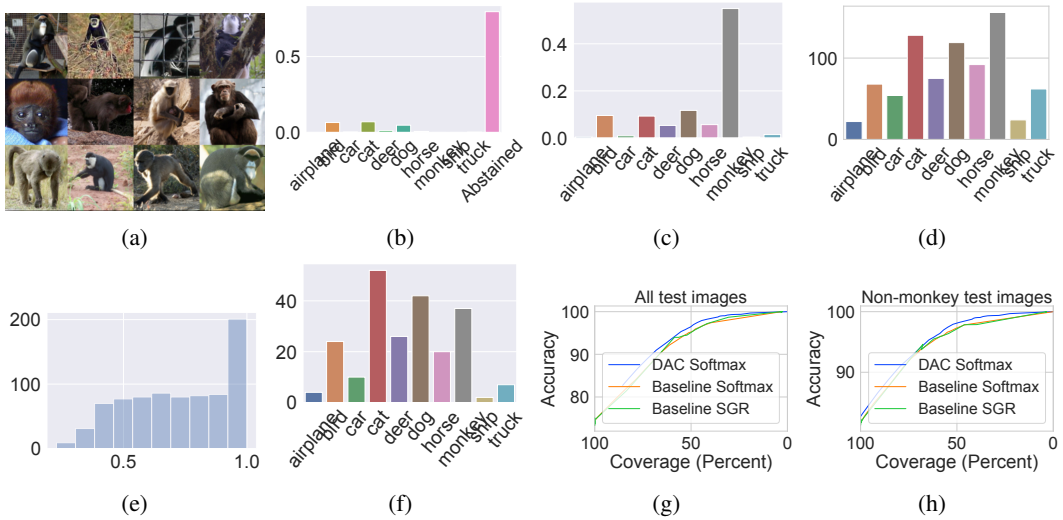


Figure 2: (a) All monkey images in the training set had their labels randomized (b) The DAC abstains on about 80% of the monkey images (abstention recall). (c) Among images that were abstained, most of the images were those of monkeys (abstention precision). (d) Distribution of baseline DNN predictions over monkey images in the test set indicating learning difficulty on this class (e) Distribution of winning softmax scores of the baseline DNN on the monkey images (f) Distribution of baseline DNN softmax scores > 0.9. Most of these confident predictions are non-monkeys (g) Comparison against various abstention methods (softmax and SGR) on all the test images (h) Same comparison, but only on those images on which the DAC did not abstain (i.e mostly non-monkey images). The DAC has a small but consistent advantage in both cases. All figures are computed on the test set.

mode.[4]. However, at the best performing validation epoch, the DAC abstains – with both high precision and recall – on precisely those set of images that have been smudged (Figures 1c and 1d)! In other words, it appears the DAC has learned a clear association between the smudge and unreliable training data, and opts to abstain whenever this feature is encountered in an image sample. Essentially, the smudge has become a separate class all unto itself, with the DAC assigning it the abstention class label. The risk-coverage curve for the DAC (Figure 1e), calculated using softmax thresholds at the best validation epoch, closely tracks the baseline DNN's softmax thresholded curve; this is not surprising, since on those images that are not abstained on, the DAC and the DNN learn in similar fashion due to the way the loss function is constructed. We do however see a strong performance boost by eliminating the data abstained on by the DAC and then re-training a DNN. This post-DAC DNN has significantly higher accuracy than the baseline DNN (Figure 1e), and also has consistently better risk-coverage curves. Not surprisingly this performance boost is also imparted to the SGR method since any improvement in the base accuracy of the classifier will be reflected in better risk-coverage curves. In this sense, the DAC is complementary to an uncertainty quantification method like SGR or standard softmax thresholding – first training with the DAC and then a DNN improves overall performance. While this experiment clearly illustrates the DAC's ability to associate a particular feature with the abstention class, it might be argued the consistency of the smudge made this particular task easier than a typical real world setting. We provide a more challenging version of this experiment in the next section.

[4]We discuss abstention and memorization in Section 5

3.3 Noisy Labels associated with a class

In this experiment, we simulate a scenario where a particular class, for some reason, is very prone to mislabeling, but it is assumed that given enough training data and clean labels, a deep network can learn the correct mapping. To simulate a rather extreme scenario, we randomize the labels over all the monkeys in the training set, which in fact include a variety of animals in the ape category (chimpanzees, macaques, baboons etc; Figure 2a) but all labeled as 'monkey'. Unlike the previous experiment, where the smudge was a relatively simple and consistent feature, the set of features that the DAC now has to learn are over a complex real-world object with more intra-class variance.

