KNOWS WHEN IT DOESN'T KNOW: DEEP ABSTAINING CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce the deep abstaining classifier – a deep neural network trained with a novel loss function that provides an abstention option during training. This allows the DNN to abstain on confusing or difficult-to-learn examples while improving performance on the non-abstained samples. We show that such deep abstaining classifiers can: (i) learn representations for structured noise – where noisy training labels or confusing examples are correlated with underlying features – and then learn to abstain based on such features; (ii) enable robust learning in the presence of arbitrary or unstructured noise by identifying noisy samples; and (iii) be used as an effective out-of-category detector that learns to reliably abstain when presented with samples from unknown classes. We provide analytical results on loss function behavior that enable automatic tuning of accuracy and coverage. We demonstrate the utility of the deep abstaining classifier using multiple image benchmarks.

1 INTRODUCTION

Machine learning algorithms are expected to increasingly replace humans in decision-making pipelines. With the deployment of AI-based systems in high risk fields such as medical diagnosis (Miotto et al. (2016)), autonomous vehicle control (Levinson et al. (2011)) and the legal sector (Berk (2017)), an erroneous prediction that should have otherwise been flagged for human intervention because the system has not robustly learned when it is likely to get the wrong answer can have severe consequences.

In these situations, the quality of "knowing when it doesn't know" and abstaining from predicting is an essential trait for a classifier to possess. This allows the decision-making to be routed to a human or another more accurate, but possibly more expensive, classifier, with the assumption being that the additional cost incurred is greatly surpassed by the consequences of a wrong prediction.

Since learning systems have been around for multiple decades, there has been extensive theoretical and empirical investigation into rejection (or abstention) classification with the bulk of this being in the area of shallow learners (Chow (1970); Cortes et al. (2016); Fumera & Roli (2002)) and multilayer perceptrons (De Stefano et al. (2000)). A framework for "self-aware learning" was analyzed in the context of Markov decision processes in Li et al. (2008). 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 risk-vs-coverage profiles based on the output of a trained model.

In this paper, we focus on exploiting the representational power of DNNs to *learn* features that are likely to lead to uncertain predictions based on associations made by the network while training. We introduce a new modified loss function that utilizes an abstention output allowing the DNN to learn when abstention is a better option, while at the same improving performance on the non-abstained samples. Learning such representations is useful when training data contains label noise that is correlated with some underlying feature (structured noise) as well as in the case of arbitrary label noise. In the former case, such features might be indicative of when the DNN is likely to get the wrong answer, enabling more robust predictions. In the latter case, the samples that have been abstained during training can be filtered out for subsequent training which often leads to increased performance.

We also consider the problem of open-set detection since real-world systems are often deployed in open-domain situations. When presented with samples from unknown classes, abstention is often the safest choice. There have been a number of recent works on open-set and the related problem of out-of-distribution detection for deep networks that use post-training calibration: (Bendale & Boult (2016)) describe an open-set detection method by learning the distribution of the pre-activations of the trained model. Hendrycks & Gimpel (2016) uses the winning softmax responses of the DNN which, while often effective, are also prone to overconfident wrong predictions, even when presented with random noise data. Liang et al.) use temperature scaling and input perturbation for state-of-the-art out-of-distribution detection. Here, we provide an alternate but effective but simple method for open-set detection by training the DNN to learn feature-based thresholds for classes of interest.

- novel loss function for training an abstaining deep neural network – a deep abstaining classifier (DAC) – that enables representation learning for abstention in the presence of structured label noise. When trained on structured noisy data, the DAC is able to pick up the features associated with noisy data with remarkable precision.
- Demonstration of the utility of the DAC as a *data cleaner* in the presence of arbitrary label noise that significantly improves downstream training.
- A simple method to use the DAC as an effective open-set detector that learns to reliably abstain when presented with samples from unknown classes.

We note that, 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.

