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

Abstract

When obtaining labels is expensive, the requirement of a large labeled training data set for deep learning can be mitigated by active learning. Active learning refers to the development of algorithms to judiciously pick limited subsets of unlabeled samples that can be sent for labeling by an oracle. We propose an intuitive active learning technique that, in addition to the task neural network (e.g., for classification), uses an auxiliary self-supervised neural network that assesses the utility of an unlabeled sample for inclusion in the labeled set. Our core idea is that the difficulty of the auxiliary network trained on labeled samples to solve a self-supervision task on an unlabeled sample represents the utility of obtaining the label of that unlabeled sample. Specifically, we assume that an unlabeled image on which the precision of predicting a random applied geometric transform is low must be out of the distribution represented by the current set of labeled images. These images will therefore maximize the relative information gain when labeled by the oracle. We also demonstrate that augmenting the auxiliary network with task specific training further improves the results. We demonstrate strong performance on a range of widely used datasets and establish a new state of the art for active learning. We also make our code publicly available to encourage further research.

1 Introduction

Deep neural networks have given quite promising results on benchmark datasets for common computer vision tasks, such as image classification, image segmentation, and object detection. However, a major drawback of the deep learning frameworks that act as a hurdle in their deployment for many real world problems is the requirement of a large amount of labeled training data from the target domain. Data labeling and annotations are often impractical, time-consuming, and expensive, especially when this process requires substantial time from experts in niche domains, such as medical image analysis.

Active learning seeks to develop algorithms to minimize labeling costs based on an access to a large set of unlabeled images. The idea is to select subsets of unlabeled samples to be sent for labeling to an oracle (e.g., a medical expert) such that the increase in performance (e.g., classification accuracy) is maximized. In a widely-studied setting, to optimize this performance for a budgeted number (i.e., a pool of size > 1) of samples to be labeled by the oracle, the sampling strategy should pursue two goals – representation and novelty (called coverage and uncertainty respectively in Sener and Savarese (2018)). Representation refers to selecting a batch of samples that are representative of the diversity of the data distribution so that various high probability regions can be discovered in an unbiased manner. Novelty refers to selecting those unlabeled samples that are least similar to the previously labeled samples, so that labeling them will maximize the information gain.

We propose to use the difficulty of solving an auxiliary self-supervised task on an unlabeled sample as a proxy for its novelty. The key idea here is that if a scoring neural network that is trained in a self-supervised manner (e.g., for predicting a random rotation of the image) on the *labeled samples* finds it difficult to perform the same self-supervision task on an unlabeled sample, then the unlabeled sample must be novel or out-of-distribution for it. By extension, the same unlabeled sample must also be informative for the task network (e.g. for classification), which is also trained on the labeled samples. Conversely, if the scoring neural network is able to perform the self-supervised task well on an unlabeled image, it means that the network is able to recognize something about the unlabeled such as its structure, color, pose, or deformation, and hence that unlabeled image will fetch little extra information when labeled by the oracle for the task network. Our approach thus departs from previous active learning strategies that use a single network for both the main task and for determining novelty, and uses self-supervised learning for active learning for the first time, according to the best of our knowledge.

In addition to proposing the idea of using self-supervised learning for estimating the novelty of an unlabeled sample, we also propose to add a safety and regularization mechanism in the auxiliary network. Because the auxiliary network is trained on a small set of labeled data, it benefits from a multi-task learning arrangement as a regularizer. We use the class prediction task itself as a second head for its regularization. This second head also allows more complete use of the information available about the labeled samples, which the self-supervision does not do. Additionally, we use a hybrid novelty scoring scheme for unlabeled samples that dynamically tilts more towards reliance on self-supervision when

labeling confidence is low, and forsakes this reliance when the labeling confidence is high (and self-supervision confidence is low, such as for images with rotational symmetry).

The proposed technique shows improved or matching results over the state-of-the-art on benchmark datasets for pool-based sampling methods (i.e., when labels for a fixed percent of additional samples are obtained in succession). Moreover, the proposed technique also performs competitively when the oracle mislabels some samples. Finally, the proposed technique also performs well even when the initial labeled data pool has no data points from a few of the classes (i.e., new classes can be added on the fly).

2 Related Work

Since our work brings ideas from self-supervised learning into active learning, we review works related to these two themes.

2.1 Active learning

The most common setting for active learning is *pool-based sampling*, where a large set of unlabeled data is available, from which the algorithm selects a pool of samples that can be sent to an oracle for labeling. The goal is to maximize the increase in some performance metric (e.g. classification accuracy) by optimally selecting a fixed number of samples for a pool, starting with an initial set of labeled samples. The proposed method fits this setting. Techniques for this setting can be further divided into two types – novelty-based and representation-based (along with their hybrids).