Detailed results are shown in Figure 2. The DAC abstains on most of the monkey images in the test set (Figure 2b), while abstaining on relatively many fewer images in the other classes (Figure 2c), suggesting good abstention recall and precision respectively. In essence, the DAC, like a non-

abstaining DNN would in a clean-data scenario, has learned meaningful representation of monkeys, but due to label randomization, the abstention loss function now encourages the DAC to associate monkey features with the abstention class. That is, the DAC, in the presence of label noise on this particular class, has learned a mapping from class features X to class $K_{abstain}$, much like a regular DNN would have learned a mapping from X to K_{monkey} in the absence of label noise. The representational power is unchanged from the DAC to the DNN; the difference is that the optimization induced by the loss function now redirects the mapping towards the abstention class.

Also shown is the performance of the baseline DNN in Figures 2d to 2f. The prediction distribution over the monkey images spans the entire class range. That the DNN does get the classification correct about 20% of the time is not surprising, given that about 10% of the randomized monkey images did end up with the correct label, providing a consistent mapping from features to labels in these cases. However the accuracy on monkey images is poor; the distribution of the winning softmax scores over the monkey images for the DNN is shown in Figure 2e, revealing a high number of confident predictions ($p >= 0.9$) but closer inspection of the class distributions across just these confident predictions (2f) reveals that most of these predictions are incorrect suggesting that a threshold-based approach, which generally works well (Hendrycks & Gimpel, 2016; Geifman & El-Yaniv, 2017), will produce confident but erroneous predictions in such cases. This is reflected in the small but consistent risk-vs-coverage advantage of the DAC in Figure 2g and 2h. As before we compare both a softmax-thresholded DAC and baseline DNN, as well as the SGR method tuned on the baseline DNN scores. Unlike the random smudging experiment, here we do not eliminate the abstained images and retrain – doing so would completely eliminate one class. Instead we additionally compare the performance of the DAC on the images that it did not abstain (mostly non-monkeys), with the baselines (Figure 2h) – the DAC has a small but significant lead in this case as well.

In summary, the experiments in this section indicate that the DAC can reliably pick up and abstain on samples where the noise is correlated with an underlying feature. In the next section, we peek into the network for better insights into the features that cause the DAC to abstain.

3.4 Visual Explanations of Abstention

It is instructive to peer inside the network for explaining abstention behavior. Convolutional filter visualization techniques such as guided back-propagation (Springenberg et al., 2014) combined with class-based activation maps (Selvaraju et al., 2017) provide visually interpretable explanations of DNN predictions. In the case of the DAC, we visualize the final convolutional filters on the trained VGG-16 DAC model that successfully abstained on smudged and monkey images described in experiments in the previous section. Example visualizations using class-based activation maps on the predicted class are depicted in Figure 3. In the smudging experiments, when abstaining, the smudge completely dominates the rest of the features (Figures 3a,b). The same image, when presented without a smudge (Figure 3c), is correctly predicted, with the actual class features being much more salient (3d) implying that the abstention decision is driven by the presence of the smudge. For the randomized monkey experiment, it is precisely the features associated with the monkey class that result in abstention(Figures 3e, 3f), visually confirming our hypothesis in Section 3.3 that the DAC has effectively mapped monkey features to the abstention label. Further experiments illustrating the abstention ability of the DAC in the presence of structured noise are described in Section B in the supplementary material.

(a) (b) (c) (d) (e) (f)

Figure 3: Filter visualizations for the DAC. When presented with a smudged image (a), the smudge completely dominates the feature saliency map (b) that causes the DAC to abstain. However for the same image without a smudge (c), the class features become much more salient (d) resulting in a correct prediction. In the random monkeys experiment, for abstention on monkeys (e), the monkey features are picked up correctly (f), which leads to abstention

4 Learning in the Presence of Unstructured Noise: The DAC as a Data Cleaner

So far we have seen the utility of the DAC in structured noise settings, where the DAC learns representations on which to abstain. Here we consider the problem of unstructured noise – noisy labels that might occur arbitrarily on some fraction of the data. Classification performance degrades in the presence of noise (Nettleton et al., 2010), with label noise shown to be more harmful than feature noise (Zhu & Wu, 2004). While there have been a number of works related to DNN training in the presence of noise (Sukhbaatar et al., 2014; Reed et al., 2014; Patrini et al., 2017), unlike these works we do not model the label flipping probabilities between classes in detail. We simply assume that a fraction of labels have been uniformly corrupted and approach the problem from a data-cleaning perspective: using the abstention formulation and the extra class, can the DAC be used to identify noisy samples in the training set, with the goal of performing subsequent training, using a regular DNN, on the cleaner set? To identify the samples for elimination, we train the DAC,