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 ith 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 which ensures that learning persists even in the presence of abstention. 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 thus 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. Depending on α, if it becomes hard during the training process for the model to achieve low

Algorithm 1: α auto-tuning

input : T, L, t, ρ, α_{final}, P^M
$\beta = (1 - P^M_{k+1})\mathcal{H}_c(P^M_{1\ldots K})$
for $t := 0$ *to* T **do**
  $\tilde{\beta} \leftarrow (1 - \mu)\tilde{\beta} + \mu\beta$
  if $t = L$ **then**
    $\alpha := \tilde{\beta}/\rho$
    $\delta_\alpha := \frac{\alpha_{final} - \alpha}{T - L}$
  end
  if $t > L$ **then**
    $\alpha \leftarrow \alpha + \delta_\alpha$
  end
end

cross-entropy on a sample, then the process can decide to allow
the model to abstain on that sample. In other words, poor cross-entropy performance on a sample drives model abstention, and vice versa, where α indicates a tradeoff.

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:

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

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 learnt to correctly classify that particular sample from class j, and will push mass into abstention class provided α satisifies the above inequality. This constraint allows us to perform auto-tuning on α during training with the algorithm is given in Algorithm 1. P_M is the output vector of mini-batch M and T is the total number of epochs. $\tilde{\beta}$ is a smoothed moving average of the α threshold (initialized to 0), and updated at every epoch. We perform abstention-free training for a few initial epochs (L) to accelerate learning, triggering abstention from epoch $L + 1$ onwards. 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 to a final value of α_{final} to trace out smooth abstention-accuracy curves where accuracy is calculated on the non-abstained samples. In the experiments in the subsequent sections, we illustrate the power of the DAC, when trained with this loss function, to learn representations for abstention remarkably well.

3 THE DAC AS A LEARNER OF STRUCTURED NOISE

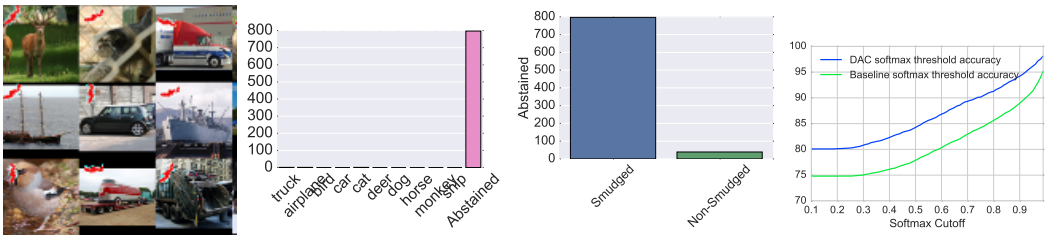
While noisy training labels are usually an unavoidable occurrence in real-world data, such noise often exhibits a pattern attributable to training data being corrupted in some non-arbitrary manner. In image data collected for training – that might have been automatically pre-tagged by a recognition system– a subset of the images might be of degraded quality, causing such labels to be unreliable[1] ; in other scenarios, certain classes might be hard to label correctly unless the labeler has a sufficient level of expertise. Further, certain extraneous features might tend to co-occur with inconsistently labeled data – for example if data labeling has been crowd-sourced, inconsistent labels might correspond with some feature of the annotator itself – certain annotation sources might turn out to be highly unreliable – but this is usually not known in advance. For a comprehensive survey of label noise see Frénay & Verleysen (2014).

In all of the above scenarios, there are 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 training data required, the processing of curating the data down to a clean, reliable set might be too expensive, or in the case of sensitive medical data, outsourcing label annotations might not be an option. However, given that DNNs can learn rich, hierarchical representations LeCun et al. (2015), the question 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. We have seen that abstention is driven by the cross-entropy in the training loss; features that are consistently picked up the DAC during abstention should thus have high feature weights with respect to the abstention class, which indicates 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 associated with difficult or confusing samples and reliably learns to abstain based on these features.

3.1 EXPERIMENTS

Setup: For all the experiments in this section and the rest of the paper, we use a deep convolutional network employing the VGG-16 Simonyan & Zisserman (2014) architecture, implemented in the

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



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

Figure 1: (a) A sample of smudged images on which labels were randomized.(b) DAC Prediction on smudged images on the test set (abstention recall) (c) Almost all of the abstained images were those that were smudged (abstention precision) (d) Classification accuracy vs softmax-threshold based cutoffs indicate better predictive power for the DAC over the actual classes.

