Task-Aware Variational Adversarial Active Learning

Kwanyoung Kim [1]   **Dongwon Park** [1]   **Kwang In Kim** [2]   **Se Young Chun** [1]

Abstract

Deep learning has achieved remarkable performance in various tasks thanks to massive labeled datasets. However, there are often cases where labeling large amount of data is challenging or infeasible due to high labeling cost such as labeling by experts or long labeling time per large-scale data sample (e.g., video, very large image). Active learning is one of the ways to query the most informative samples to be annotated among massive unlabeled pool. Two promising directions for active learning that have been recently explored are data distribution-based approach to select data points that are far from current labeled pool and model uncertainty-based approach that relies on the perspective of task model. Unfortunately, the former does not exploit structures from tasks and the latter does not seem to well-utilize overall data distribution. Here, we propose the methods that simultaneously take advantage of both data distribution and model uncertainty approaches. Our proposed methods exploit variational adversarial active learning (VAAL), that considered data distribution of both label and unlabeled pools, by incorporating learning loss prediction module and RankCGAN concept into VAAL by modeling loss prediction as a ranker. We demonstrate that our proposed methods outperform recent state-of-the-art active learning methods on various balanced and imbalanced benchmark datasets.

1. Introduction

Deep learning has achieved remarkable performance in various computer vision tasks such as image classification (Krizhevsky et al., 2012; He et al., 2016), object detection (Ren et al., 2015; Redmon et al., 2016), semantic segmentation (Long et al., 2015; Chen et al., 2018), and so

on thanks to massive labeled datasets with annotations such as ImageNet dataset for image classification (Deng et al., 2009) and PASCAL VOC datasets for classification, detection, and segmentation (Everingham et al., 2010). Obtaining good annotations is a challenging task and has often been a large-scale project. However, there are often cases where labeling large amount of data is even more challenging or often infeasible due to high labeling cost such as labeling by experts (Esteva et al., 2017) or long labeling time per large-scale data sample such as very large pathology images (Campanella et al., 2019). High labeling cost seems to be one of the factors to limit the scope of applicability of deep learning to more research areas and more institutes with less labeling budget.

There are several approaches to overcoming limited labeling budget such as utilizing small amount of labeled data (few-shot learning) (Ravi & Larochelle, 2017), exploiting both small amount of labeled data and large amount of unlabeled data (semi-supervised learning) (Zhu et al., 2003; Kingma et al., 2014), and selecting data to label for the best possible performance (active learning) (Settles, 2009; Gal et al., 2017). While the first two approaches are applicable after labeling, the last approach is applicable before labeling. Active learning has been widely investigated in machine learning. While active learning has been extensively in relatively traditional machine learning settings (non-deep learning approaches) (Cohn et al., 1996; Tong & Koller, 2001; Brinker, 2003; Melville & Mooney, 2004; Nguyen & Smeulders, 2004; Settles, 2009; Houlsby et al., 2011; Mac Aodha et al., 2014; Wang & Ye, 2015; Sundin et al., 2019; Pinsler et al., 2019) and has recently been investigated actively in deep learning settings (Gal et al., 2017; Sener & Savarese, 2018; Yang et al., 2017; Yoo & Kweon, 2019; Tran et al., 2019; Sinha et al., 2019; Kirsch et al., 2019).

In this work, we revisit two recent state-of-the-art active learning methods that are learning loss (Yoo & Kweon, 2019) and variational adversarial active learning (VAAL) (Sinha et al., 2019) and investigated active learning methods to exploit the pros of both methods. While VAAL is exploiting the feature structures of both labeled and unlabeled dataset in a task-agnostic way without any information from task learner (Sinha et al., 2019), learning loss is utilizing loss prediction module to predict losses of task learner for unlabeled data (Yoo & Kweon, 2019). We

---

[1]Department of Electrical Engineering [2]Department of Computer Science & Engineering, UNIST, Ulsan, Republic of Korea. Correspondence to: Se Young Chun <sychun@unist.ac.kr>.

propose to incorporate task-related information on predicted loss into VAAL using a concept from RankCGAN (Saquil et al., 2018) so that both global data structure from VAAL and local task-related information from learning loss can be used to select samples to annotate. We evaluated our proposed methods on various benchmark datasets such as CIFAR-10, SVHN, Fashion-MNIST that have the same number of images per class (balanced datasets) as well as our modified CIFAR-10 that had different numbers of images for classes (imbalanced). Here are the contributions:

1. Revisiting learning loss work / VAAL work and proposing new active learning method called task-aware VAAL (TA-VAAL) to exploit both global data structure from labeled / unlabeled datasets as well as local task-aware information from loss prediction module for active learning with deep neural networks,

2. Extensive empirical analyses on the proposed method on both balanced datasets and imbalanced dataset not only using performance measures, but also using information theoretic measures for the proposed active sampling method, showing its state-of-the-art performance as well asl reliable active sampling results.