Dataset	Method	Label Noise Fraction			
		0.2	0.4	0.6	0.8
CIFAR-10 (ResNet-34)	Baseline	88.94	85.35	79.74	67.17
	$\mathcal{L}_q$	89.83	87.13	82.54	64.07
	Trunc $\mathcal{L}_q$	89.7	87.62	82.7	67.92
	Forward T	88.63	85.07	79.12	64.30
	Forward $\hat{T}$	87.99	83.25	74.96	54.64
	DAC	**92.91 (0.24/0.01)**	**90.71 (0.41/0.03)**	**86.30 (0.56/0.07)**	**74.84 (0.75/0.16)**
	Oracle	92.56	90.95	88.92	86.43
CIFAR-10 (Wide Res-Net 28x10)	Baseline	91.53	88.98	82.69	64.09
	MentorNet	92.0	89.0	-	49.0
	DAC	**93.35 (0.25/0.01)**	**90.93 (0.43/0.01)**	**87.58 (0.59/0.04)**	**70.8 (0.77/0.17)**
	Oracle	95.17	94.38	92.74	91.01
CIFAR-100 (ResNet-34)	Baseline	69.15	62.94	55.39	29.5
	$\mathcal{L}_q$	66.81	61.77	53.16	29.16
	Trunc $\mathcal{L}_q$	67.61	62.64	54.04	29.60
	Forward T	63.16	54.65	44.62	24.83
	Forward $\hat{T}$	39.19	31.05	19.12	8.99
	DAC	**73.55 (0.18/0.05)**	**66.92 (0.25/0.01)**	**57.17 (0.77/0.03)**	**32.16 (0.87/0.33)**
	Oracle	77.15	73.85	69.48	58.5
CIFAR-100 (Wide Res-Net 28x10)	Baseline	71.24	65.24	57.56	30.43
	MentorNet	73.0	68.0	-	**35.0**
	DAC	**75.75 (0.2/0.05)**	**68.2 (0.57/0.01)**	**59.44 (0.76/0.06)**	34.06 (0.87/0.33)
	Oracle	78.76	76.23	72.11	63.08
Fashion-MNIST (ResNet-18)	Baseline	93.91	93.09	91.83	88.61
	$\mathcal{L}_q$	93.35	92.58	91.3	88.01
	$\mathcal{L}_q$	93.21	92.6	91.56	88.33
	Forward T	93.64	92.69	91.16	87.59
	Forward $\hat{T}$	93.26	92.24	90.54	85.57
	DAC	**94.76 (0.25/0.01)**	**94.09 (0.48/0.01)**	**92.97 (0.66/0.03)**	**90.79 (0.88/0.04)**
	Oracle	95.22	94.87	94.64	93.63

Table 1: Comparison of performance of DAC against related work for data corrupted with uniform label-noise. The DAC is used to first filter out noisy samples from the training set and a DNN is then trained on the cleaner set. Each set also shows the performance of the baseline DNN trained on the original data. Also shown is the performance of a hypothetical oracle data-cleaner that has perfect information about noisy labels. The parenthetical numbers next to the DAC indicate the fraction of training data removed by the DAC and the remaining noise level. $\mathcal{L}_q$, truncated $\mathcal{L}_q$, and Forward results are from (Zhang & Sabuncu, 2018); MentorNet results are from (Jiang et al., 2018)

observing the performance of the non-abstaining part of the DAC on a validation set (which we assume to be clean). As mentioned before, this non-abstaining portion of the DAC is simply the DAC with the abstention mass normalized out of the true classes. The result in Lemma 1 assures that learning continues on the true classes even in the presence of abstention. However at the point of best validation error, if there continues to be training error on the non-abstaining portion of the DAC, then this is likely indicative of label noise; it is these samples that are eliminated from the training set for subsequent training using regular cross-entropy loss.