PyTorch pyt (2016) framework. We train the network for 200 epochs using SGD accelerated with Nesterov momentum and employ weight decay and learning rate annealing. We performing abstention free training during the first 20 epochs which allows for faster training[2] For the experiments 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 L (abstention-free epochs) to 20, ρ (α initialization factor) to 64 and μ to 0.05.

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

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 afore-mentioned 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. For the DAC we only measure accuracy over the non-abstained samples.

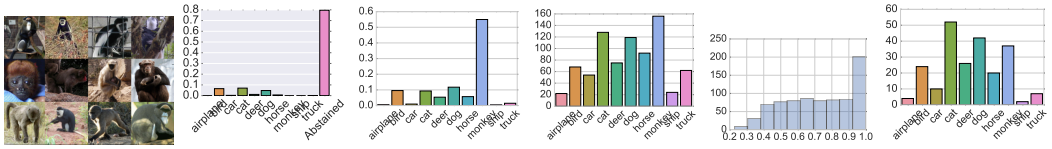
Results (Table 1:) When trained over a non-corrupted set, the baseline (i.e non-abstaining) DNN has an average accuracy over 82%, but this drops to under 75% when trained on the smudged set. On the smudged images alone, the prediction of the baseline was 27.5% (not shown) which, while better than random (since some consistent learning has occurred on similar non-smudged images), is nevertheless a severe degradation in performance.

The DAC however exhibits very different behavior (accuracy and coverage are reported at the end of training) – the reduction in overall accuracy is much lower, while maintaining a high coverage. What is remarkable is that the DAC abstains – with both high precision and recall – on precisely those set of images that have been smudged (Figures 1b and 1c)! In other words, it appears the DAC has learned a clear association between the smudge and unreliable ground truth, 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. Further, in terms of prediction reliability, the DAC has significantly higher accuracy than the baseline DNN (Figure 1d).

	Accuracy	**Coverage**
DNN (Trained on clean)	82.19%	100%
DNN (Trained on corupted)	74.81%	100%
DAC (Trained on clean)	82.66%	100%
DAC (Trained on Corrupted)	80.09%	89.5%
DAC Abstention Precision	93.6%	
DAC Abstention Recall	98.4	

Table 1: DAC vs Baseline DNN on random smudging experiments.

[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.



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

Figure 2: (a) A sample of the monkey class in the trainset that had their labels randomized (b) Prediction on monkey images on the test set for the DAC (abstention recall). (c) For the DAC, 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 (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

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.

3.3 Noisy Labels associated with a class

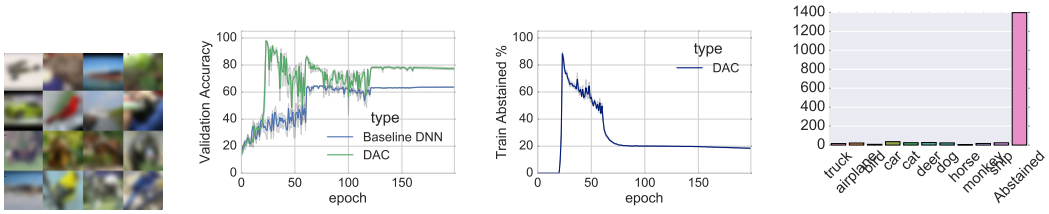
In this experiment, we simulate a scenario where a particular class, for some reason, is very prone to mislabeling, but it is assumed that given enough training data and clean labels, a deep network can learn the correct mapping. To simulate a rather extreme scenario, we randomize the labels over all 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 much 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 a clean-data scenario, has learned meaningful representation of monkeys, but due to label randomization, the abstention loss function now forces 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 abstention class $K+1$, 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.