2. Related Work

There have been many works on active learning to select the most informative data points and we categorized them into two different approaches: model certainty-based approach and data distribution-based approach. The first approach is using unlabeled data in a passive way while the second approach is using unlabeled data in an active way. In other words, the former has sample selection rules that are not affected by unlabeled data, but simply apply those rules to unlabeled data, while the latter exploit both labeled and unlabeled data to build up (or to train) sample selection rules (or deep neural networks).

Model uncertainty-based approach defined and used metrics for sample selection with labeled data. For example, the minimal distance from decision boundaries (or classification hyperplanes) can be used to select samples with the most ambiguous classification results (Tong & Koller, 2001; Brinker, 2003). Empirical risk minimization is used to minimize an upper bound of the true risk so that one can query the most informative samples that are the most uncertain and representative (Wang & Ye, 2015). Bayesian active learning by disagreement (BALD) maximizes the mutual information between model predictions and model parameters for active learning (Houlsby et al., 2011). However, these method were not well-suited to deep neural networks.

Recently, there have been some works on model uncertainty-based active learning with deep neural networks. BALD was extended to accommodate deep neural networks with

Bayesian neural network and Monte-Carlo dropout (Gal et al., 2017). As a practical active learning method, two-step approach was proposed to select the K-most uncertain samples by using bootstrapping and and then to find the most representative subset among K samples by using greedy method (Yang et al., 2017). Bayesian generative active deep learning was proposed to utilize both labeled data and labeled fake data (generated images from deep generative model) to train a classifier (a task-learner) as well as a discriminator for real / fake images (Tran et al., 2019). Yoo & Kweon 2019 proposed active learning loss method that attaches "loss prediction module" to a task-learner and loss prediction module is trained to predict target losses of unlabeled data samples. In learning loss method, predicted loss was used as a surrogate for model uncertainty and was obtained from feature information in mid-layers.

Data distribution-based approach exploits both labeled and unlabeled data to form sample selection rules so that selected samples are far from the distribution of labeled data and have the most well-representative information of unlabeled pool. For example, clustering for unlabeled data could help to choose samples from clusters so that they are from diverse clusters, not from one or small number of clusters (Nguyen & Smeulders, 2004). Expected error reduction using hierarchical clustering was developed to select active samples in a graph based semi-supervised framework (Mac Aodha et al., 2014). An objective function with diversity constrain was proposed to impose more diversity on the subset of data pool for multi-class active learning (Yang et al., 2015).

Recently, there have also been some works on data distribution-based active learning with deep neural networks. Core set approach was developed that minimizes the distance between labeled data point and unlabeled data pool with intermediate feature information of trained convolutional neural network models (Sener & Savarese, 2018). Recently, Sinha et al. 2019 proposed VAAL to train variational auto encoder (VAE) that captures the representing information of both labeled data and unlabeled data and to adopt adversarial learning to discriminate unlabeled data point from labeled data using the information from the latent space of the trained encoder of the VAE.

Our work falls into the data distribution-based approach, an extension of VAAL to incorporate task related information into the VAE framework. We conjecture that our proposed framework can accommodate both global data distribution structure and local task-related information so that high performance and reliability can be achieved. Both our proposed work and the work of Tran et al. 2019 use deep generative models, but the ways of using them are quite different and our work is using unlabeled data to train the VAE while Tran et al. 2019 does not. Thus, both methods are not competing with each other, rather could be complementing each other.
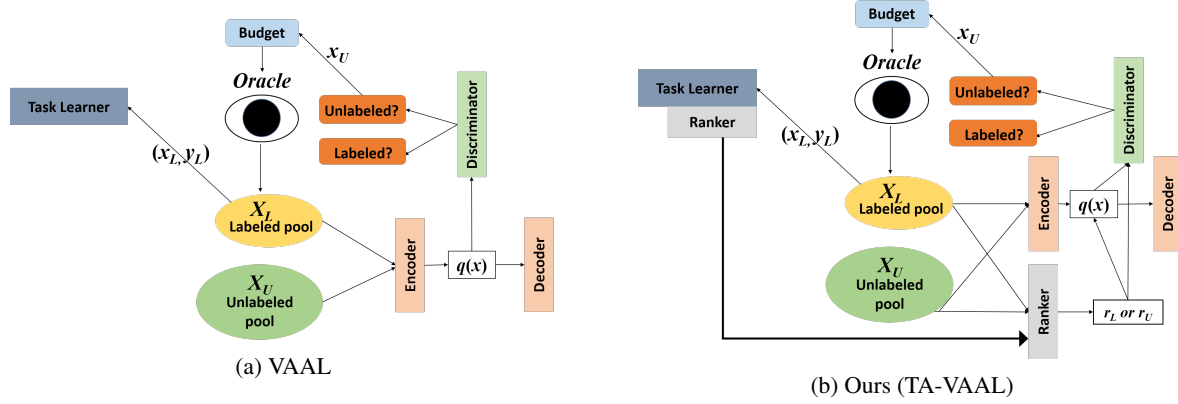
(a) VAAL

(b) Ours (TA-VAAL)

Figure 1. Schematic diagrams of VAAL and our proposed method, TA-VAAL: VAAL is effective at learning the overall influence of labels propagated to the entire (unlabeled) data distribution. Injecting the capability of capturing fine-grained relative label rank information, TA-VAAL helps focus on more influential/informative labels and adjust how they are propagated.