To test the performance of the DAC, our comparisons include two recent models that achieve state-of-the-art results in training with noisy labels on image data: MentorNet (Jiang et al., 2018) that uses a data-driven curriculum-learning approach involving two neural nets – a learning network (StudentNet) and a supervisory network (MentorNet); and (Zhang & Sabuncu, 2018), that uses a noise-robust loss function formulation involving a generalization of the traditional categorical cross-entropy loss function. We also compare the performance against the Forward method described in (Patrini et al., 2017) which uses a loss correction approach; for the latter we use the numbers reported in (Zhang & Sabuncu, 2018) using the same setup.

Experimental Setup We use the CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009) and the Fashion-MNIST (Xiao et al., 2017) datasets with an increasing fraction of arbitrarily randomized labels, using the same networks as the ones we compare to. In the DAC approach, both the DAC and the downstream DNN (that trains on the cleaner set) use the same network architectures. The downstream DNN is trained using the same hyperparameters, optimization algorithm and weight decay as the models we compare to. As a best-case model in the data-cleaning scenario, we also report the performance of a hypothetical oracle that has perfect information of the corrupted labels, and eliminates only those samples. To ensure approximately the same number of optimization steps as the comparisons when data has been eliminated, we appropriately lengthen the number of epochs and learning rate schedule for the downstream DNN (and do the same for the oracle.)

Results are shown in Table 1. By identifying and eliminating noisy samples using the DAC and then training using the cleaner set, noticeable – and often significant – performance improvement is achieved over the comparison methods in most cases. Interestingly, in the case of higher label randomization, for the more challenging data set like CIFAR-10 and CIFAR-100, we see the noisy baseline outperforming some of the comparison methods. The DAC is however, consistently better than the baseline. On CIFAR-100, for 80% randomization, the other methods often have very similar performance to the DAC. This is possibly due to substantial the amount of data that has been eliminated by the DAC leaving very few samples per class. The fact that the performance is comparable even in this case, and the high hypothetical performance of the oracle illustrate the effectiveness of a data cleaning approach for deep learning even when a significant fraction of the data has been eliminated. Additional results in the case of non-uniform, class-dependent label noise are reported in Section C of the Appendix.

While data cleaning (or pruning) approaches have been considered before in the context of shallow classifiers (Angelova et al., 2005; Brodley & Friedl, 1999; Zhu et al., 2003), to the best of our knowledge, this is the first work to show how abstention training can be used to identify and eliminate noisy

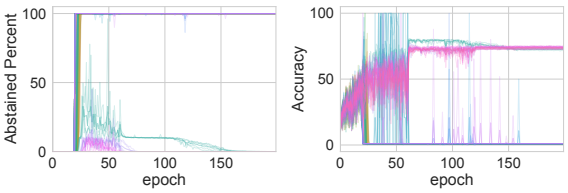


Figure 4: DAC behavior for the random smudging experiments for different fixed α's (indicated by color) showing the tendency for all-or-nothing abstention (left) if α is kept constant. In the zero-abstention case, accuracy is the same as a baseline non-abstaining DNN (right)

labels for improving classification performance. Besides the improvements over the work we compare to, this approach also has additional advantages: we do not need to estimate the label confusion matrix as in (Sukhbaatar et al., 2014; Reed et al., 2014; Patrini et al., 2017) or make assumptions regarding the amount of label noise or the existence of a trusted or clean data set as done in (Hendrycks et al., 2018) and (Li et al., 2017b).

The DAC approach is also significantly simpler than methods based on the mentor-student networks in (Jiang et al., 2018; Han et al., 2018), or the graphical model approach in (Vahdat, 2017). The results in this section not only demonstrate the performance benefit of a data-cleaning approach for robust deep learning in the presence of significant label noise, but also the utility of the abstaining classifier as an effective way to clean such noise.

5 Abstention and Memorization

In the structured-noise experiments in Section 3, we saw that the DAC abstains, often with near perfection, on label-randomized samples by learning common features that are present in these samples. However, there has been a considerable body of recent work that shows that DNNs are also perfectly capable of memorizing random labels (Zhang et al., 2016; Arpit et al., 2017). In this regard, abstention appears to counter the tendency to memorize data; however it does not generally prevent memorization.

