PAL : Pretext-based Active Learning

Shubhang Bhatnagar,[1] **Sachin Goyal**,[2,*] **Darshan Tank**,[1,*] **Amit Sethi** [1]
[1]Indian Institute of Technology, Bombay
[2]Microsoft Research, India
[*]Equal contribution

Abstract

The goal of active learning algorithms is to judiciously select subsets of unlabeled samples to be labeled by an oracle, in order to reduce the time and cost associated with supervised learning. Previously, active learning techniques for deep neural networks have used the same network for the task at hand (e.g., classification) as well as sample selection, which can be conflicting goals. To address this issue, we use a separate sample scoring network to capture the relevant information about the distribution of the labeled samples, and use it to assess the novelty of unlabeled samples. Specifically, we propose to efficiently train the scoring network using a self-supervised learning (pretext) task on the labeled samples. To make the scoring network more robust, we added to it another head, which is trained using the supervised (task) objective itself. The scoring network was paired with a scoring function that allows an appropriate trade-off between the two heads. We also ensure that the selected samples are diverse by selectively fine-tuning the scoring network in sub-rounds of each query round. The resulting scheme performs competitively with the state-of-the-art on benchmark datasets. More importantly, in realistic scenarios when some labels are erroneous and new classes are introduced on the fly, the performance of the proposed method remains strong.

1. Introduction

In spite of their unprecedented accuracy, a hurdle in the deployment of deep learning for many real world problems is the requirement of large labeled data. While it is easy to access large repositories of unlabeled data, extensive data labeling and annotation is often impractical, time-consuming, and expensive, such as for medical images. Strategies to reduce the labeling requirement include transfer, semi-supervised, few-shot, and active learning. Active learning algorithms are used to decide whether or not to send an unlabeled sample for labeling to an oracle (e.g., a medical expert), such that the increase in the task performance (e.g., classification accuracy) is maximized with respect to a labeling cost (e.g., cumulative number of labels acquired).

We propose a pool-based sampling technique for active learning in which the learning progresses iteratively in rounds. In each round, up to a budgeted number of N additional samples can be selected from the unlabeled pool for labeling [7, 29, 27]. The selection strategy is usually based on picking *novel* and *diverse* samples from the unlabeled pool. Novelty (a.k.a. *uncertainty* and *confusion*) refers to selecting samples that are least similar to the previously labeled samples in order to maximize the information gain by getting them labeled. Diversity refers to selecting samples that explore various high probability regions of an appropriate latent space, instead of being similar to each other.

It is well known that a task network is a poor estimator of its own uncertainty on unlabeled samples that are unlike the labeled samples [16]. Yet, previous active learning methods have relied on it for estimating uncertainty [29, 10, 7, 2]. Because task performance and uncertainty estimation can be conflicting goals, we propose using an auxiliary *scoring network*, in line with previous studies [31]. Secondly, we use the ease of solving a self-supervised learning task on unlabeled samples by the scoring network trained on labeled samples as a proxy for the relative likelihood of unlabeled samples to belong to the distribution of the labeled samples. Low-likelihood samples are good candidates for labeling as they are likely to bring new information. SSL has been used to find out-of-distribution (OOD) samples for anomaly detection [18, 14]. Surprisingly, SSL has not been used for scoring novelty in active learning. The self-supervision labels can be generated inexpensively for testing the uncertainty of the unlabeled samples. Simply training a second network adds minimal additional computational cost compared to other methods that train variational or adversarial models [29, 7], which we confirmed empirically. Because of the use of this pretext (self-supervised) task, we call this technique *pretext-based active learning (PAL)*.

Because SSL works best with a large dataset, our third innovation is to improve the performance of the scoring

network using multi-task learning. Multi-task learning is known to help regularize its constituent networks in the case of limited labeled data [3], which is the case for the previously labeled samples in active learning. We used the classification task itself as the second objective of the scoring network. Additionally, our formulation of the scoring function allows for a usable trade-off between the two heads for mutual correction.

Finally, we ensure diversity among the N samples selected during a query by breaking it into K sub-queries, and checking for novelty of the next sub-query. We tune only the self-supervision head between sub-queries to ensure diversity without increasing the labeling cost. PAL showed accuracy that was competitive with the state-of-the-art [29, 27, 10] in benchmarks, without the use of computationally expensive training (Section 4.1). More importantly, we also tested PAL for the following two realistic scenarios described below.

First, the data labels obtained in the real world are hardly ideal due to human or machine errors in the labeling (e.g., using NLP to label images). A good sampling technique should be able to make out truly confusing samples in presence of label noise, and yet active learning techniques have not been tested on noisy labels, except [29]. We hypothesized and confirmed that PAL is more resilient to label noise as compared to other techniques because it has a label-independent pretext task head (Section 4.2).

Second, data from several classes may be underrepresented or, worse, absent in a *biased initial pool*. Alternatively, as the labeling process proceeds, previously unseen classes might be discovered in the unlabeled data, and one might want to include them in the process from then on. Yet, most works on active learning assume that all classes are represented in the initial labeled pool. A good active learning strategy should rapidly procure samples from such classes to be labeled and match its performance on these classes to the previously well-represented. We found that PAL quickly recovered accuracy on previously unseen classes by an appropriate momentary increase in their sampling rates in the first few queries after their discovery. This is further discussed in Section 4.3.

We also show that PAL is insensitive to the choice of the scoring network architecture (Section 4.4), but the three components – self-supervision, supervision, and diversity – are important (Section 4.5).