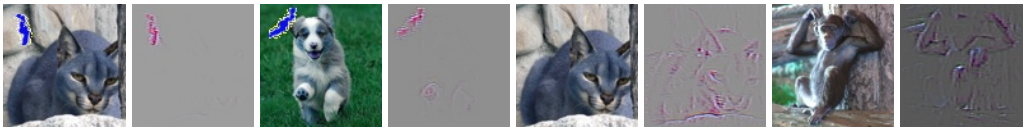
3.4 Noisy Labels associated with data transformation

Here we simulate a scenario where a subset of the training data, due to feature degradation, ends up with unreliable labels. We apply a Gaussian blurring transformation to 20% of the train and test images across all the classes (Figure 3a), and randomize the labels on the blurred training set. This is similar to the smudging experiment, but lacks the presence of a consistent, conspicuous feature that the DAC can associate with abstention. On the other hand, the lack of high frequency components, or



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

Figure 3: Results on blurred-image experiment with noisy labels (a)20% of the images are blurred in the train set, and their labels randomized (b) Validation accuracy for baseline vs DAC (non-abstained) (c)Abstention behavior for the DAC during training (d) Distribution of predictions on the blurred validation images for the DAC. We also observed (not shown) that for the baseline DNN, the accuracy on the blurred images in the validation set is no better than random.



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

Figure 4: Filter visualizations for the DAC. When presented with a smudged image(a,c), the smudge completely dominates the feature saliency map(b,d) that cause the DAC to abstain. However for the same image without a smudge(e), the class features become much more salient (f) resulting in a correct prediction. For abstention on monkeys(g), the monkey features are picked up correctly (h)which leads to abstention

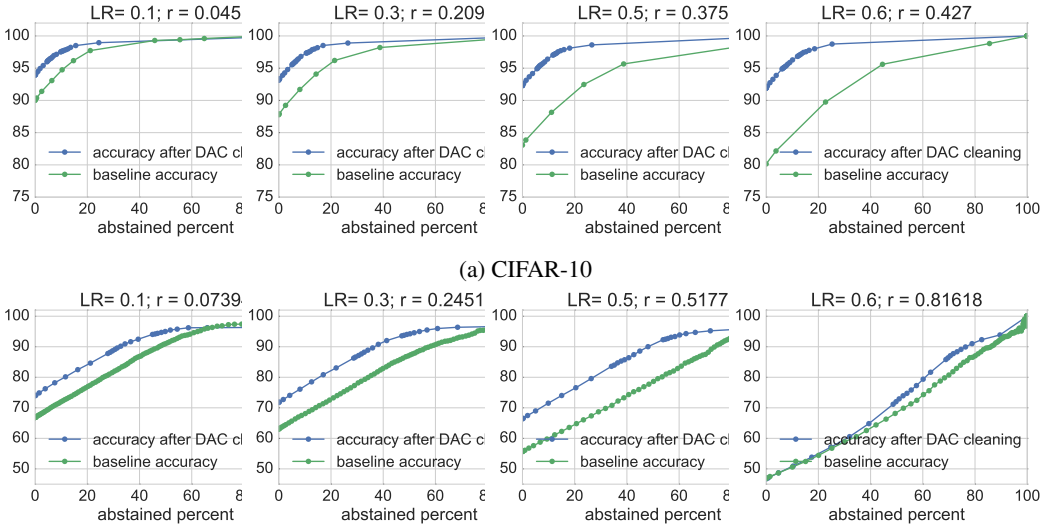
conversely the abundance of low frequency components, might itself be thought of as a feature that is consistent across the samples that have had their label randomized.

Results The DAC abstains remarkably well on the blurred images in the test set (Figure 3d), while maintaining classification accuracy over the remaining samples in the validation set ($\approx 79\%$). The baseline DNN accuracy drops to 63% (Figure 3b), while the basline accuracy over the smudged images alone is no better than random ($\approx 9.8\%$) . The abstention behavior of the DAC on the blurred images in the test set can be explained by how abstention evolves during training (Figure 3c). Once abstention is introduced at epoch 20, the DAC initially opts to abstain on a high percentage of the traning data, while continuing to learn (since the gradients w.r.t the true-class pre-activations are always negative.). In the later epochs, sufficient learning has taken place on the non-randomized samples but the DAC continues to abstain on about 20% of the training data, which corresponds to the blurred images indicating that a strong association has been made between blurring and abstention.

3.5 What does the DAC "See"? 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 4. In the smudging experiments, when abstaining, the smudge completely dominates the rest of the features while the same image, when presented without a smudge, is correctly predicted – with the actual class features being much more salient – implying that the abstention decision is driven by the presence of the smudge. For the randomized monkey experiment, while abstaining, it is precisely the features associated with the monkey class that result in abstention, visually confirming our hypothesis in Section 3.3 that the DAC has effectively mapped monkey features to the abstention label.