Lemma 2. *For the loss function $\mathcal{L}$ given in Equation 1, for a fixed α, and trained over t epochs, as $t \to \infty$, the abstention rate $\gamma \to 0$ or $\gamma \to 1$.*

Proof Sketch. Intuitively, if α is close to 0, p_{k+1} quickly saturates to unity, causing the DAC to abstain on all samples, driving both loss and the gradients to 0 and preventing any further learning. Barring this situation, and given that the gradient $\frac{\partial \mathcal{L}}{\partial a_j} \leq 0$, where j is the true class (see Lemma 1), the condition for abstention in Section 2 eventually fails to be satisfied. After this point, probability mass is removed from the abstention class $k+1$ for all subsequent training, eventually driving abstention to zero. □

Experiments where α was fixed confirm this; Figure 4 shows abstention behavior and the corresponding generalization performance on the validation set for different values of fixed alpha in the random-smudging experiments (Section 3.2). The desired behavior of abstaining on the smudged samples (whose labels were randomized in the training set) does not persist indefinitely. At epochs 60 and 120, there are steep reductions in the abstention rate, coinciding with learning rate decay. At this point, apparently, the DAC moves into a memorization phase, finding more complex decision boundaries to fit random labels, as the lower learning rate enables it to descend into a possibly narrow minima. Generalization performance also suffers once this phase begins. This behavior is consistent with the discussion in (Arpit et al., 2017) – the DAC does indeed first learn patterns before descending into memorization. Auto-tuning of α described in Section 2.1 prevents the abstention rate from saturating to 1; however as we saw in Section 3.2, it does not prevent abstention from converging to 0. A sufficiently small learning rate and long training schedule eventually results in memorization. As discussed earlier, tracking the loss of the non-abstaining portion of the DAC on a validation set can used to determine when a desirable abstention level has been reached.

6 Conclusions

There is little work discussing abstention in the context of deep learning and even less discussing abstention approaches for combating label noise. Here, we demonstrated the effectiveness of such an approach when training deep neural networks. We showed the utility of the DAC under multiple types of label noise: as a representation learner in the presence of structured noise and as an effective data cleaner in the presence of arbitrary noise. Results indicate that data-cleaning with the DAC significantly improves classification performance for downstream training. Furthermore, the loss function formulation is simple to implement and can work with any existing DNN architecture; this makes the DAC a useful addition to real-world deep learning pipelines.

Acknowledgments: The authors were supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Los Alamos National Laboratory under Contract DE-AC5206NA25396. This work was supported in part by the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References

Angelova, A., Abu-Mostafam, Y., and Perona, P. Pruning training sets for learning of object categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pp. 494–501. IEEE, 2005.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. *arXiv preprint arXiv:1706.05394*, 2017.

Brodley, C. E. and Friedl, M. A. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

Chapelle, O., Scholkopf, B., and Zien, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Chow, C. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223, 2011.

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pp. 67–82. Springer, 2016.

De Stefano, C., Sansone, C., and Vento, M. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30 (1):84–94, 2000.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.

Frénay, B. and Verleysen, M. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2014.

Fumera, G. and Roli, F. Support vector machines with embedded reject option. In *Pattern recognition with support vector machines*, pp. 68–82. Springer, 2002.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in neural information processing systems*, pp. 4885–4894, 2017.

Görnitz, N., Porbadnigk, A., Binder, A., Sannelli, C., Braun, M., Müller, K.-R., and Kloft, M. Learning and evaluation in presence of non-iid label noise. In *Artificial Intelligence and Statistics*, pp. 293–302, 2014.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: robust training deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. *arXiv preprint arXiv:1802.05300*, 2018.

ImageNet. Imagenet statistics. URL `http://image-net.org/about-stats`.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2309–2318, 2018.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Li, W., Wang, L., Li, W., Agustsson, E., and Van Gool, L. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017a.

Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., and Li, L.-J. Learning from noisy labels with distillation. In *ICCV*, pp. 1928–1936, 2017b.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. *arXiv preprint*, 2017.

Nettleton, D. F., Orriols-Puig, A., and Fornells, A. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33(4):275–306, 2010.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, pp. 2233–2241, 2017.

Porbadnigk, A. K., Görnitz, N., Sannelli, C., Binder, A., Braun, M., Kloft, M., and Müller, K.-R. When brain and behavior disagree: Tackling systematic label noise in eeg data with machine learning. In *Brain-Computer Interface (BCI), 2014 International Winter Workshop on*, pp. 1–4. IEEE, 2014.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Rolnick, D., Veit, A., Belongie, S., and Shavit, N. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Turk, A. M. https://www.mturk.com.

Vahdat, A. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems*, pp. 5596–5605, 2017.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, Z. and Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836*, 2018.

Zhu, X. and Wu, X. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, 22(3): 177–210, 2004.

Zhu, X., Wu, X., and Chen, Q. Eliminating class noise in large datasets. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 920–927, 2003.