2. Related Work

2.1. Active learning

There are several settings for active learning, such as *membership query synthesis* and *stream based sampling*. In the former, the learner generates new samples to query the oracle [1, 34, 19], while in the latter the unlabeled dataset is presented as a stream, and is evaluated online [5, 6]. However, unlike these settings, *pool-based sampling* makes a more complete use of unlabeled and labeled data that is often available [29, 27, 10]. In this setting, starting with a set of labeled samples, a fixed number of samples from the unlabeled pool is selected for querying the oracle for labels so that a performance metric (e.g., classification accuracy) can be maximized. The proposed method fits this setting.

Pool-based active learning techniques aim to pick samples that are novel and/or representative. **Novelty** (a.k.a., uncertainty, confusion, perplexity, non-triviality, out-of-distribution, and informativeness) refers to an unlabeled sample's ability to provide new information, if labeled, *independently of other samples selected*. Among early proposals for measures of novelty, entropy of the estimated class probability mass function (PMF) [28] is too simplistic prone to calibration error [16], discordance between a committee of classifiers [13] can be computationally expensive, and distance from a linear decision boundary [30] is not directly applicable to deep neural networks because of their complex decision boundaries. Distance from an adversarial example has been proposed as an approximation of distance from decision boundary [7], but also it requires computationally expensive gradient descent on image pixels. Uncertainty estimations based on Bayesian frameworks, such as MC-dropout [9, 11], are also computationally expensive. Surprisingly, no one has used the difficulty of solving a self-supervised pretext task as a measure of novelty, which requires only up to one additional network to be trained in parallel with the task network.

Methods based on **diversity** (a.k.a. representation and coverage) seek to select samples that can represent data distribution well. Even if the selected samples are individually novel, their similarity to each other can limit the joint information gained from labeling them as a query group. A method based on identifying a *core-set* has been proposed that models the empirical loss over the set of already labeled samples combined with the pool of query samples on the empirical loss over the whole dataset [27]. However, this approach suffers when the representations are high-dimensional, because the Euclidean distance is a poor local distance estimator in high dimensional spaces. An alternative approach called variational adversarial active learning (*VAAL*) aims to learn a good representation using a variational autoencoder (VAE) trained adversarially using a discriminator that tries to predict if a sample is already labeled [29].

2.2. Self-supervised learning

Self-supervised learning (SSL) has shown great promise in learning usable data representations without needing explicit data labels. The learned representation can later be fine-tuned with a smaller labeled dataset. Most of the pro-

posed SSL techniques automatically create a supervised pretext task by degrading an unlabeled image, and training a neural network to recover the original image. Some commonly used randomized degradations on an image for SSL are removing color [21], reducing resolution [22], occluding parts of an image [25], jumbling the spatial order of its sub-images [24], and applying geometric transforms [12].

SSL can also be used to assess the degree to which a given image is plausible in a probabilistic sense with respect to the distribution of the training images, in order to find out-of-distribution (OOD) samples [18, 14]. If a synthetically degraded version of an image can be restored close to its original, then it follows that images similar to it must have been encountered and modeled during SSL. We use this idea to identify unlabeled images that are most unlike the previously labeled images, and therefore likely to maximize gain in information when labeled. The additional problem that we had to solve was that our training dataset was small due to the active learning scenario.

2.3. Multi-task learning

It has been theoretically proven that, if tasks are related and training dataset is limited, multi-task learning can help improve the accuracy of both tasks [3]. Consequently, multi-task learning has been a subject of vigorous research in deep learning [33, 4, 26]. In our experience, multi-task learning often requires smaller computational costs (fewer networks) in comparison to ensembles for the same task for comparable gains in accuracy using comparable constituent multi-net architectures. We used this idea in the scoring network, for the above mentioned reason, as well as to include the label information, while modeling the distribution of the labeled samples.

3. Method: Pretext based Active Learning

Our method is an instance of pool-based active learning, essence of which can be described as follows. Let the pool of currently labeled samples be $\mathcal{D}_L$ and the pool of unlabeled samples be $\mathcal{D}_U$. A task network $f_\theta(\mathbf{x}_l)$ parameterized by θ is trained on all samples $\mathbf{x}_l \in \mathcal{D}_L$. The active learning algorithm selects a budgeted set of N or fewer samples from $\mathcal{D}_U$ in each query. The queried samples are then labeled by an oracle (assumed ideal, unless stated otherwise), added to $\mathcal{D}_L$, and removed from $\mathcal{D}_U$. The task network is retrained on the expanded $\mathcal{D}_L$ and its increase in accuracy is examined. This process is repeated until a specified number of samples $|\mathcal{D}_L|$ are labeled or a desired accuracy level is achieved.

Previous methods estimated the uncertainty of the unlabeled samples using the task network itself, e.g., based on the entropy of computed class PMF. As shown in Figure 1, we use a different neural network than task network for our selection strategy, which we refer to as the scoring network hereafter. The scoring network has two heads, one each for self-supervision and classification, whose outputs are used to assign a *confusion* score S to an unlabeled image $\mathbf{x}$.