3. Background

3.1. Active learning

Let us denote the pool of labeled data and annotation by (X_L, Y_L), respectively, and the pool of unlabeled data by X_U. The goal of active learning is to select the samples x_U from the unlabeled data pool X_U and to annotate them to add the paired samples / annotations (x_L, y_L), respectively, to the labeled data pool (X_L, Y_L) for the best possible performance of the given task learner (or classifier, deep neural network) T under limited annotation budget. At each k stage, the task leaner T is trained by minimizing the objective function with k-stage labeled data set $(X_L, Y_L)^k$. Given a pair of training data point (x_L, y_L) from the labeled pool, the task learner T predicts $\hat{y}_L = T(x_L)$ and the task learner T is trained by minimizing the loss function $l_L = L_T(\hat{y}_L, y_L)$ for all pairs in (X_L, Y_L).

3.2. Learning loss for active learning

While l_L can be calculated when the ground truth label y_L is available, it can not be without y_L. Yoo & Kweon 2019 proposed loss prediction module to predict the loss value $\hat{l}_U$ for x_U in the unlabeled data pool X_U and used it to perform active learning. Loss prediction module h_{loss} consists of global average pooling, fully connected layer and ReLU to predict l_L, ground-truth target loss.

Loss prediction module is optimized using the objective function $L_{loss}(\hat{l}_L, l_L)$, where $\hat{l}_L = h_{loss}(x_L)$. Mean squared error (MSE) could be naively selected as the loss function $L_{loss}(\cdot)$, but training the loss prediction module h_{loss} with MSE often has a tendency to minimize the scale of entire objective function. To avoid scaling issue, Yoo & Kweon 2019 used ranking between two sample points as another loss function. They compared a pair of samples to obtain ranking: (X_L, Y_L) was re-grouped into a set of pairs

$(x_P, l_P) = \{(x_i, x_j), (l_i, l_j)\}$. If there were B elements in the original (X_L, Y_L), this re-grouped set will have $B/2$ elements (pairs) of (x_P, l_P). Thus, the marginal ranking loss $L_{loss}(\hat{l}_P, l_P)$ in (Yoo & Kweon, 2019) that should be minimized will be

$$-\frac{2}{B} \sum_{i=1}^{B/2} (\max(0, -I_i \cdot (\hat{l}_i - \hat{l}_j) + \epsilon)) \tag{1}$$
$$\text{with} \quad I_i = \begin{cases} +1 & \text{if} \quad l_i > l_j \\ -1 & \text{otherwise} \end{cases}$$

where ϵ is a positive scalar that was set to 1.

Learning loss only utilizes labeled data to train loss prediction module and applies it to unlabeled data to select samples that have fairly high predicted losses.

3.3. Variational adversarial active learning

A schematic diagram of VAAL is illustrated in Figure 1a (Sinha et al., 2019). VAAL firstly trained the VAE with both labeled dataset and unlabeled dataset to learn the representations of both datasets without using any information about the task learner T for performing transductive learning. Then, a discriminator for deciding labeled / unlabeled data is connected to the latent space of the VAE and then both the VAE and the discriminator are jointly trained in an adversarial manner (Goodfellow et al., 2014). Thus, the VAE will encode both the labeled and unlabeled data pools into the same latent space and the discriminator will be trained to predict the probability that a point in the latent space belongs to unlabeled data pool.

VAAL is task-agnostic active learner and the proposed discriminator for labeled / unlabeled samples seems to work well as a surrogate for task learner. However, we conjecture that direct task-related information could be helpful to further improve the overall performance of active learning.

4. Method

4.1. Ranking conditional GANs in VAAL

Generative adversarial network (GAN) consists of generator G and discriminator D (Goodfellow et al., 2014) and has been applied to various tasks such as image style transfer, image super resolution, and image editing. The generator G takes a latent variable $z \sim p_z$ as an input to generate a sample data $G(z)$ and is trained to fool the discriminator D while the discriminator $D(x)$ is trained to distinguish if the input data x is real (or $x \sim p_{data}$) or fake (or $G(z), z \sim p_z$). Thus, the objective function for this minimax game can be expressed as:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}}[\log D(x)]$$
$$+ \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))] \quad (2)$$

where the latent variable z is drawn randomly from a probability distribution such as normal distribution $\mathcal{N}(0,1)$ or uniform distribution $\mathcal{U}(-1,1)$.

To control the distribution of latent space, conditional GAN (CGAN) was proposed to introduce additional latent variable, r, to both the generator G and the discriminator D (Mirza & Osindero, 2014). The objective function of CGAN is as follows:

$$\min_G \max_D \mathbb{E}_{r,x \sim p_{data}}[\log(D(r,x)]$$
$$+ \mathbb{E}_{r,z \sim p_z}[\log(1 - D(r, G(r,z)))]. \quad (3)$$

RankCGAN incorporated "Ranker" into CGAN so that it produces ranking attribute to provide subjective control over the latent subspace containing r (Saquil et al., 2018). By using one or more subjective attributes, generative model can be controlled by ranking information. For example, typical ranking information could be attributes such as "sporty", "black" attributes that are subjective for shoes classification (Saquil et al., 2018).