Novelty refers to an unlabeled sample's ability to provide new information, if labeled. Other terms used for the idea of novelty are uncertainty, confusion, perplexity, non-triviality, out-of-distribution, and informativeness etc. Novelty-based methods use a measure of additional new information of an unlabeled sample such as the entropy of a classifier neural network output (class probability mass function), distance from a decision boundary, expected risk, etc. (Settles 2009). In query-by-committee, multiple instances of a classifier network are trained on different subsets of data, and the variance on the outputs of learned classifiers is taken as a measure of uncertainty, and by extension, novelty (Giladbachrach, Navot, and Tishby 2006). Uncertainty estimations based on Bayesian frameworks have also been proposed for active learning. For instance, MC-dropout has been used as an uncertainty measure in deep and shallow neural networks (Gal and Ghahramani 2016; Gal, Islam, and Ghahramani 2017a). Distance from the decision boundary has also been used for active learning with support vector machines (Tong and Koller 2002). However, such a measure of uncertainty based on distance does not easily translate to convolutional neural nets where the decision boundary is often convoluted. As a remedy, the distance from an adversarial example has been proposed as an approximation of distance from the decision boundary (Ducoffe and Precioso 2018). However, this method is quite slow as finding the adversarial example, which requires perturbing the image based on the gradient of the loss with respect to its pixels, is a time-consuming process.

Representation-based methods seek to query samples that are diverse and can represent the data distribution more accurately. A method based on identifying a core-set has been proposed (Sener and Savarese 2018). This method chooses a pool of query samples such that the empirical loss over the set of already labeled samples combined with the pool of query samples is as close as possible to the empirical loss over the whole dataset. It was shown that for networks with loss functions that obey Lipschitz-continuity, such a pool is given by the points with maximum L-distance, where L is the distance from the nearest labeled point. However, this approach suffers when the representations are high-dimensional, since the Euclidean distance is a poor local distance estimator in high dimensional spaces, especially when the underlying data does not lie close to a low-dimensional manifold. An alternative approach called variational adversarial active learning (VAAL) tries to learn a good representation using a variational autoencoder (VAE) (Sinha, Ebrahimi, and Darrell 2019). The latent variable of the VAE is trained adversarially using a discriminator, which tries to predict if a sample is already labeled. Data points having a representation that is closer to samples in the unlabeled pool as compared to the labeled pool are selected for labeling.

Other settings for active learning include *stream-based pooling*, in which the unlabeled dataset is evaluated in an streaming fashion. With each incoming unlabeled data, the algorithms should decide whether to query the oracle or discard the sample, one at a time. Online updates of an explicit region of uncertainty (based on a measure of novelty) have been proposed in which samples that fall in this region are queried, while those outside the region are discarded (D. Cohn and Ladner 1994; Dasgupta, Hsu, and Monteleoni 2008). These approaches are quite relevant and useful in real-time real-world scenarios such as speech tagging (Dagan and Engelson 1995) and sensor scheduling (Krishnamurthy 2002), but they are not able to use more complete information available about the unlabeled samples that are available in a pool-based setting.

Another setting for active learning is *membership query synthesis*, in which the learner generates new samples to query the oracle (Angluin 1988), including by using generative adversarial networks (Zhu and Bento 2017; Huijser and Gemert 2017), instead of sampling from an existing pool on unlabeled data. A method to generate interpretable queries by restricting the hamming distance of a generated query with a randomly sampled training instance has also been proposed (Awasthi and Kanade 2012). However these approaches have had limited success because of the difficulty in interpreting and subsequently labeling the generated queries even by human annotators, specially in niche data domains such as medical imaging. Additionally, these techniques do not make use of unlabeled data, which is often available in today's age.

2.2 Self-supervised learning

Self-supervised learning (SSL) has shown great promise in learning good data representations without the requirement of explicit data labels. The idea is to learn the underlying domain-specific spatial structure of unlabeled images using

a neural network, and then transferring the learning to the actual supervised task, so that higher accuracy can be achieved with limited labeled data. Most of the proposed SSL techniques create a supervised pretext task based on unlabeled data. Labels for the pretext task can be generated automatically without the need of a human annotator. Some of these pretext tasks synthetically degrade an unlabeled image and train a neural network to recover the original image. The choice of the pretext task is based on the know-how of the input domain. Some commonly used degradations are removing color (Larsson, Maire, and Shakhnarovich 2017), reducing resolution (Ledig et al. 2017), and occluding parts (Pathak et al. 2016) of an image. Other pretext tasks apply randomized transforms to images and train neural networks to predict the transform. These transforms include jumbling the spatial order of sub-images of an image (Noroozi and Favaro 2016), jumbling the sequences of frames of videos (Misra, Zitnick, and Hebert 2016), and applying randomized geometric transforms (Gidaris, Singh, and Komodakis 2018).

Self-supervised learning techniques not only reduce the labeling requirement by training domain-specific feature extraction layers, these can also be used to assess the degree to which a given image is predictable (i.e., likely to belong to the data distribution), in a probabilistic sense. That is, if a synthetically degraded version of an image can be restored close to its original, or if the applied random transform can be correctly predicted by an auxiliary neural network, then it follows that images similar to it must have been encountered during the training of the auxiliary network. This idea has been exploited in identifying out-of-distribution (OOD) samples for its own sake (Hendrycks et al. 2019; Golan and El-Yaniv 2018). We use this idea to identify unlabeled images that are most unlike the previously labeled images, and therefore likely to yield maximize gain in information when the oracle is queried for their labels in an pool-based active learning setting.