3.1. Self-supervision head

The self-supervision head estimates the likelihood of the unlabeled data to be sampled from the distribution of the labeled samples. In particular, we rotate the images by $90i^\circ$ for $i \in \{0, 1, 2, 3\}$ and train a network g_ϕ parameterized by ϕ to predict i on only the images from $\mathcal{D}_L$ so that the head learns the distribution of the labeled data. Using this head, the following confusion score S_R is assigned to each unlabeled image $\mathbf{x}$:

$$S_R(\mathbf{x}) = -\sum_{i \epsilon \{0,1,2,3\}} g_\phi(\text{rot}_{90i}(\mathbf{x}))_i, \tag{1}$$

where $\text{rot}_{90i}(.)$ is the rotation function and $g_\phi(.)_i$ is the i^{th} component of the estimated PMF of rotation angles. We hypothesize that an image $\mathbf{x} \in \mathcal{D}_U$ for which S_R is closer to its minimum value -4 will likely be similar to the labeled points in $\mathcal{D}_L$, and will fetch little extra information, if labeled. Conversely, for OOD points S_R will be closer to 0.

3.2. Classification head and hybrid score

A scoring network trained just on a self-supervised task may not work very well for the following reasons. Firstly, the score S_R is not reliable for images that have a rotational symmetry, and might be high even if the scoring network has modeled the semantic features well. Moreover, the labels of $\mathcal{D}_L$ are left unused by the scoring network. To correct for the mistakes of the rotation head and to use the labels in $\mathcal{D}_L$, we introduce a classification head $h_\psi(\mathbf{x})$ parameterized by ψ in the scoring network. We compute the degree to which the outputs of h_ψ for an unlabeled sample $\mathbf{x}$ are close to a uniform PMF U, using KL divergence [17], as a second measure of confusion $S_C(\mathbf{x})$, to give a *hybrid confusion score* $S(\mathbf{x})$:

$$\begin{aligned} S(\mathbf{x}) &= S_R(\mathbf{x}) + \lambda S_C(\mathbf{x}) \text{ where} \\ S_C(\mathbf{x}) &= -\text{KL}(U \mid\mid h_\psi(\mathbf{x})), \end{aligned} \tag{2}$$

where $\lambda \geq 0$ is a relative importance hyperparameter. For an OOD sample, we expect the KL divergence to be low and S_C to be high.

The negative of KL divergence of the class PMF from uniform distribution is a more suitable alternative to entropy – a more popular measure of confusion – because its magnitude as S_C on classifiable images appropriately overshadows S_R (e.g., in case of rotationally symmetric images) as shown by the following proposition:

Proposition 1: *Negative of KL-divergence of a class PMF from a uniform distribution can overshadow the confusion score from S_R, but entropy cannot.*

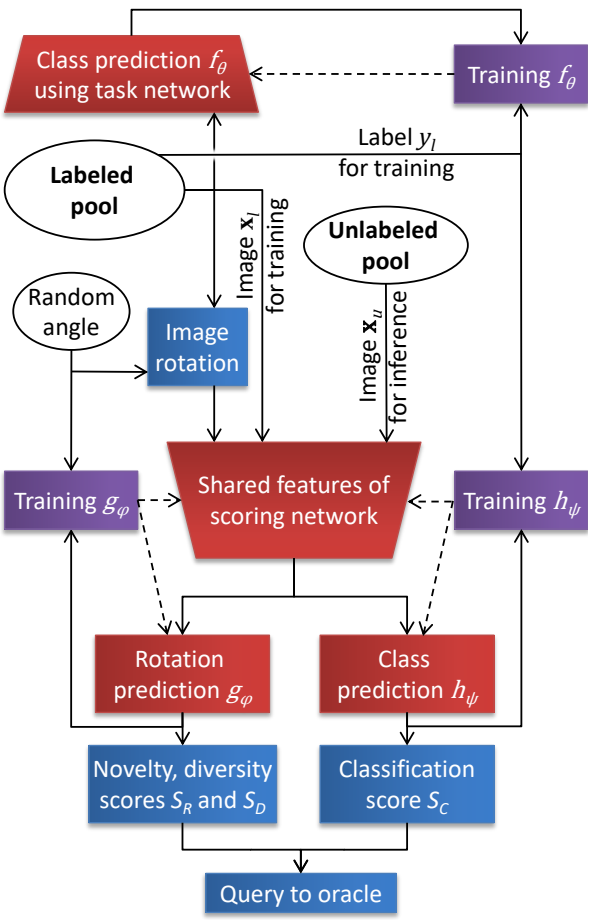

Figure 1. The proposed pretext-based active learning (PAL) has a scoring network in addition to a task network. The two heads of the scoring network are trained on the labeled data, but they assign novelty scores to the unlabeled data points. Dashed lines are training signals. Top N scoring samples are labeled by the oracle.

Proof: Consider a binary classification problem for analysis, with p as the predicted probability score by the task network for the correct class. When the unlabeled sample is almost correctly classified with $p \to 1$, we get the following for the hybrid confusion score of Equation 2:

$$\begin{aligned}\lim_{p\to 1} S &= \lim_{p\to 1}\left(S_R - \frac{\lambda}{2}\log\left(\frac{1}{2p}\right) - \frac{\lambda}{2}\log\left(\frac{1}{2(1-p)}\right)\right)\\ &= S_R - \frac{\lambda}{2}\log\left(\frac{1}{2}\right) - \frac{\lambda}{2}\lim_{p\to 1}\log\left(\frac{1}{2(1-p)}\right)\\ &= -\infty.\end{aligned}$$