First of all, we propose to modify the VAE in the original VAAL (Figure 1a) to incorporate a rank variable r from RankCGAN as illustrated in Figure 1b. By inputting the ranking information about predicted loss from the task learner with loss prediction module, RankCGAN framework allows us to control the latent subspace with loss predictions so that the overall latent space can be reshaped. We argue that this structure will allow to reflect the global data structure from both labeled and unlabeled datasets in the overall latent space and to control the loss prediction latent subspace inside the overall latent space.

4.2. Ranker as loss prediction module in task learner

We argue that RankCGAN (Saquil et al., 2018) is a method to connect between VAAL (Sinha et al., 2019) and learning loss (Yoo & Kweon, 2019). We firstly revisited VAAL with

RankCGAN and here we revisit learning loss with RankC-GAN. The concept of "Ranker" in RankCGAN (Saquil et al., 2018) is similar to the loss prediction module in learning loss for active learning (Yoo & Kweon, 2019) in a few aspects and one of them is their ranking losses for training. Rather than directly predicting $loss$ or value itself, they are predicting $ranking$ of losses or values.

Secondly, we propose to modify the original learning loss for active learning (Yoo & Kweon, 2019) with the ranking loss $L_R(\hat{l}_P, l_P)$ from (Saquil et al., 2018) that is as follows:

$$-\frac{2}{B} \sum_{i=1}^{B/2} (I_i \log[\text{sig}(\hat{l}_i - \hat{l}_j)]$$
$$+ (1 - I_i) \log[1 - \text{sig}(\hat{l}_i - \hat{l}_j)]) \quad (4)$$
$$\text{with} \quad I_i = \begin{cases} 1 & \text{if} \quad l_i > l_j \\ 0 & \text{otherwise} \end{cases}$$

where R is the Ranker to predict loss ($e.g.$, $\hat{l}_i = R(x_i)$) and "sig" is the sigmoid function.

Thus, the total loss function of task learner with the Ranker R is expressed as

$$L_T(\hat{y}_L, y_L) + \eta L_R(R(x_P), l_P) \quad (5)$$

where η is a scaling parameter.

We empirically found that training a task learner with the ranking loss from (Saquil et al., 2018) was more stable and yielded better performance than the original learning loss work (Yoo & Kweon, 2019) with the original ranking loss in (1) for the task of active learning.

4.3. Task-aware variational adversarial active learning

Finally, we propose our task-aware VAAL that combines the conditional VAE with a rank variable (predicted loss or true loss) and the task learner with Ranker to yield loss ranking information by feeding Ranker output into a rank variable in the conditional VAE as illustrated in Figure 1b. Our proposed method bridges between model uncertainty-based approach and data distribution-based approach in a tight way by using conditional GAN (RankCGAN) so that the information about data distribution accounts for model uncertainty information (predicted loss). We argue that our proposed approach will have an advantage to use more data (unlabeled data) over typical model uncertainty-based approach trained only with labeled data. We also argue that our approach will have another advantage of exploiting task-related information over data distribution-based approach.

To train the conditional VAE in a conditionally adversarial manner with ranking (the output of Ranker), the objective function L_{VAE}^{trans} of the conditional VAE for learning features

of both labeled and unlabeled pools can be reformulated as

$$\mathbb{E}[\log p_\theta(x_L|z_L, r_L)] - \lambda KL(q_\theta(z_L|x_L)||p(z))+ \\ \mathbb{E}[\log p_\theta(x_U|z_U, r_U)] - \lambda KL(q_\theta(z_U|x_U)||p(z)) \quad (6)$$

where q_θ and p_θ are the encoder and decoder of the VAE that are parameterized by θ, r_L, r_U are the output of Ranker from labeled, unlabeled data points, respectively, $p(z)$ is a Gaussian distribution and λ is a hyper-parameter, $KL(\cdot)$ is Kullback-Leibler distance, and reparameterization technique was used for training. Another loss function for training the VAE is the adversarial loss to fool the binary discriminator D by representing the same probability distribution of labeled and unlabeled pools. The objective function for the conditional adversarial loss is

$$L_{VAE}^{adv} = -\mathbb{E}[\log D(r_L, q_\theta(z_L|x_L)] \\ - \mathbb{E}[\log(D(r_U, q_\theta(z_U|x_U)]. \quad (7)$$

Thus, the final loss function for training the conditional VAE is the sum of (6) and (7).