3 Method

Active learning techniques need to select a group of most informative samples to be picked for labeling from a set of unlabeled samples for maximal improvement on a supervised learning task. Let the pool of currently labeled samples be D_L and the pool of unlabeled samples be D_U. Pool-based sampling involves selecting a budgeted set of N samples from D_U in each query. The queried samples are then labeled by an oracle (assumed to be ideal, unless otherwise stated), added to D_L, and removed from D_U. This process is repeated followed by labeling and addition to the labeled pool D_L until a desired number of samples are sampled, or a desired level of accuracy is achieved. This iterative process is the basic essence of pool-based active learning.

Previously proposed methods used for estimating the uncertainty of an unlabeled sample reuse the task model itself, e.g., entropy of the computed class probability mass function. As shown in Figure 1, We use a different model than the task model for our selection strategy, which we refer to as the scoring model hereafter. The scoring model has two heads – a self-supervision head and a classification head – whose outputs are used to assign a *confusion* score to the unlabeled query image. We describe both the heads next.

3.1 Self-supervision head

The self-supervision head serves as a means to estimate the likelihood of the unlabeled data to be sampled from the distribution of the labeled samples. We use a particular set of geometric transforms for our self-supervised task – rotations by multiples of $90°$– because these transformations do not introduce any kind of artifacts in images. During the training phase, this head is trained on the three rotated and the original versions of the data – that is, on rotation by $0°$, $90°$, $180°$ and $270°$.

The self-supervised training is done only on the data points from D_L so that the head learns the labeled data distribution. Based on the output of the head, a confusion score is assigned to each unlabeled query image, as given by:

$$S_R(x) = - \sum_{i \epsilon \{0,1,2,3\}} f_{\theta_S,R}(x_i)_i, \qquad (1)$$

where $f_{\theta_S,R}$ is the scoring model (self supervision head) parametrized by θ_S, x_i is the data point x rotated by $90i$ degrees, $f_{\theta_S,R}(x)_i$ is the softmax output (rotation probability mass function) of the head for the rotation $90i$ for a input x. We hypothesized that on an unlabeled data point from D_U on which this head predicts all the four applied rotations correctly, score S_R will be closer to -4, and will likely be similar to the labeled points. Conversely, for an OOD unlabeled point the rotation head is unlikely to predict any of the rotation correctly, and the score would be nearer to 0, as it would not know their semantic features such as pose and structure. The strategy would be to pick those unlabeled points for the next pool to be labeled, on which this head exhibits the greatest confusion score S_R.

3.2 Classification head

Using a scoring function just based on the self-supervised task of predicting the applied rotation is sub-optimal because of the following reasons. Firstly, The labels of D_L are left unused for building the active learning strategy. Moreover, the score S_R is not robust to the data points which have a rotational symmetry, which leads to a poor performance and a high score in the transform prediction head, although the network might know all about the lower semantic features of that image. For the reasons state above, we use another measure of confusion (novelty), which is the degree to which the outputs of a classification head are closer to a uniform probability mass function. That is, from S_R we subtract the KL divergence of the softmax outputs of the classification head from a uniform distribution as shown below:

$$S_{hybrid}(x) = S_R(x) - \lambda \, \mathrm{KL}(U \mid\mid f_{\theta_S,C}(x)) \qquad (2)$$

where, $f_{\theta_S,C}$ is the scoring model (classification head) trained on the target set of classes on the labeled samples and parametrized by θ_S, $f_{\theta_S,C}(x)$ is the softmax (probability mass function) output of the head for input x, U is the uniform probability distribution over all classes, and $\lambda > 0$ is a relative importance hyperparameter.