On the other hand, if S_C is replaced by the entropy of the PMF h_ψ, then the hybrid score S_E would be finite because:

$$\begin{aligned}\lim_{p\to 1} S_E &= \lim_{p\to 1}\left(S_R - \lambda p\log(p) - \lambda(1-p)\log(1-p)\right)\\ &= S_R - 0 - \lambda\lim_{p\to 1}(1-p)\log(1-p)\\ &= S_R - \lambda\lim_{p\to 1}\frac{\log(1-p)}{\frac{1}{(1-p)}} = S_R,\end{aligned}$$

using L'Hôpital's rule to equate the second term to 0. □

An added advantage of using a multi-task setting for the scoring network is getting better ordinal estimates of a true latent score due to an ensemble-like effect, as long as the correlations between the two components of the score and their correlation with the underlying score are positive. This can follows from the following proposition:

Proposition 2: *There exists a trade-off parameter in Equation 2 that maximizes the correlation between the true underlying score and the hybrid score, which is greater than or equal to the correlation of the true score with either of the components, as long as all correlations between the scores are positive.*

Proof: Note that the requirement of a positive correlation is only a weak one for any reasonably trained networks g_ϕ and h_ψ, as we empirically show in Table 1. Now, without loss of generality, let us assume that some monotonic transformations of the true underlying score, the self-supervision score, and the classification score give standardized random variables u, v, and w respectively, such that their means $\mu_u = \mu_v = \mu_w = 0$, and their variances $\sigma_u^2 = \sigma_v^2 = \sigma_w^2 = 1$. Further, we assume that the covariances σ_{uv}, σ_{uw}, and σ_{vw} are positive. Let an analog of the hybrid score s be a positive combination of the two given by $s = \alpha v + \sqrt{1-\alpha^2}w$, where $\alpha \in [0,1]$ has a monotonic relation with the $\lambda \geq 0$ in Equation 2, and the variance $\sigma_s^2 = 1$. Then, the correlation between u and s, which is the same as the cosine between them, is $\mathbf{E}[u.s] = \alpha\sigma_{uv} + \sqrt{1-\alpha^2}\sigma_{uw}$. If we maximize this correlation by setting its derivative with respect to α to zero, we get:

$$\begin{aligned}&\frac{d\mathbf{E}[u.s]}{d\alpha} = 0\\ \implies &\frac{d}{d\alpha}\left(\sigma_{uv}\alpha + \sigma_{uw}\sqrt{1-\alpha^2}\right) = 0\\ \implies &\sigma_{uv} + \frac{-\alpha}{\sqrt{1-\alpha^2}}\sigma_{uw} = 0\\ \implies &\sigma_{uv}^2(1-\alpha^2) = \sigma_{uw}^2\alpha^2\\ \implies &\alpha = \pm\frac{\sigma_{uv}}{\sqrt{\sigma_{uv}^2 + \sigma_{uw}^2}}\end{aligned}$$

Clearly, a maxima for $\mathbf{E}[u.s]$ exists, because its second derivative is negative for $\alpha^* = +\frac{\sigma_{uv}}{\sqrt{\sigma_{uv}^2+\sigma_{uw}^2}}$ when the covariances are positive, and $\alpha^* \in (0,1)$. □

	SVHN	CIFAR-10	CIFAR-100	Caltech-101
r_p	0.42	0.49	0.46	0.63
r_s	0.44	0.50	0.34	0.57

Table 1. Pearson's correlation (r_p) and Spearman's correlation (r_s) between S_R and S_C of unlabeled data points for the four datasets using a model trained on 10% of the samples.

In summary, we can select the N most informative samples from $\mathcal{D}_U$ with the highest $S(\mathbf{x})$ as per Equation 2 in each query round, after finding a good setting for the hyperparameter $\lambda \geq 0$ based on validation.

3.3. Diversity score

To ensure that the most informative N samples are not similar to each other, we divide a query of size N into K sub queries of size $\frac{N}{K}$ samples each. For selecting the first sub query, we select the top $\frac{N}{K}$ samples using the confusion score S from Equation 2. After picking these, we fine-tune the scoring network using self-supervision without asking the oracle for their labels in the middle of the query. Using this fine-tuned network, we generate a score S_D, quite similar to Equation 1, where $g_{\phi'}$ replaces g_{ϕ}. S_D promotes diversity as it would be small for data points which are similar to the points already selected in the previous sub-queries.

Now, we define an updated score S:

$$S(\mathbf{x}) = S_R(\mathbf{x}) + \lambda_1 S_C(\mathbf{x}) + \lambda_2 S_D(\mathbf{x}) \tag{3}$$

where $S_R(\mathbf{x})$ and $S_C(\mathbf{x})$ are the previously defined confusion score components. Using Equation 3 we select another sub-query of $\frac{N}{K}$ samples, and the process repeats $K-2$ times.

The process of selecting the query samples $\mathcal{D}_Q$ is described in Algorithm 1 dubbed *pretext-based active learning (PAL)*. While $g_{\phi'}$ is fine-tuned during a sub-query, all networks are trained from scratch using the cross entropy loss $\mathcal{L}$ after the oracle labels $\mathcal{D}_Q$ outside Algorithm 1.

4. Experiments and Results

In this section, we empirically show the effectiveness of the proposed *pretext-based active learning (PAL)*. We discuss the experimental setup, datasets used, techniques compared and the implementation details.

Datasets: We performed experiments on four datasets: (1) SVHN [23], where classification task has to be performed for ten digit classes (house numbers) with color images of size 32×32 pixels from google street view images, (2) CIFAR-10 [20], where classification task has to be performed on ten classes in this widely-used computer vision benchmark that contains color images of size 32×32 pixels, (3) CIFAR-100 [20], which is similar to the CIFAR-10 dataset in image size, but is much more difficult with 100 classes and only 600 images per class, and (4) Caltech-101 [8], where classification has to be performed on color images of size 300×200 pixels belonging to 101 different classes, with between only 40 to 800 images per class.