The loss for training the discriminator D was designed as follows:

$$L_D = -\mathbb{E}[\log D(r_L, q_\theta(z_L|x_L)] \\ - \mathbb{E}[\log(1 - D(r_U, q_\theta(z_U|x_U)]. \quad (8)$$

Finally, at each sampling step, the data points $(x_1^\dagger, ..., x_b^\dagger)$ to be labeled will be selected by performing the following operation:

$$(x_1^\dagger, ..., x_b^\dagger) = \underset{(x_1,...,x_b) \subset X_U}{\operatorname{argmin}} D(r_U, q_\theta(z_U|x_U)). \quad (9)$$

Note that a subset method, replacing X_U in (9) with a random subset of X_U, was used to avoid outliers as much as possible, which was also used in (Yoo & Kweon, 2019).

5. Experimental Results

5.1. Balanced benchmark datasets

We evaluated our proposed TA-VAAL method on various (balanced) benchmark datasets such as CIFAR-10 (Krizhevsky, 2009), SVHN (Netzer et al., 2011) and Fashion MNIST (Xiao et al., 2017). CIFAR-10 consisted of 50,000 training images and 10,000 test images whose size is 32×32 with 10 object categories. SVHN contains 73,257 train images and 26,032 test images with the size of 32×32 and with 10 object categories. Fashion-MNIST consists of the training set of 60,000 images and the test set of 10,000 images, which are 28×28 grayscale images, associated with annotations from 10 classes.

We set initially labeled pool with randomly selected 1,000 images and the query data size b as 1,000 at each stage. To avoid overlapped samples and introduce diversity to selected

samples from unlabeled pool, we adopt the subset method that obtains random subsets from unlabeled pool at each stage before applying active learning methods. We set the subset size to be 10,000, 10 times larger than the query size.

5.2. Imbalanced dataset

We also performed experiments on imbalanced datasets whose data set sizes are different for classes. We choose the CIFAR-10 dataset that has 5,000 sample images per class for 10 classes (thus, the total number of images are 50,000). Images were randomly removed in an imbalanced manner as follows. We firstly removed images randomly for the first 4 classes to yield 3,000, 3,000, 4,000, 5,000 images for the classes 1, 2, 3, 4, respectively. Then, more images were removed randomly so that there are three cases of different levels of imbalances as the following Table 1 illustrates:

Table 1. The numbers of images per class for all 10 classes in order on our generated imbalanced CIFAR-10 datasets. The entropies for the first, second, third rows are 1.65, 1.89, 2.11, respectively.

NUMBERS OF IMAGES PER CLASS ($\times 100$)	TOTAL
30, 30, 40, 50, 0.5, 1, 1, 1, 2, 10	16,550
30, 30, 40, 50, 5, 5, 2, 3, 5, 20	19,000
30, 30, 40, 50, 10, 10, 10, 10, 10, 30	23,000

The smaller the entropy is, the more imbalanced the dataset is. The amount of images for the first 4 classes is 15,000 images and the remaining 6 classes contain 1,550, 4,000, 8,000 images. The entropies of three datasets are 1.65,1.89 and 2.11, respectively, corresponding to the first, second, third rows of Table 1.

5.3. Implementation details

For training, various techniques were used such as 32×32 random crop from 36×36 zero-padded images, normalization with mean and standard deviation of training set, and horizontal flip and flop augmentation for CIFAR-10 and SVHN. Normalization only was applied for Fashion MNIST. ResNet18 (He et al., 2016) was used for all task learners and stochastic gradient descent (SGD) optimizer was used with momentum of 0.9 and weight decay of 0.005. Learning rate was 0.1 for the first 160 epochs and then 0.01 for the rest of 40 epochs. For the VAE, a modified Wasserstein auto-encoder (Tolstikhin et al., 2018) for taking ranking information was used and the discriminator was constructed as a 5-layer multi-layer perceptron (MLP). For both the VAE and the discriminator, the Adam optimizer (Kingma & Ba, 2015) with learning rate of 5×10^{-4} was used. Mini-batch size was 128 and the epochs was 200 for all datasets.

5.4. Results for balanced benchmark datasets

Four active learning methods were evaluated including random sampling, learning loss (Yoo & Kweon, 2019), VAAL (Sinha et al., 2019) and our proposed TA-VAAL method on three benchmark datasets, CIFAR-10, SVHN and Fashion MNIST. Figure 2 presents three graphs of the number of labeled images versus accuracy, the mean and standard deviation of 5 trials.

In Figure 2a on CIFAR-10, our proposed method achieved the mean accuracy of 90.32% in the last stage (10k) while other methods yielded less than the mean accuracy of 90%. Our proposed methods outperforms against state-of-the-art methods for all stages. In particular, our proposed method yielded the mean accuracy of over 80% only with 4,000 labeled data while other methods were not able to achieve.

The results on SVHN are shown in Figure 2b. Our proposed method outperforms all state-of-the-art methods in almost all stages except for one stage (5k). Our proposed method yielded substantially higher mean accuracies than other methods at early stages such as 3k and 4k.

The results on Fashion MNIST are depicted in Figure 2c. Our proposed method outperforms all other compared methods over almost all stages except for the first stage and outperformed all other methods significantly at the 3k stage by more than 5% margin.

Note that the performances of learning loss and ours yielded slightly higher or lower mean accuracies at the first stage due to additional loss prediction module attached to the task learner. We performed ablation studies and will show the results in the later section to show that this additional loss is not the most important factor for the overall performances of our proposed method.