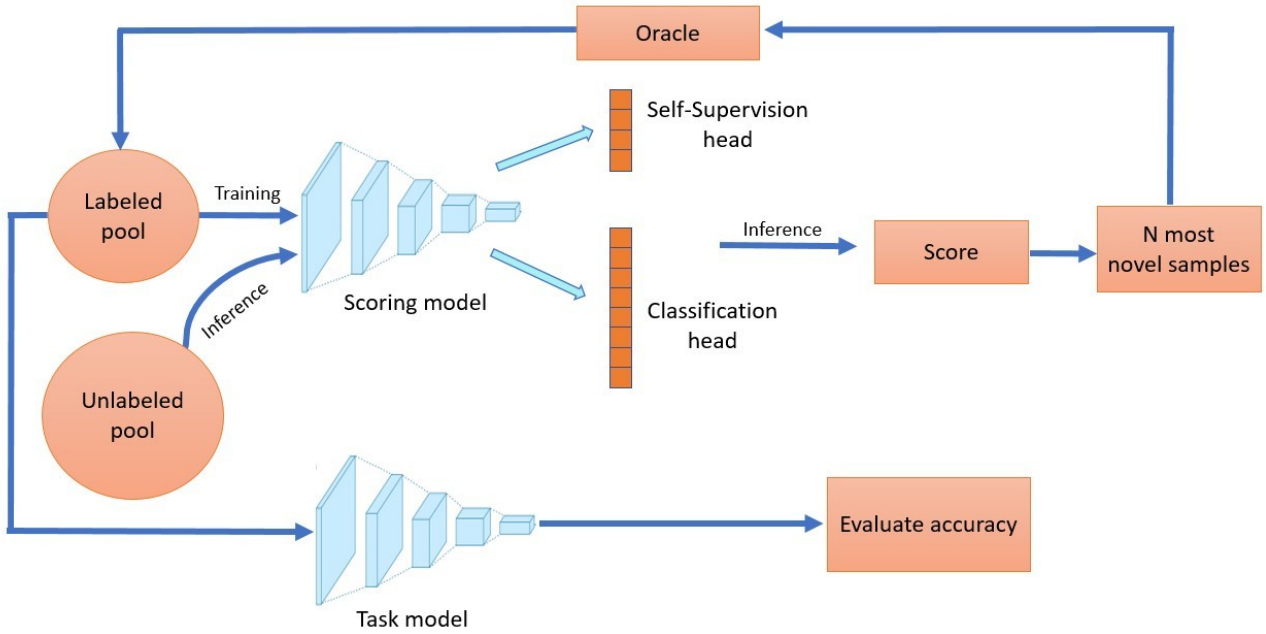

Figure 1: The proposed network pretext-based active learning (PAL) has a scoring model in addition to a task model. The two heads of the scoring model are trained jointly on the labeled data, which are then used to assign novelty scores to the unlabeled data points. The N most novel samples are labeled by an oracle.

The KL divergence with uniform distribution would be low for the samples on which the classification head performs poorly and is therefore an alternative measure of confusion of the sample, as previously described (Hendrycks and Gimpel 2016). In Section 5, we discuss why this measure of uncertainty is qualitatively different from the often-used entropy of the class probability mass function.

Finally, the most informative samples which we would want to pick and get labeled by the oracle will either have a high S_R score due to poor performance in the rotation prediction head or a very low divergence with a uniform distribution due to low confidence predictions by the classification head. Hence, we pick the samples from D_U which have the highest S_{hybrid} score and add them to D_L. Algorithm 1 details one query round of our method dubbed *pretext-based active learning (PAL)*.

3.3 Training and loss function

The proposed scoring model is trained using a multi-task loss. The loss has two components – one for the self-supervision (rotation prediction) task and another for the classification task.

Let f_{θ_S} be the scoring model parameterized by θ_S, with $f_{\theta_S,R}$ being the self-supervision head and $f_{\theta_S,C}$ being the classification head. Then we estimate the optimal parameter θ_S^{opt} as : $\min_{\theta_S} \ell_S(\theta_S)$, where,

$$\ell_S(\theta_S) = \sum_{i \in \{0,1,2,3\}} \ell(f_{\theta_S,R}(x_i), i) + \lambda\, \ell(f_{\theta_S,C}(x), y) \tag{3}$$

where $\lambda > 0$ is a hyperparameter, $(x, y) \in D_L$, x_i is the data point x rotated by $90i$ degrees, and ℓ is the cross-entropy loss.

Algorithm 1 Pretext-based Active Learning Query Round

Inputs: Labeled pool $D_L := \{X_L, Y_L\}$, Unlabeled pool $D_U := \{X_U\}$
Parameters: Initialized task model $f_{\theta_T}(X)$ with parameters θ_T and scoring model $f_{\theta_S}(X)$ with parameters θ_S, sampling budget N
Output: Trained task model f_{θ_T}
Training Task and Scoring Model:
  for $n = 1, \ldots,$ epochs
    sample a training pair (x_l, y_l) (or a mini-batch) from D_L
    $x_{l,i} \leftarrow x_l$ rotated by $90i°, i \in \{0, 1, 2, 3\}$
    $\theta_S \leftarrow \theta_S - \alpha(\nabla_{\theta_S} \ell_S(f_{\theta_S}(x_{l,i}), (i, y_l))$
    $\theta_T \leftarrow \theta_T - \alpha \nabla_{\theta_T} \ell(f_{\theta_T}(x_l), y_l)$
  end for
Sampling Strategy:
  $S_U \leftarrow \{S_{hybrid}(x), x \in D_U\}$
  $X_{\text{query}} \leftarrow \arg\min_{x}{}_N\ S_U$
  $Y_{\text{query}} \leftarrow \text{Oracle}(X_{\text{query}})$
  $D_L \leftarrow D_L \cup \{X_{\text{query}}, Y_{\text{query}}\}$, $D_U \leftarrow D_U - X_{\text{query}}$

4 Empirical Evaluation

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

compared and the implementation details.

Datasets: We performed experiments on four datasets: (1) SVHN (Netzer et al. 2011), 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 (Krizhevsky 2009), 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 (Krizhevsky 2009), 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 (Fei-Fei, Fergus, and Perona 2004), 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 (Sinha, Ebrahimi, and Darrell 2019). (3) *DBAL:* This method uses Bayesian CNNs to estimate uncertainty of unlabeled points and to pick the most uncertain samples (Gal, Islam, and Ghahramani 2017b). (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 (Sener and Savarese 2018).