Techniques compared: We compared the performance of our approach with the following active learning sampling strategies. (1) *Random sampling:* This is the simplest but nevertheless a strong baseline involving randomly picking samples to be labeled. (2) *VAAL:* This technique involves using a VAE to learn a feature space and then adversarially training a discriminator on it. It is a current state-of-the-art technique for active learning [29]. (3) *DBAL:* This method uses Bayesian CNNs to estimate uncertainty of unlabeled points and to pick the most uncertain samples [10]. (4) *Core-set:* This is a strong representation-based method for selecting the samples most different than the labeled samples to maximize both uncertainty and diversity of the samples to be picked for labeling [27].

Experimental setup: A fair comparison between various active learning techniques was ensured in line with prior works [29, 27]. All techniques were used to iteratively expand the labeled dataset for training a common classifier architecture – VGG16 [32] – from scratch during each query round. An average of results starting with five random initializations are reported. The initial labeled pool of samples was common to all techniques, and comprised 10% of the whole dataset. Each query added another 5% of the samples selected by the individual active learning algorithms to their respective labeled sets that diverged after the initial 10%. The oracle can be assumed to provide error-free labels, unless stated otherwise.

For the scoring network of the proposed PAL approach, we used a ResNet-18 [15] architecture. The relative importance hyperparameters in Equation 3) were selected from $\{0.5\ ,\ 1.0\}$. Learning rates were in the range $[10^{-1}, 10^{-4}]$. Optimizers were either ADAM or SGD, depending on validation results. The hardware included an NVIDIA GeForce GTX 1080 GPU running CUDA 10.2 and cuDNN 7.6 using PyTorch.

4.1. Performance versus fraction of data labels

Figure 2 shows the mean accuracy for five runs for different fractions of the data labeled, for different active learning techniques. Our PAL strategy outperformed random sampling by a wide margin and consistently seems to outperform VAAL [29], DBAL [10], and core-set [27]. For instance, PAL requires only 20% of labeled SVHN images to achieve performance equal to that achieved by VAAL or DBAL using 30% labels, thus saving a 33% of labeling effort and cost. Additionally, PAL requires only about 2 hours per query round to train on a single GTX 1080 GPU with 11GB memory for SVHN, whereas more computationally expensive methods such as VAAL [29] take more than 24 hours for the same. Out of the techniques compared only

Algorithm 1: Pretext-based Active Learning (PAL)

Result: Set of additional samples to be labeled $\mathcal{D}_Q$
Data: Labeled pool $\mathcal{D}_L := \{\mathbf{X}_L, Y_L\}$, unlabeled pool $\mathcal{D}_U := \{\mathbf{X}_U\}$, query size N
Set: Num. epochs E_Q and E_S, num. sub-queries K
 # Training task and scoring networks
for $t \in \{1, \ldots, E_Q\}$ **do**
  for $\{\boldsymbol{x}_l, y_l\} \in \mathcal{D}_L$ **do**
    $\theta \leftarrow \theta - \eta\nabla_\theta \mathcal{L}(f_\theta(\mathbf{x}_l), y_l)$ # Task network
    $\psi \leftarrow \psi - \eta\nabla_\psi \mathcal{L}(h_\psi(\mathbf{x}_l), y_l)$ # S_C
    for $i \in \{0, 1, 2, 3\}$ **do**
      $\phi \leftarrow \phi - \eta\nabla_\phi \mathcal{L}(g_\phi\,(\text{rot}_{90i}(\mathbf{x}_l)), i)$ # S_R
for $\boldsymbol{x}_u \in \mathcal{D}_U$ **do**
  Use g, h to compute and save $S_R(\mathbf{x}_u)$, $S_C(\mathbf{x}_u)$
 # Diversity-based sub-query sampling
Initialize: $\mathcal{D}_Q = \emptyset; \phi' = \phi$
for $k \in \{1, \ldots, K\}$ **do**
  for $n \in \{1, \ldots, \frac{N}{K}\}$ **do**
    if $k == 1$ **then**
      $\mathbf{x}_q \leftarrow \arg\min_{\mathbf{x}_u \in \mathcal{D}_U} S_R(\mathbf{x}_u) + \lambda_1 S_C(\mathbf{x}_u)$
    else
      $\mathbf{x}_q \leftarrow \arg\min_{\mathbf{x}_u \in \mathcal{D}_U} S_R(\mathbf{x}_u) + \lambda_1 S_C(\mathbf{x}_u) + \lambda_2 S_D(\mathbf{x}_u)$
    $\mathcal{D}_Q \leftarrow \mathcal{D}_Q \cup \{\mathbf{x}_q\}$
    $\mathcal{D}_U \leftarrow \mathcal{D}_U - \{\mathbf{x}_q\}$
  for $t \in \{1, \ldots, E_S\}$ **do**
    for $\boldsymbol{x}_q \in \mathcal{D}_Q$ **do**
      for $i \in \{0, 1, 2, 3\}$ **do**
        $\phi' \leftarrow \phi' - \eta\nabla_{\phi'} \mathcal{L}(g_{\phi'}\,(\text{rot}_{90i}(\mathbf{x}_q)), i)$
  for $\boldsymbol{x}_u \in \mathcal{D}_U$ **do**
    Use $g_{\phi'}$ to compute and save $S_D(\mathbf{x}_u)$
Get oracle to label $\mathcal{D}_Q$ and update $\mathcal{D}_L$

core-set [27] was faster than PAL, but its performance was quite variable across the datasets as can be seen in Figure 2.