5.5. Results on imbalanced dataset

Four active learning methods were evaluated including random sampling, learning loss (Yoo & Kweon, 2019), VAAL (Sinha et al., 2019) and our proposed TA-VAAL method on three imbalanced datasets by modifying CIFAR-10 as illustrated in Table 1. Figure 3 presents three graphs of the number of labeled images versus accuracy, the mean and standard deviation of 5 trials. Due to reduced data size, we experimented until the labeled data pool size of 5,000. Initial data size of labeled pool and incremental budget size were not changed.
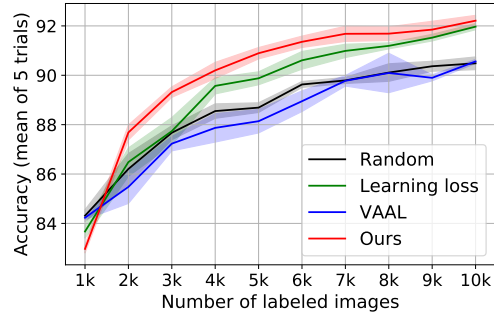
Figure 3 on modified CIFAR-10 with the dataset entropies of 1.65 (the most imbalanced), 1.89 and 2.11 (the least imbalanced) illustrates the results to show that our proposed method outperformed all other state-of-the-art methods over all stages. For more imbalanced dataset, our proposed method was able to achieve higher performances than other
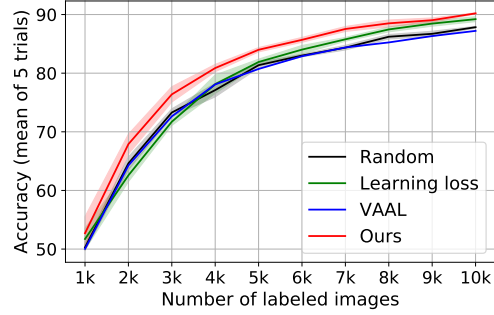

(a) CIFAR-10
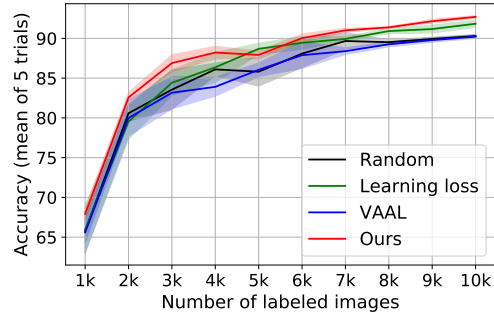
(b) SVHN

(c) Fashion MNIST

Figure 2. Accuracy of active learning algorithms (random sampling, Learning loss, VAAL and ours (TA-VAAL)) on balanced datasets. The vertical width of the shaded region along each curve corresponds to twice the standard deviations.

compared methods and these differences became smaller for less imbalanced dataset with higher dataset entropy. Even though the final dataset size for all cases were 5,000, note that there are some classes with 50-100 total images per class in the case of the dataset entropy of 1.65. Thus, in this most harsh case, the standard deviation of the performance was large and the final accuracy at 5k stage was still lower than 70% due to these bottlenecks. In the case with dataset entropy of 1.89, our proposed method outperformed all other methods with substantial margins. Note that model uncertainty based approaches such as ours or learning loss were
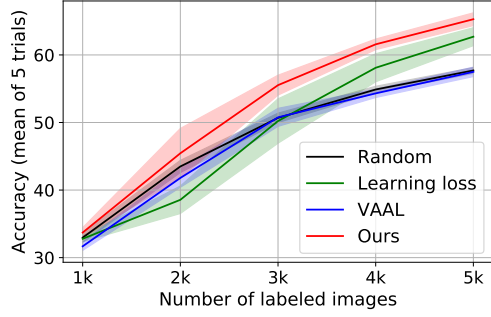
(a) Dataset entropy = 1.65
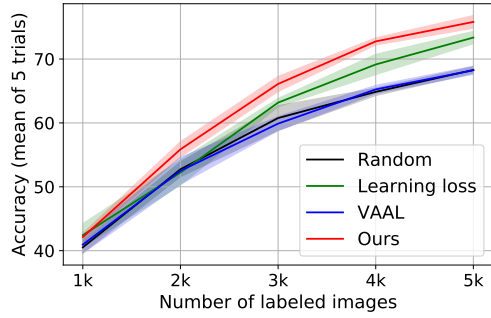
(b) Dataset entropy = 1.89
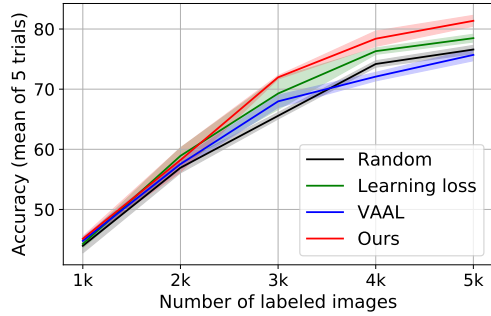
(c) Dataset entropy = 2.11

Figure 3. Active learning results on imbalanced benchmark datasets: random sampling, Learning loss, VAAL and ours (TA-VAAL). Higher dataset entropy implies more balanced dataset.

able to achieve substantially higher performances than other task-agnostic approaches in the case of imbalanced data sizes for classes. Moreover, our proposed method yielded substantially higher performances than learning loss method due to the incorporation of the unlabeled data information using generative models.