4 LEARNING IN THE PRESENCE OF UNSTRUCTURED NOISE: THE DAC AS A DATA CLEANER



(a) CIFAR-10

(b) CIFAR-100

Figure 5: Softmax-thresholded abstention curves of a baseline DNN trained on an increasing fraction of randomized ground-truth for CIFAR-10 (top) and CIFAR-100 (bottom) vs a DNN trained with data that was first cleaned by DAC abstention. LR indicates the fraction of labels that were randomized initially; r is the fraction of labels that were removed by the DAC before the second-pass training.

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-cleansing perspective: can the DAC be used to identify noisy samples with the goal of performing subsequent training, using a regular DNN, on the cleaner set?

Data Here we use the CIFAR-10 and CIFAR-100 datasetsKrizhevsky & Hinton (2009), with an increasing fraction of arbitrarily randomized labels, keeping the network, optimization and learning schedule the same as the experiments in Section 3. We first train the DAC on the noisy set, and then eliminate the training samples that were abstained on at the end of traning. We then perform subsequent training on the cleaner set using a DNN with regular cross-entropy loss. The results are compared to a baseline DNN trained on the entire set to illustrate performance improvement.

Results are shown in 5. By identifying and eliminating noisy samples using the DAC, significant performance improvement is achieved up to a high fraction of label randomization. At 60% randomization for CIFAR-100(Figure 5b), a substantial fraction of the clean data (over 80%) has also been discarded leading to a degradation in performance, but still no worse than the baseline DNN. Recent work in Rolnick et al. (2017) suggests that DNNs are resistant to label noise, but the number of clean samples was kept fixed, as increasing amounts of noisy data were added. While the results here indicate that the baseline DNN does manage to maintain high levels of accuracy while training on a significantly corrupted set, utilizing the DAC to filter out noisy samples confers a

noticeable performance boost to the classifier trained with the cleaner sample.[3]. These results here are in agreement with other observations in the literature, for example Barandela & Gasca (2000)

5 ABSTENTION AND MEMORIZATION

In the experiments in the previous sections, 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 capably 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 1. *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 $\frac{\partial \mathcal{L}}{\partial a_j} \geq 0$, where j is the true class, 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. □

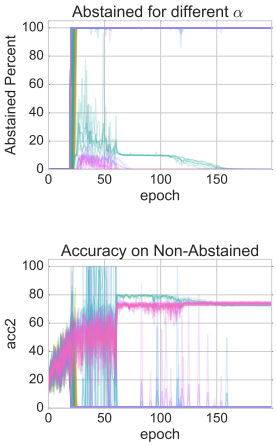


Figure 6: DAC behavior for the rwandom smudging experiments for different fixed α's

Experiments where α was fixed confirm this; Figure 6 shows abstention behavior, and the corresponding generalization performance, on the validation set for different values of fixed alpha in the random-smudging experiments described in 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. It appears that at this point, the DAC moves into a memorization phase, finding more complex decision boundaries to fit the 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 literally descending into memorization. Unlike the controlled experiments in this paper, where we are able to determine when the ideal amount of abstention is occurring, in general, one would need to track the accuracy subject to some user-specified coverage on the validation set. Auto-tuning of α helps avoid the all-or-nothing behavior, and allows the DAC to delay memorization on noisy samples (see for example Figure 3c).

6 OPEN-WORLD DETECTION WITH THE DAC

In this section, we present a simple, effective method for using the DAC in an open-world detection scenario, where the desired behavior when encountering a sample from an unseen class is to abstain. Previous work in this area, and in the related problem of out-of-distribution detection use various post-training calibration techniques – modeling distributions of the pre-softmax layers (Bendale & Boult (2016)), softmax thresholding (Hendrycks & Gimpel (2016)) and temperature scaling with input perturbation (Liang et al.). Here we present an alternate mechanism based on a feature-learning approach that is used to train the DAC to decide when to abstain.