Experimental setup: For all the baselines and datasets, we trained the model initially with a randomly sampled labeled pool of 10% of the whole dataset. All strategies started with the same pool of initially labeled samples when being compared. We fixed the query size to 5% of the whole dataset. We assume that the labeling by oracle is error-free, unless stated otherwise. A classifier is trained on the labeled pool after each query and its accuracy is evaluated. The labeled pool thus increases by 5% of the total data after each query, and all models are retrained on the new expanded labeled dataset from scratch in each query round.

An average performance of five runs with random initialization of the networks was taken while reporting the results. For fairness of comparison, we used VGG16 (Zhang et al. 2016) as the classifier model for all the baselines, in line with Sinha, Ebrahimi, and Darrell (2019); Sener and Savarese (2018).

For the scoring model of the proposed PAL approach, we used ResNet-18 (He et al. 2016) as the architecture to learn the self-supervised geometric transform prediction task. Relative importance hyperparameter λ (equation 3) was chosen from $\{0.5\ ,\ 1.0\}$. We tuned the classifier learning rate in the range $[10^{-1}, 10^{-4}]$ and we also selected the optimizer $\in \{\text{Adam, SGD}\}$. The experiments were run on an Intel(R) i7 CPU with 6 cores clocked at 3.60 GHz and with NVIDIA GeForce GTX 1080 GPU running CUDA 10.2 and cuDNN 7.6.

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 and random sampling. Our PAL strategy outperformed random sampling by a wide margin and consistently seems to outperform VAAL (Sinha, Ebrahimi, and Darrell 2019), DBAL (Gal, Islam, and Ghahramani 2017b), and core-set (Sener and Savarese 2018). 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.

4.2 Robustness to noise

Real labeling processes are prone to human error. Natural language processing is also being used to mine labels from unstructured text associated with image repositories and web crawls, which also leads to label noise. Therefore, an active learning strategy should be resilient to such noise. We model label noise by using an imperfect oracle which assigns incorrect labels to a random subset of query points. We hypothesize that PAL is more resilient to such noise as compared to other techniques because it has a label-independent rotation prediction head. A wrongly labeled image may cause active learning strategies that depend on the label alone to degrade a lot in performance. A good sampling technique should be able to instead make out truly confusing samples in presence of label noise.

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 auxiliary rotation task. Less informative images in the unlabeled pool (similar to an incorrectly labeled image in the labeled pool) would still be scored low by the rotation head, unaffected by the label noise, and thus allow more informative images to be picked. A similar observation that self-supervised tasks add robustness in anomaly detection tasks has been reported previously (Hendrycks et al. 2019).

4.3 Biased initial pool

All experiments reported so far use randomly selected samples as the initial labeled pool. This makes the initial pool likely to have all classes of data in representative proportions. However, in practice, this might not be true, and data from several classes may be underrepresented or worse, absent in the *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. A good active learning strategy should rapidly procure samples from such classes to be labeled and match its performance to the scenario where such classes were well-represented in the initially labeled data.