6. Empirical Analysis Results

6.1. Task learners without loss prediction modules

In (Yoo & Kweon, 2019), the task learner with loss prediction module was trained to show the performance on actively
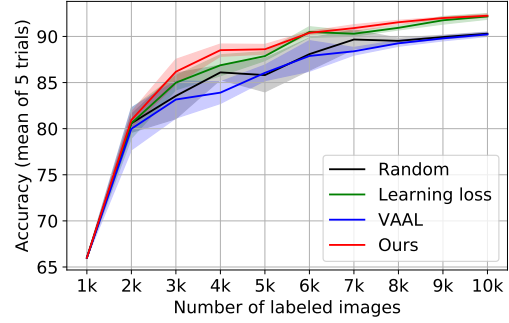

Figure 4. Active learning results on SVHN using task learners without loss prediction modules.

selected samples and was compared with other methods using the task learners without loss prediction modules. Since loss prediction modules are part of learning loss method and our proposed TA-VAAL method and it seems that these loss prediction modules influence the overall performance as shown in Figure 2, often yielding slightly higher or lower performances.

To measure the quality of the selected samples using active learning methods, we trained task learners of learning loss method and our proposed method again without attaching loss prediction module (or Ranker) on the selected datasets up to 10k as shown in Figure 4. Our proposed method was not able to yield the best performance at 5k stage on SVHN as shown in Figure 2b, but now our proposed method outperforms all other methods on SVHN with the effect of loss prediction module for task learner training is minimized as illustrated in Figure 4. However, at the last stage (10k), our proposed method yielded slightly dropped performance by 0.48% compared with the result of task learner with Ranker. Learning loss method has slightly improved performance by 0.24%. Despite the changes, our proposed method still outperformed other state-of-the-art methods by achieving mean accuracy of 92.27 % in last active learning cycle.

6.2. Ablation studies

Figure 5 shows the performance results of our proposed methods with and without proposed components and proposed structures along with other state-of-the-art methods. The means and standard deviations of 5 trials are reported. First of all, learning loss method without proposed ranking loss (4), called learning loss_v2, yielded substantially higher performances at later active learning stages and comparable performances at early stages as compared to the original learning loss. Thus, it seems that using our proposed ranking loss (4) along with task learners is advantageous.

Another study is to incorporate ranking information into VAAL by using the original learning loss architecture,
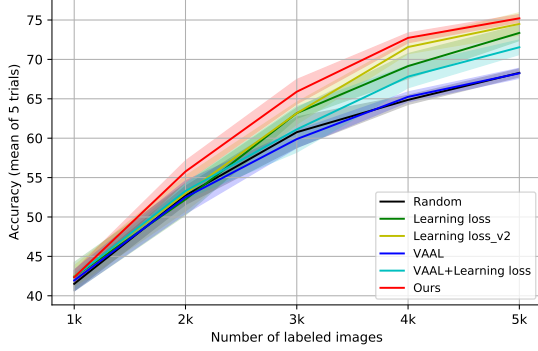
Figure 5. Results of our proposed methods with and without proposed components and proposed structures on imbalanced dataset, CIFAR-10, whose dataset entropy is 1.89.

rather than our proposed Ranker (4). This combination of VAAL+learning loss still yielded substantially better performances than VAAL over all stages, but was not able to yield better performance than the original learning loss method at later stages while this combination seems to yield good performances at early stages.

### 6.3. On selected samples using active learning methods

Figure 6 shows the bar graphs for the number of labeled images (selected sampled) versus the entropies of the data class counts of selected samples over 10 classes. The higher the entropy is, the better samples are selected over classes. Figure 6a shows that our proposed method selected good samples with high data class count entropy on the most severely imbalanced dataset and Figure 6b also shows similar results on the least severely imbalanced dataset with robustness or small standard deviations.

These results are also corresponding to the performance results in Figure 3. For example, learning loss method yielded substantially low performance at 2k stage as shown in Figure 3a due to its data selection with very low data count output entropy over classes at the same stage as illustrated in Figure 6a. This is possibly due to limited number of data for certain classes in the case of dataset entropy = 1.65 so that task learner in learning loss method was not well-trained. In the meanwhile, our proposed method was able to select good samples at the same stage thanks to the structure from VAAL so that good performance as well as high data class count entropy was able to be achieved.

