DISTRIBUTION AWARE ACTIVE LEARNING VIA GAUSSIAN MIXTURES

[1]**Younghyun Park,** [2]**Dong-Jun Han,** [3]**Jungwuk Park,** [4]**Wonjeong Choi,** [5]**Humaira Kousar,** [6]**Jaekyun Moon**
[1,3,4,5,6] Korea Advanced Institute of Science and Technology (KAIST), [2] Purdue University
{[1]dnffkf369,[3]savertm,[4]dnjswjd5457,[5]humairakousar32}@kaist.ac.kr, [2]han762@purdue.edu, [6]jmoon@kaist.edu

ABSTRACT

In this paper, we propose a distribution-aware active learning strategy that captures and mitigates the distribution discrepancy between the labeled and unlabeled sets to cope with overfitting. By taking advantage of gaussian mixture models (GMM) and Wasserstein distance, we first design a distribution-aware training strategy to improve the model performance. Then, we introduce a hybrid informativeness metric for active learning which considers both likelihood-based and model-based information simultaneously. Experimental results on four different datasets show the effectiveness of our method against existing active learning baselines.

1 INTRODUCTION

Active learning (AL), the method of actively selecting informative unlabeled samples to be labeled, is one of the important and challenging problems in machine learning due to the scarcity of labeled data in the learning process. The key idea of active learning is that some training samples are more informative than others so deep neural network (DNN) models can achieve higher performance by using only informative samples. Therefore, the common goal of AL algorithms is to design a good informativeness criterion to label the most informative samples in the unlabeled pool.

Despite extensive progress in active learning, we point out that existing works do not directly handle the distribution discrepancy between labeled set X_L and unlabeled set X_{UL} in the latent space which causes overfitting, when (i) training the model or (ii) evaluating the informativeness of unlabeled samples. In Fig. 1, it is unveiled that although DNN seems to be well trained since features are meaningfully separated in X_L, there exists a distinct discrepancy between X_L and X_{UL}. In other words, given a feature extractor f_θ, $p(f_\theta(X_L)) \neq p(f_\theta(X_{UL}))$. It becomes problematic when (i) we make evaluation on test set or (ii) we estimate the informativeness of unlabeled samples since the model is already overfitted to the distorted and limited embedding region. Unfortunately, existing AL works tend not to take the distribution discrepancy into account carefully. Most of them (Ash et al., 2019; Parvaneh et al., 2022; Agarwal et al., 2020; Liu et al., 2021) train DNN models by using only labeled set and do not consider the overfitting to the labeled set. As for informativeness estimation, some works (Sener & Savarese, 2017; Parvaneh et al., 2022) consider features of unlabeled samples but disregard the distribution discrepancy; they rely on Euclidean distance (e.g., L2 norm) which cannot reflect distorted latent space compared to probabilistic distance.

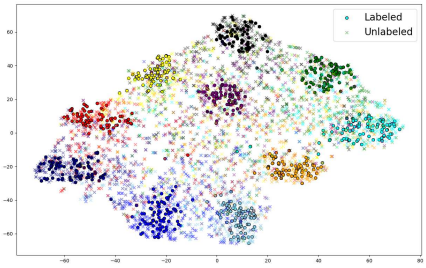


Figure 1: T-SNE visualization of embedding space of labeled and unlabeled samples.

Goal and contribution. The general goal of this paper is to mitigate the overfitting caused by distribution discrepancy between X_L and X_{UL}, which is pathologically prominent in active learning. To this end, we suggest (i) novel training strategies and (ii) new informativeness metrics; these methods should be able to capture and resolve the discrepancy. Our key idea is to adopt a well-known unsupervised clustering method, gaussian mixture models (GMM), which is an useful tool to probabilistically interpret the latent distribution. We propose to characterize each dataset as a probabilistic model by fitting a GMM and measure likelihoods of unlabeled samples x_{ul} from the perspective of X_L and X_{UL}. Along with GMM, our overall contributions are summarized as follows:

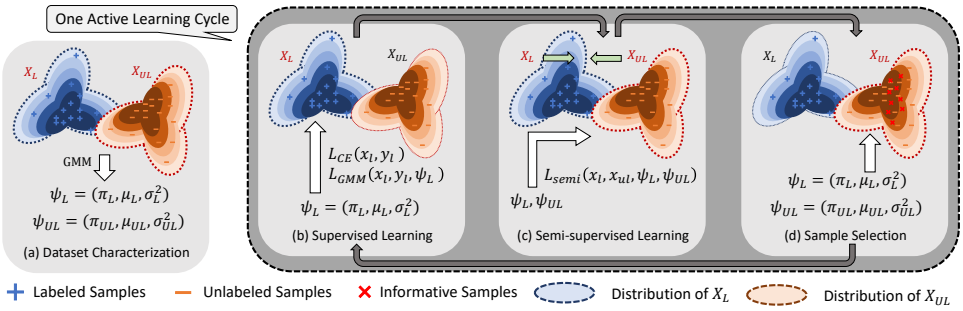


Figure 2: Proposed active learning strategy using gaussian mixtures.

- We propose training strategies that consider the distribution discrepancy using GMM. We introduce a semi-supervised learning stage to reduce the Wasserstein distance between X_L and X_{UL}.
- We propose a hybrid informativeness metric that takes both likelihood-based and model-based information simultaneously. This metric prioritizes unlabeled samples which are expected to resolve the distribution discrepancy the most by means of distributional knowledge obtained from GMM.

2 Proposed Method

In AL, DNN is trained through successive active learning cycles $t = 0, 1, .., T$. During t^{th} cycle, DNN is trained using a small labeled set X_L^t and a large unlabeled set X_{UL}^t. At the end of the t^{th} cycle, the DNN selects a small set of unlabeled samples I^t with the highest informativeness. Then, human experts annotate I^t to update the labeled set as $X_L^{t+1} = X_L^t \cup I^t$ and the unlabeled set as $X_{UL}^{t+1} = X_{UL}^t \setminus I^t$. This process is repeated until the labeling budget is depleted. In the following, we describe our distribution-aware training strategy (in Sections 2.1 & 2.2 & 2.3) and our distribution-aware informativeness scores (in Section 2.4) for the distribution-aware active learning.

2.1 Dataset Characterization via GMM

We assume feature representations of training data follow an isotropic Gaussian mixture so that likelihood of each sample can be statistically estimated according to the learned mixture models. Given a batch of samples $\{x_i\}_{i=1}^N$ and its feature representations $\{z_i = f_\theta(x_i)\}_{i=1}^N$, GMM's parameter set $\psi := (\pi, \mu, \sigma^2)$ can be optimized by Expectation Maximization algorithm as:

(**E Step.**) Compute the responsibility γ from the current parameter set $\psi := (\pi, \mu, \sigma^2)$

$$\gamma(z_{ik}) := p(y_i = k|z_i) = \frac{\pi_k \mathcal{N}(z_i|\mu_k, \sigma_k^2)}{\sum_{j=1}^K \pi_j \mathcal{N}(z_i|\mu_j, \sigma_j^2)} \tag{1}$$

(**M Step.**) Update the parameter set $\psi := (\pi, \mu, \sigma^2)$ based on the current responsibilities

$$\pi_k = \frac{\sum_{i=1}^N \gamma(z_{ik})}{\sum_{k=1}^K \sum_{i=1}^N \gamma(z_{ik})}, \quad \mu_k = \frac{\sum_{i=1}^N \gamma(z_{ik}) z_i}{\sum_{i=1}^N \gamma(z_{ik})}, \quad \sigma_k^2 = \frac{\sum_{i=1}^N \gamma(z_{ik})(z_i - \mu_k)^2}{\sum_{i=1}^N \gamma(z_{ik})}$$

where π_k, μ_k, σ_k denote mixing coefficient, mean, diagonal covariance of k^{th} modality and K is the number of categories. γ_{ik} stands for 'responsibility' of k^{th} Gaussian $\mathcal{N}(z|\mu_k, \sigma_k^2)$ for generating the data z_i. To characterize X_L, X_{UL}, we divide datasets into multiple batches and fit a GMM to each batch. Afterwards, we average the learned ψ from all batches and regard the average as a probabilistic characteristic of the datasets (i.e., ψ_L and ψ_{UL} characterize X_L and X_{UL}, respectively).