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

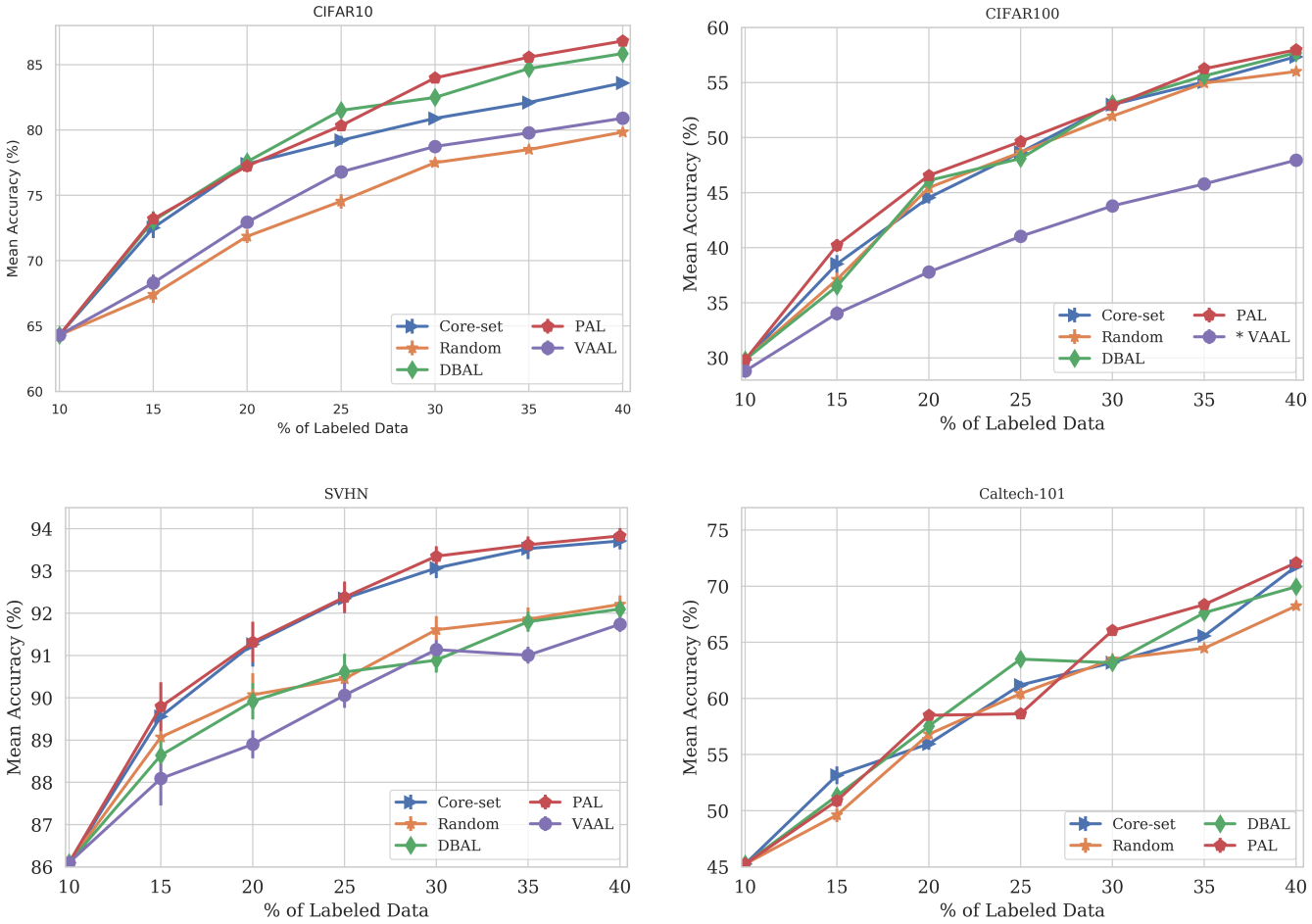

Figure 2: Performance of random sampling, VAAL (Sinha, Ebrahimi, and Darrell 2019), DBAL (Gal, Islam, and Ghahramani 2017b), and core-set (Sener and Savarese 2018) compared with PAL (proposed) on CIFAR-10, CIFAR-100, SVHN and Caltech-101. Markers show mean accuracy of five runs, and vertical bars show 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.*

up the performance when it is allowed to sample from the previously missing classes as well, and matches the performance of the unbiased initial pool case. 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.

5 Conclusion and Discussion

We proposed a new method to measure novelty of an unlabeled point by introducing pretext-based active learning (PAL) to develop effective sampling strategies for active learning. The proposed method uses the difficulty in solving a self-supervised task as a proxy measure for the novelty of an unlabeled sample. For this purpose, it departs from previous proposals to estimate novelty by using an auxiliary neural network called the scoring network.

The scoring network is trained on the labeled samples, even though it is learning in a self-supervised manner, so that it models the distribution of the labeled samples alone. Further, the scoring network itself is trained in a multi-task setting, by including a supervised classification head, to regularize and boost the performance of the self-supervised head.

Among other techniques that were compared, it remains unclear as to which method of determining uncertainty is better, apart from empirical evidence, as shown in this paper. This lack of clarity stems from the lack of visibility into a good representation of the unlabeled samples, which in turn will shape the data distribution to which the labeled data can be compared. However, the paper presents early evidence that over-reliance on only one measure of uncertainty may not be judicious, and hybrid methods where individual components compensate where the others fail are likely to

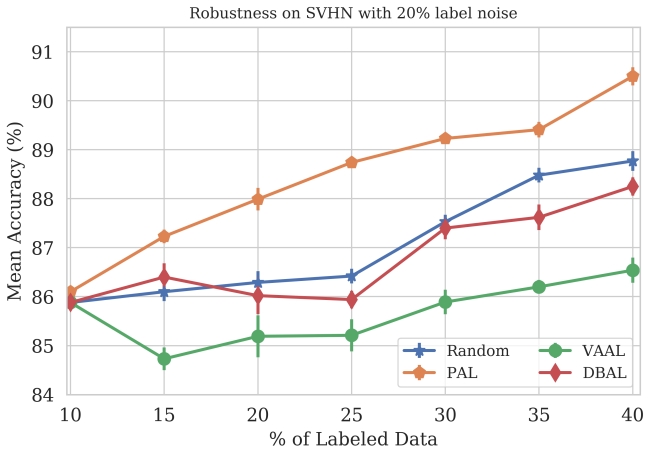

Figure 3: PAL performance on SVHN with 20% label noise

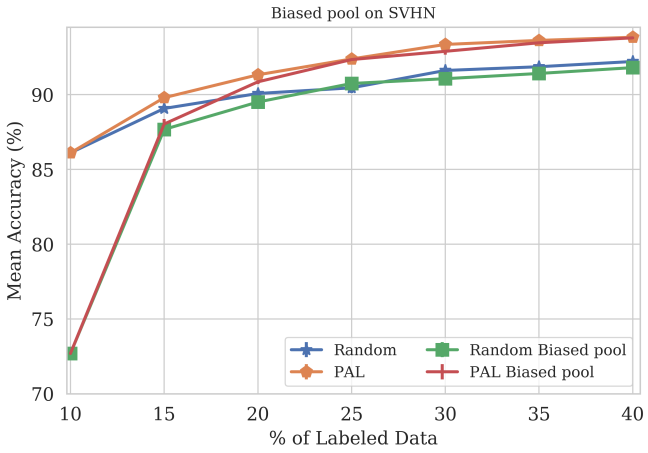

Figure 4: PAL performance on a biased initial pool (two classes missing for the initial 10% of the labeled samples) on SVHN