Lastly, Figure 7 shows that at the last stage, the discriminator of VAAL yielded concentrated output values (counts) at the last stage so that the active learning selection became almost random, while our proposed method still yielded wide range of discriminator outputs for good selections and performances, as illustrated in Figures 6 and 3.
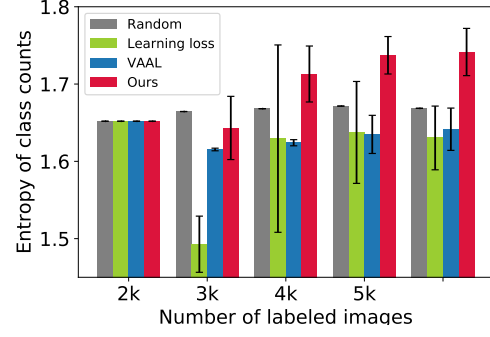
(a) Dataset entropy = 1.65
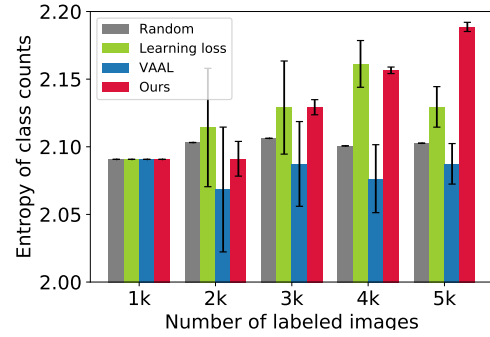
(b) Dataset entropy = 2.11

Figure 6. Bar graphs of number of labeled images (selected samples) versus data class count entropy on the imbalanced datasets with (a) dataset entropy of 1.65 and (b) dataset entropy of 2.11.
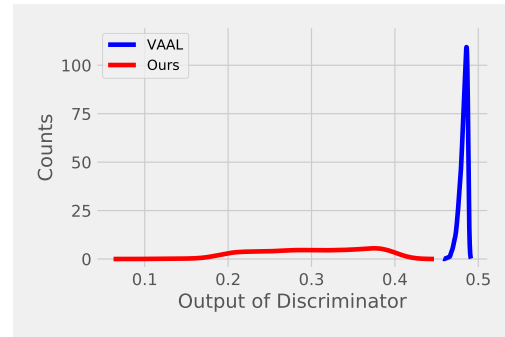

Figure 7. Output of discriminator versus number of samples at the last stage on the imbalanced dataset with entropy 1.65.

7. Conclusion

We proposed the methods that simultaneously take advantage of both promising directions for active learning that have been recently explored are data distribution-based approach to select samples that are far from current labeled pool and model uncertainty-based approach that relies on task models. We demonstrate that our proposed methods outperform state-of-the-art active learning methods on various balanced and imbalanced benchmark datasets.

Acknowledgments

References

Brinker, K. Incorporating diversity in active learning with support vector machines. In *ICML*, pp. 59–66, 2003.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309, 2019.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pp. 801–818, 2018.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In *ICML*, pp. 1183–1192, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *NIPS*, pp. 3581–3589, 2014.

Kirsch, A., van Amersfoort, J., and Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *NeurIPS*, pp. 7024–7035, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105, 2012.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *CVPR*, pp. 3431–3440, 2015.

Mac Aodha, O., Campbell, N. D., Kautz, J., and Brostow, G. J. Hierarchical subquery evaluation for active learning on a graph. In *CVPR*, pp. 564–571, 2014.

Melville, P. and Mooney, R. J. Diverse ensembles for active learning. In *ICML*, pp. 74, 2004.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, pp. 1–9, 2011.

Nguyen, H. T. and Smeulders, A. Active learning using pre-clustering. In *ICML*, pp. 79, 2004.

Pinsler, R., Gordon, J., Nalisnick, E., and Hernández-Lobato, J. M. Bayesian batch active learning as sparse subset approximation. In *NeurIPS*, pp. 6356–6367, 2019.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. In *ICLR*, 2017.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *CVPR*, pp. 779–788, 2016.

Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pp. 91–99, 2015.

Saquil, Y., Kim, K. I., and Hall, P. Ranking CGANs: subjective control over semantic image attributes. In *BMVC*, 2018.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

Settles, B. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Sinha, S., Ebrahimi, S., and Darrell, T. Variational adversarial active learning. In *ICCV*, pp. 5972–5981, 2019.

Sundin, I., Schulam, P., Siivola, E., Vehtari, A., Saria, S., and Kaski, S. Active learning for decision-making from imbalanced observational data. In *ICML*, pp. 6046–6055, 2019.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. In *ICLR*, 2018.

Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Tran, T., Do, T.-T., Reid, I., and Carneiro, G. Bayesian generative active deep learning. In *ICML*, pp. 6295–6304, 2019.

Wang, Z. and Ye, J. Querying discriminative and representative samples for batch mode active learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9 (3):1–23, 2015.

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

Yang, L., Zhang, Y., Chen, J., Zhang, S., and Chen, D. Z. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *MICCAI*, pp. 399–407. Springer, 2017.

Yang, Y., Ma, Z., Nie, F., Chang, X., and Hauptmann, A. G. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.

Yoo, D. and Kweon, I. S. Learning loss for active learning. In *CVPR*, pp. 93–102, 2019.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pp. 912–919, 2003.