2.2 Distribution Aware Regularization

Most of the existing AL works supervise the output of softmax classifier by minimizing cross-entropy loss ensuring that outputs of the classifier are well separated according to the target class. On top of the cross-entropy loss, we propose to minimize an auxiliary regularization loss L_{GMM} to supervise latent space as well. L_{GMM} is defined as NLL loss: $L_{GMM}(z, y; \psi_L) = \sum_{k=1}^K -y_k log(p(y = k|z; \psi_L))$. Thus, the training objective for supervised learning stage now becomes $L_{CE}(y, \hat{y}) + \alpha L_{GMM}(z, y; \psi_L)$ where α is a constant weight. Benefits of this regularization are two-fold. First, by supervising the latent space in itself, we can guide gaussian modalities to be kept apart. Secondly, supervising the latent space during the supervised training stage improves the interpretability of the feature extractor to understand uneven distribution of feature representations.

2.3 DISTRIBUTION ALIGNMENT VIA WASSERSTEIN DISTANCE

Distribution alignment via adversarial learning. We introduce a semi-supervised adversarial learning stage after the supervised learning stage in Section 2.2. Advantages of this method are two-fold. 1) By aligning two distributions $p(z_L)$ and $p(z_{UL})$, DNN is expected to learn a general latent space which reduces the distribution discrepancy. 2) At the sample selection stage, the learned discriminator can measure the model-based informativeness. For distribution alignment, we minimize Wasserstein distance $W(z_L, z_{UL}) = \inf_{\delta \in \Pi(z_L, z_{UL})} \mathbb{E}_{(z_l, z_{ul}) \sim \delta}[||z_l - z_{ul}||]$, which can be modified into following adversarial training objective based on Kantorovich-Rubinstein Duality Theorem as:

$$\min_\theta \max_\phi \mathbb{E}_{z_{ul} \sim f_\theta(X_{UL})}[D_\phi(z_{ul})] - \mathbb{E}_{z_l \sim f_\theta(X_L)}[D_\phi(z_l)] \tag{2}$$

where f_θ is the feature extractor and D_ϕ is the 1-Lipschitz discriminator. The above training objective guides D_ϕ to output $D_\phi(z_l) \to 0, D_\phi(z_{ul}) \to 1$, while f_θ is trained to confound D_ϕ. Also, we minimize the CE loss using labeled data so that adversarial learning process in Eq. 2 does not harm the performance of the main task excessively. Note that we resolve the 1-Lipshitz continuous constraints on D_ϕ by adding the gradient penalty term of (Gulrajani et al., 2017) in Eq. 2.

Distribution aware distribution alignment via GMM. We propose to feed D_ϕ with the distributional information from (ψ_L, ψ_{UL}) on top of z. This helps D_ϕ consider global data distribution rather than memorizing each of a small number of labeled samples. Considering ψ is a group of high-dimensional vectors (e.g., $\mu_L \in R^{100 \times 512}$ for CIFAR100), it's hard to simply concatenate (ψ_L, ψ_{UL}) to z, so we instead propose to indirectly exploit (ψ_L, ψ_{UL}) for the reduction of computational burdens. Specifically, we first compute posterior probability $p(y = k|z; \psi_L)$, $p(y = k|z; \psi_{UL})$ and likelihoods $p(z; \psi_L)$, $p(z; \psi_{UL})$, and then concatenate z to the above probabilities and pass the concatenated vectors to the discriminator. As for the computation of (ψ_L, ψ_{UL}), we found that it is stable enough to sporadically characterize X_L, X_{UL} (e.g., every 500 out of 10,000 iterations).

2.4 HYBRID METHOD FOR INFORMATIVE IMAGE SELECTION

Likelihood-based informativeness metric. We prioritize samples that can mitigate the distribution discrepancy between X_L and X_{UL}. In other words, these samples should be dissimilar to X_L, yet best represent X_{UL} at the same time. Accordingly, we define the likelihood-based metric as follows:

$$I_{Like}(x_{ul}; \psi_L, \psi_{UL}) = p(z_{ul}; \psi_{UL}) - p(z_{ul}; \psi_L) \quad where \quad p(z; \psi) = \sum_{j=1}^{K} \pi_j \mathcal{N}(z|\mu_j, \sigma_j^2) \tag{3}$$

Here, log-likelihood $p(z; \psi)$ reflects the probability that data z comes from a Gaussian mixture ψ.