The experimental results from the previous section indicate that depending on image content, there is context-dependent suppression of features based on what associations have been made during training. For example, from the filter visualizations in the smudging experiments (Figure 4), when the DAC is presented with the image of a cat without a smudge, the cat features are expressed clearly

[3]Another option is to just discard the noisy labels and perform semi-supervised training for even better performance

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 7

in the filters, while the same features get suppressed in the presence of the smudge since the smudge was strongly indicative of abstention. We exploit this phenomena to use the DAC as an effective out-of-category detector.

The DAC is trained with both in-category and out-of-category (OoC) data, but a fixed feature (a smudge, for instance) is added to all the OoC data. However, to ensure that the fixed feature is suppressed in the presence of in-class data, we also add this feature to a fraction of the in-sample data, i.e $\tilde{N}_{in} = \tilde{N}_{out}$, $\tilde{N}_{in}$ and $\tilde{N}_{out}$ denoting the number of in and out-of-category samples that have the fixed feature. If $\tilde{X}$ denotes the fixed feature, then $P(Y_{in}|\tilde{X}) = P(Y_{out}|\tilde{X})$, but the probability of any particular class $P(Y_k|\tilde{X}) < P(Y_{out}|\tilde{X}$. That is, given the fixed feature alone, the DAC abstains unless there are other class-specific features that cause non-abstention. During training, the DAC learns the optimal weights on these activations, per-class and per-feature. Conceptually, this is a threshold-based detector, but optimized at the feature level during training. During inference, all incoming samples are augmented with $\tilde{X}$ in a pre-processing step and whether the DAC abstains or not is determined by the presence of known class features. Filter visualizations illustrating this idea are shown in Figure 7. We re-use the smudge here as our fixed feature; when presented with in-class data that is also smudged (horse or car), the DAC activations for the class features are much more salient than the smudge. On the other hand, for out-of-class data, the smudge is more salient driving the prediction to abstention. We also present detection results for two real-world image datasets (STL-10 and Tiny ImageNet) treating each of these as in-sample and out-of-sample in separare experiments. We choose 5k images from the Tiny Imagenet set, picking 10 categories that do not overlap with those in STL-10. We also further use Gaussian and uniform random noise as additional out-of-category data. Results showing detection performance are shown in Table 2 illustrating the effectiveness of the DAC in open-domain situations.

In/Out	**FPR (95% TPR)**	**AUROC**
STl-10/TinyImageNet	29.7	89.1
STL-10/Gaussian	0.0	100.0
STL-10/Uniform	0.0	100.0
TinyImageNet/STL-10	25.6	90.3
TinyImageNet/Gaussian	0.0	100.0
TinyImageNet/Uniform	0.0	100.0

Table 2: Out-of-category detection results for the DAC on STL-10 and Tiny ImageNet.

7 Conclusion

We introduced and the illustrated the utility of a deep abstaining classifier – a DNN trained on a novel loss function that *learns* to abstain as opposed to abstention calibration after training. We illustrated the utility of the DAC in multiple situations: as a representation learner in the presence of structured noise, as an effective data cleaner in the presence of arbitrary noise, and as an effective out-of-category detector. While adversarial settings were not considered in this work, the DAC, and abstention in general, might be considered as part of the defense arsenal against adversarial attacks; we leave this for future work. In summary, our results here indicate that the representational power of DNNs can be used very effectively as a means of self-calibration – "knowing when it doesn't know".

References

October 2016. URL `https://pytorch.org`.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. *arXiv preprint arXiv:1706.05394*, 2017.

Ricardo Barandela and Eduardo Gasca. Decontamination of training samples for supervised pattern recognition methods. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 621–630. Springer, 2000.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

Richard Berk. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13(2):193–216, 2017.

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

Adam Coates, Andrew Ng, and Honglak Lee. 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.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pp. 67–82. Springer, 2016.

Claudio De Stefano, Carlo Sansone, and Mario Vento. 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.

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

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

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

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

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

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pp. 163–168. IEEE, 2011.

Lihong Li, Michael L Littman, and Thomas J Walsh. Knows what it knows: a framework for self-aware learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 568–575. ACM, 2008.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *arXiv preprint*, 2017.

David F Nettleton, Albert Orriols-Puig, and Albert Fornells. 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.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 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.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 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.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

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

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

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