work better. 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. The KL divergence has an infinite derivative and infinite dynamic range when the prediction confidence (probability) of one of the class tends to 1. Consequently, such a sample is deemed certain, and is given a very low novelty score as the main task network will also most likely perform well on such a sample. 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. Entropy of class predictions is therefore unable to dynamically balance out the shortcomings of the cross-entropy of rotation predictions, when the need arises, when the rotation of an unlabeled sample cannot be predicted but its classification is relatively certain.

We empirically validated our hypotheses by showing strong performance of PAL on a variety of computer vision

Figure 5: PAL rapidly samples data from new classes till they are proportionately present in the labeled pool. (Zeroth query is the initial labeled pool. Query number shown instead of 5% increments of labeled data to avoid confusion.)

datasets. We also showed that PAL is robust to a noisy oracle and also performs well even when the initial labeled data pool has no data points from a few of the classes. These robustness experiments, once again, demonstrate that a hybrid scoring method is able to break the reliance on labels, which may be noisy, by going lower in the semantic hierarchy and tapping into the knowledge gained by self-supervision.

Better methods are needed to balance between the twin goals of using novelty as well as representation to generate samples for the queries. By relying on novelty alone, there is a real danger of picking a lot of novel samples that do not reasonably cover all regions of the data distribution. Conversely, methods that rely on representation, such as coreset (Sener and Savarese 2018), can be hijacked by outliers.

6 Ethical Impact

We believe that research on active learning solves a very important problem encountered while developing AI solutions for the real world. Domains where labeling is expensive, such as medical image analysis, remain a stumbling block in the democratization of AI, which could be serving under-served communities and patients. To the best of our knowledge and imagination, we do not foresee a potential negative societal impact of our work, or active learning in general. We caveat this by adding that the due diligence applicable to machine learning, such as auditing for data and outcome biases, is also applicable to active learning.

References

Angluin, D. 1988. Queries and Concept Learning. In *Machine Learning*, volume 2.

Awasthi, P.; and Kanade, V. 2012. Learning using Local Membership Queries under Smooth Distributions. *CoRR* abs/1211.0996. URL http://arxiv.org/abs/1211.0996.

D. Cohn, L. A.; and Ladner, R. 1994. Improving generalization with active learning. In *Machine Learning*, volume 15.

Dagan, I.; and Engelson, S. P. 1995. Committee-Based Sampling for Training Probabilistic Classifiers. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, 150–157. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1558603778.

Dasgupta, S.; Hsu, D. J.; and Monteleoni, C. 2008. A general agnostic active learning algorithm. In Platt, J. C.; Koller, D.; Singer, Y.; and Roweis, S. T., eds., *Advances in Neural Information Processing Systems 20*, 353–360. Curran Associates, Inc. URL http://papers.nips.cc/paper/3325-a-general-agnostic-active-learning-algorithm.pdf.

Ducoffe, M.; and Precioso, F. 2018. Adversarial Active Learning for Deep Networks: a Margin Based Approach. *CoRR* abs/1802.09841. URL http://arxiv.org/abs/1802.09841.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *Computer Vision and Pattern Recognition Workshop* .

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. volume 48 of *Proceedings of Machine Learning Research*, 1050–1059. New York, New York, USA: PMLR. URL http://proceedings.mlr.press/v48/gal16.html.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017a. Deep Bayesian Active Learning with Image Data. *CoRR* abs/1703.02910. URL http://arxiv.org/abs/1703.02910.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017b. Deep Bayesian Active Learning with Image Data.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=S1v4N2l0-.

Gilad-bachrach, R.; Navot, A.; and Tishby, N. 2006. Query by Committee Made Real. In Weiss, Y.; Schölkopf, B.; and Platt, J. C., eds., *Advances in Neural Information Processing Systems 18*, 443–450. MIT Press. URL http://papers.nips.cc/paper/2916-query-by-committee-made-real.pdf.

Golan, I.; and El-Yaniv, R. 2018. Deep Anomaly Detection Using Geometric Transformations. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*, 9758–9769. Curran Associates, Inc. URL http://papers.nips.cc/paper/8183-deep-anomaly-detection-using-geometric-transformations.pdf.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hendrycks, D.; and Gimpel, K. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks.

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty.

Huijser, M.; and Gemert, J. C. V. 2017. Active Decision Boundary Annotation with Deep Generative Models. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 5296–5305.

Krishnamurthy, V. 2002. Algorithms for optimal scheduling and management of hidden Markov model sensors. *IEEE Transactions on Signal Processing* 50(6): 1382–1397.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2017. Colorization as a Proxy Task for Visual Understanding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 840–849.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 105–114.