4.2. Robustness to labeling errors

We simulated labeling errors by randomly assigning incorrect labels to a subset of the labeled pool and the queried set. We performed experiments on the SVHN dataset, corrupting 20% of the data labels. We compared our technique to other active learning techniques and to random sampling, whose sampling performance is unaffected by label noise. In Figure 3, we observe that our technique fares better compared to the others tested. We attribute this robustness of PAL to the use of the pretext task in the scoring network.

4.3. Introduction of new classes

We performed experiments with a biased initial pool consisting of only eight out of the ten classes in the SVHN dataset. As seen in Figure 4, PAL is able to rapidly ramp up the performance when it is allowed to sample from the previously missing classes (after the initial 10% labels). In fact, it quickly catches up with its own strong performance on the unbiased initial pool case (i.e., the upper graph of Figure 4 is same as that of SVHN results in Figure 2). As shown in Figure 5, it procures samples from the previously missing classes more rapidly (over-samples) than random sampling and consequently, achieves higher overall accuracy at par with PAL that started with training on all ten classes. On the other hand, the representation of the two missing classes remains around 20% for random sampling, once those classes are available for queries, as expected.

4.4. Robustness to scoring network architecture

We now demonstrate that the key ideas behind the performance of PAL are what have been described thus far, and not the backbone architecture of the scoring network. We replaced ResNet-18 with VGG-16, and found no significant change in performance on the SVHN dataset, as shown in Figure 6.

4.5. Importance of the score components

We examine the effect of the different components of our score used to formulate the query by performing an ablation study. We compare performance of 1) without the diversity score S_D ($\lambda_2 = 0, \lambda_1 \neq 0$), 2) without the diversity score S_D and the supervision score S_C for querying with the pretext task based uncertainty score ($\lambda_1 = 0, \lambda_2 = 0$), and 3) the original scenario with both diversity and supervision included ($\lambda_1 \neq 0, \lambda_2 \neq 0$). We observe that using uncertainty estimates from both the pretext task and the classification task give a much better performance, as is hypothesized in Proposition 2. Adding the diversity score results in a further improvement in the performance, as is expected. These results are summarized in Figure 7.

5. Conclusion and Discussion

We proposed a new active learning method to estimate the novelty of unlabeled samples by introducing pretext-based scoring (using self-supervised learning). The proposed method uses the error in solving a self-supervised task as a proxy measure for the novelty of an unlabeled sample and departs from previous proposals by using an auxiliary scoring network that resolve potential conflicts between good task performance and query formation.

The scoring network is trained on the labeled samples in order to model their distribution instead of that of the entire data. Furthermore, the scoring network itself is trained in a multi-task manner, by including a supervised classification head to regularize and boost the performance of the self-supervised head. The multi-task training is necessary due to the small number of samples in active learning.

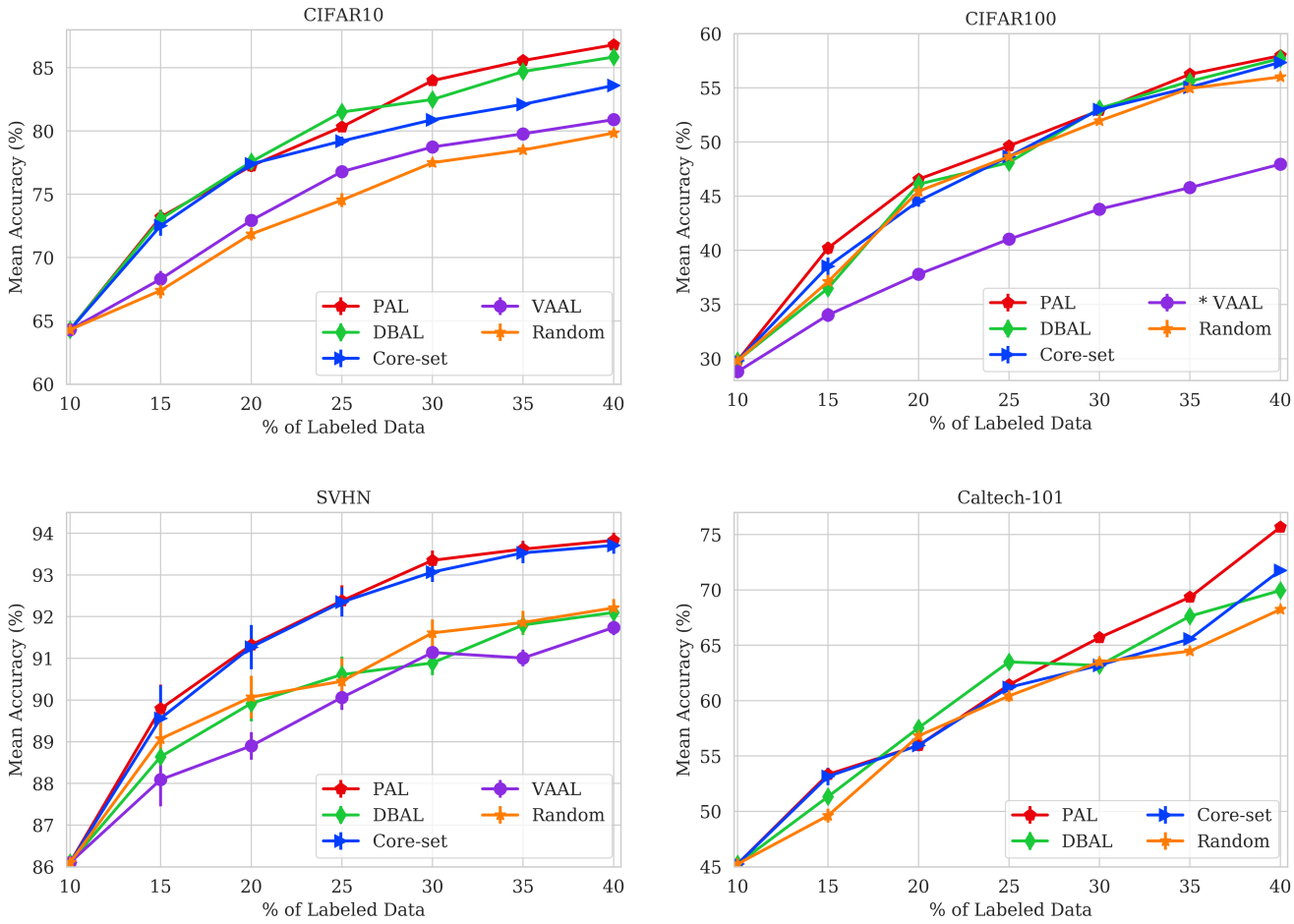

Figure 2. Performance of random sampling, VAAL [29], DBAL [10], and core-set [27] compared with PAL (proposed) on CIFAR-10, CIFAR-100, SVHN and Caltech-101. Markers show mean accuracy of five runs, and vertical bars show standard deviation (some are too small to be visible). **Note that VAAL takes prohibitively long to train due to the use of a VAE. Therefore, we report results on CIFAR-100 from the original paper, and exclude results of VAAL on Caltech-101.*

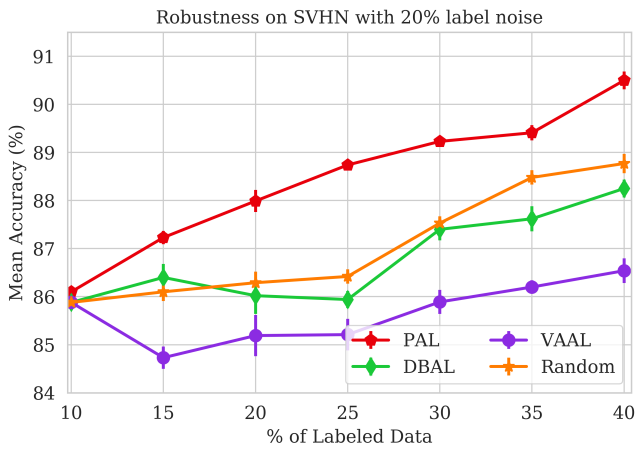

Figure 3. PAL performance on SVHN with 20% label noise.

This work presents early evidence that over-reliance on only one measure of uncertainty may not be judicious, and hybrid methods, where individual components compensate for each other, are likely to work better. A hybrid scoring method is able to break the reliance on labels, which may be noisy, for modeling the data distribution by going lower in the semantic hierarchy and tapping into the knowledge gained by self-supervision. A similar observation that self-supervised tasks add robustness in anomaly detection tasks has been reported previously [18]. Specifically, the proposed hybrid novelty scoring method uses KL divergence from uniform distribution of the classification head for a dynamic trade-off between classification uncertainty and difficulty of predicting rotation. We considered using entropy of class probabilities of the classification head instead of the KL divergence, but the range of entropy values and its derivative is finite. Thus, entropy would have been unable to

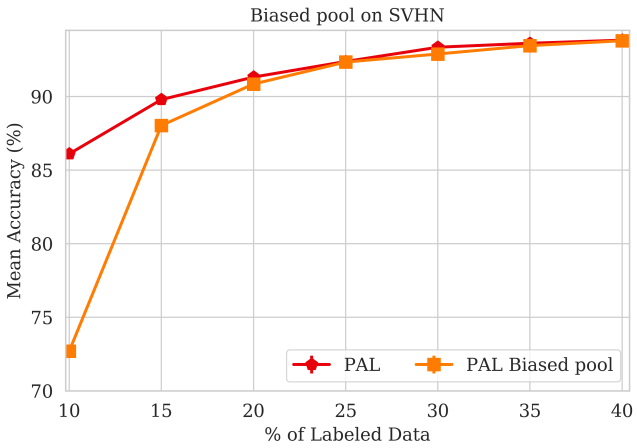

Figure 4. PAL performance on a biased initial pool (two classes missing for the initial 10% of the labeled samples) on SVHN. The upper graph is without a biased initial pool, and the lower graph catches up quickly when PAL is allowed to sample the missing classes in the queries.

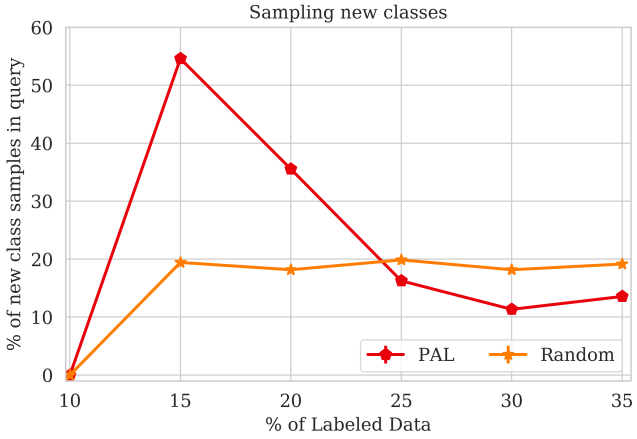

Figure 5. PAL oversamples data from new classes in the first query itself (total 15% of the dataset) as it finds it appropriately more novel than the previously overrepresented classes, before settling into a more equitable sampling by the third query (25% of the dataset).

dynamically balance out the shortcomings of the rotational head on samples on which it makes an error due to reasons other than good data representation. We empirically validated our hypotheses by showing strong performance of PAL on a variety of datasets. Furthermore, we also showed that PAL is robust to labeling an imperfect oracle and also performs well even when the initial labeled data pool has no samples from a few of the classes.