Misra, I.; Zitnick, C. L.; and Hebert, M. 2016. Unsupervised Learning using Sequential Verification for Action Recognition. *CoRR* abs/1603.08561. URL http://arxiv.org/abs/1603.08561.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *ECCV*.

Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; and Efros, A. 2016. Context Encoders: Feature Learning by Inpainting.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=H1aIuk-RW.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational Adversarial Active Learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Tong, S.; and Koller, D. 2002. Support Vector Machine Active Learning with Applications to Text Classification. *J.*

Mach. Learn. Res. 2: 45–66. ISSN 1532-4435. doi:10.1162/153244302760185243. URL https://doi.org/10.1162/153244302760185243.

Zhang, X.; Zou, J.; He, K.; and Sun, J. 2016. Accelerating Very Deep Convolutional Networks for Classification and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(10): 1943–1955.

Zhu, J.; and Bento, J. 2017. Generative Adversarial Active Learning. *CoRR* abs/1702.07956. URL http://arxiv.org/abs/1702.07956.

A KL divergence vs Entropy

Proposition 1: *KL-divergence with uniform is better suited than entropy as a score combination with self supervision based task score.*

Consider a binary classification problem for analysis, with p as the predicted probability score by the task model for the correct class. Let S_1 be the novelty score based on the hybrid of self supervision and entropy and let S_2 be the one based on the hybrid of self supervision and KL divergence.

We calculate the scores for an unlabelled sample which is classified confidently by the task model, that is with p $\rightarrow$ 1.

$$\begin{aligned}\lim_{p\to 1} S_1 &= \lim_{p\to 1} (S_R - \lambda p \log(p) - \lambda(1-p)\log(1-p)) \\ &= \lim_{p\to 1} S_R - 0 - \lim_{p\to 1} \lambda(1-p)\log(1-p) \\ &= S_R^* - \lambda \lim_{p\to 1} \frac{\log(1-p)}{\frac{1}{(1-p)}}\end{aligned}$$

Using L'Hopitals rule for evaluating the limit

$$\begin{aligned}&= S_R - 0 \\ &= S_R\end{aligned}$$

$$\begin{aligned}\lim_{p\to 1} S_2 &= \lim_{p\to 1} \left(S_R - \lambda\, 0.5\, \log\left(\frac{0.5}{p}\right) - \lambda\, 0.5\, \log\left(\frac{0.5}{1-p}\right)\right) \\ &= \lim_{p\to 1} S_R - \lim_{p\to 1} \left(\lambda\, 0.5\, \log\left(\frac{0.5}{p}\right) + \lambda\, 0.5\, \log\left(\frac{0.5}{1-p}\right)\right) \\ &= S_R^* - 0.5\,\lambda\, \log(0.5) - 0.5\,\lambda \lim_{p\to 1} \log\left(\frac{0.5}{1-p}\right) \\ &= -\infty\end{aligned}$$

*S_R is the score from the scoring model which might perform poor even when task model performs well (for ex. symmetrical images), although $S_R \in \{0, -4\}$ and hence is finite.

Clearly, the KL divergence based score is able to drop down the novelty score to very low values for a sample on which the task model is confident which is not the case with entropy based score which limits to S_R which can be both high or low.

B Invariance to model architecture

PAL uses a scoring network trained using self supervision for predicting a novelty score for each unlabeled sample, with its key idea being the training technique.PAL is invariant to changes in the scoring networks architecture and works well even if a different backbone architecture is used. We demonstrate this by performing experiments using VGG-16 as the scoring model in place of Resnet-18 used. They both show a similar performance on SVHN and CIFAR-10 datasets, as seen in Figure 1.

C Hyperparameters

We share the hyperparameters used for training the task and the scoring models for our different experiments in Table 1. All hyperparameters were obtained through a grid search.

Dataset	α_T	α_S	**task & scoring model epochs**	**batch size**	λ	**optim**
CIFAR-10	0.01	0.01	100	64	1	SGD
CIFAR-100	0.01	0.1	100	64	1	SGD
SVHN	0.01	0.01	100	64	1	SGD
Caltech-101	0.01	0.01	100	32	1	SGD

Table 1: Parameters for experiments on various datasets

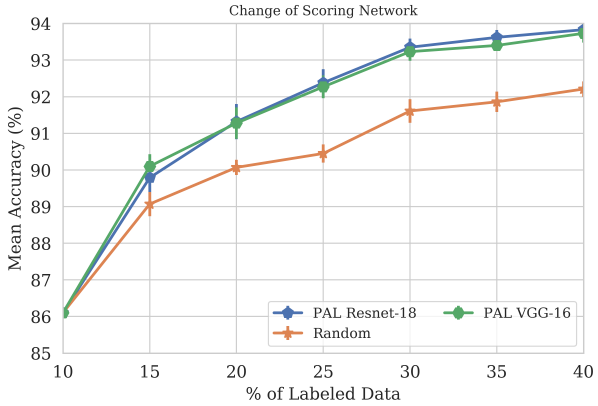

Figure 6: Performance of random sampling compared to PAL with Resnet-18 and VGG-16 as the scoring network on SVHN. A change of scoring model architecture does not affect the method by much