There is a need to balance between the twin goals of assessing novelty and diversity to select samples for the queries. By relying on novelty alone, there is a danger of picking a lot of novel samples that do not reasonably cover all regions of the data distribution. Conversely, methods

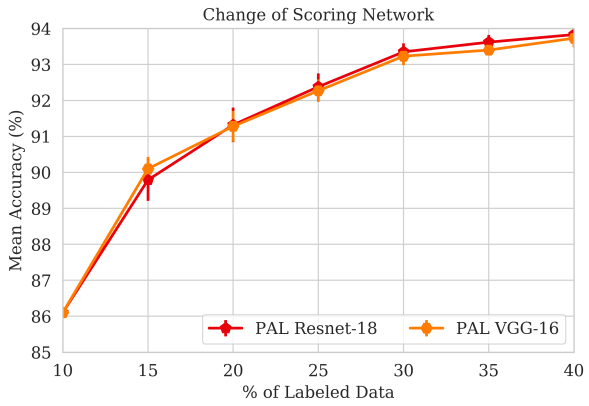

Figure 6. Performance of random sampling compared to PAL with Resnet-18 and VGG-16 as the backbone scoring network on SVHN shows robustness to change in network architecture.

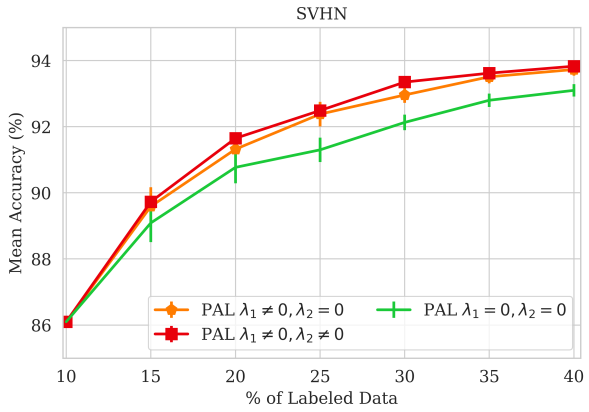

Figure 7. A comparison of performance of PAL using only self-supervision for uncertainty ($\lambda_1 = 0, \lambda_2 = 0$), self-supervision and supervision for uncertainty ($\lambda_1 \neq 0, \lambda_2 = 0$), and all three proposed components: self-supervision and supervision for uncertainty, and sub-query self-supervision for diversity ($\lambda_1 \neq 0, \lambda_2 \neq 0$).

that rely on diversity, such as core-set [27], can be hijacked by outliers in higher dimensions. While more research is needed to jointly pursue both goals, our method takes a sub-query approach to ensure that at least the samples in a sub-query are different from that of another sub-query.

References

[1] Dana Angluin. Queries and concept learning. In *Machine Learning*, volume 2, 1988. 2

[2] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[3] Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008. 2, 3

[4] Joachim Bingel and Anders Søgaard. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, 2017. 3

[5] L. Atlas D. Cohn and R. Ladner. Improving generalization with active learning. In *Machine Learning*, volume 15, 1994. 2

[6] Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 353–360. Curran Associates, Inc., 2008. 2

[7] Melanie Ducoffe and Frédéric Precioso. Adversarial active learning for deep networks: a margin based approach. *CoRR*, abs/1802.09841, 2018. 1, 2

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Wkshp*, 2004. 5

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, ., volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. 2

[10] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data, 2017. 1, 2, 5, 7

[11] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *CoRR*, abs/1703.02910, 2017. 2

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 3

[13] Ran Gilad-bachrach, Amir Navot, and Naftali Tishby. Query by committee made real. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 443–450. MIT Press, 2006. 2

[14] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9758–9769. Curran Associates, Inc., 2018. 1, 3

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[16] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019. 1, 2

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2016. 3

[18] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty, 2019. 1, 3, 7

[19] M. Huijser and J. C. V. Gemert. Active decision boundary annotation with deep generative models. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5296–5305, 2017. 2

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. 5

[21] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 840–849, 2017. 3

[22] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 3

[23] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 5

[24] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3

[25] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3

[26] Michael L Seltzer and Jasha Droppo. Multi-task learning in deep neural networks for improved phoneme

recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6965–6969. IEEE, 2013. 3

[27] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 1, 2, 5, 6, 7, 8

[28] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 2

[29] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 5, 7

[30] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, Mar. 2002. 2

[31] Donggeun Yoo and In So Kweon. Learning loss for active learning. ., 2019. 1

[32] X. Zhang, J. Zou, K. He, and J. Sun. Accelerating very deep convolutional networks for classification and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1943–1955, 2016. 5

[33] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014. 3

[34] Jia-Jie Zhu and José Bento. Generative adversarial active learning. *CoRR*, abs/1702.07956, 2017. 2