Model-based informativeness metric. We make use of the output of the discriminator D_ϕ learned during distribution alignment as: $I_{model} = D_\phi(z)$. Note that I_{model} prioritize samples that have not been seen during the training process. Finally, our proposed acquisition function is:

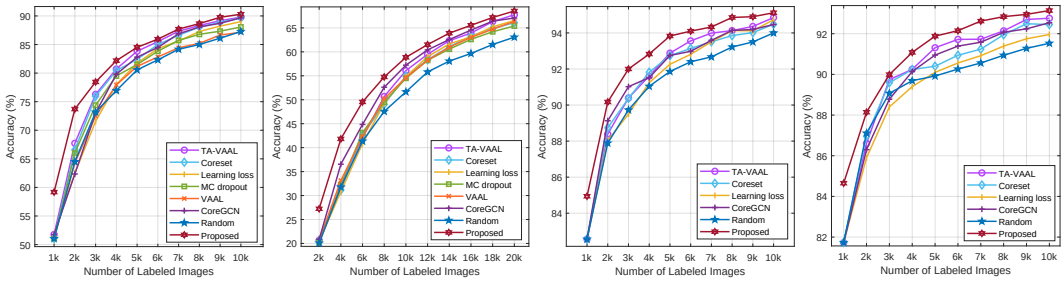
$$I_{total}(x_{ul}; \psi_L, \psi_{UL}) = I_{Like}(x_{ul}; \psi_L, \psi_{UL}) + \beta I_{model}(x_{ul}; \psi_L, \psi_{UL}) \tag{4}$$

where β is the mixing coefficient. Now we can select informative samples at the end of every cycle by selecting a candidate set S whose samples have the largest I_{total}. To keep S balanced, we select the same number of samples per each category based on pseudo-labels. Lastly, following (Parvaneh et al., 2022; Ash et al., 2019), we apply K-means clustering to S for the diversity of selected samples.

3 EXPERIMENTS

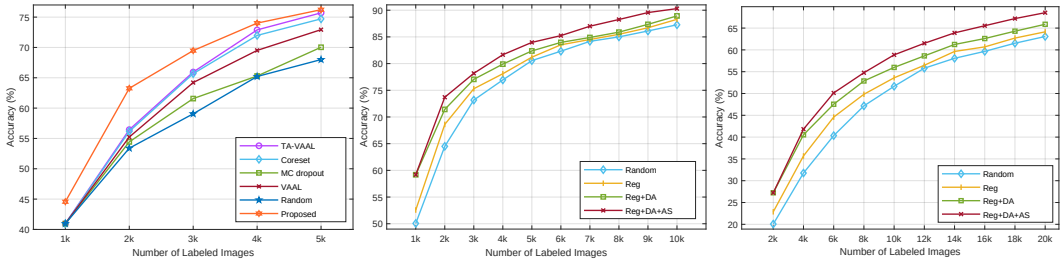
Dataset. Our proposed work is evaluated on four popular benchmark datasets: SVHN (Netzer et al., 2011), Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky et al., 2009) and CIFAR-100 (Krizhevsky et al., 2009). For all datasets except CIFAR-100, 1,000 samples are initially labeled and additionally 1,000 samples are labeled at the end of every cycle until the size of labeled set reaches 10,000. As for CIFAR-100, size of the X_L increases from 2,000 to 20,000 in steps of 2,000.

Implementation details. Following the experimental setup of (Kim et al., 2021; Caramalau et al., 2021), we implement the main classifier using ResNet-18 (He et al., 2016) which is combined with a single linear layer softmax classifier. The classifier is optimized via SGD optimizer with learning rate of 0.1; momentum of 0.9; batch size of 100; epoch number of 200. The discriminator is composed of three linear layers with a sigmoid activation and optimized by Adam optimizer for 10,000 iterations with a learning rate of 5e-4 which is decayed to 5e-5 for the last 2,000 iterations.



(a) CIFAR-10 (b) CIFAR-100 (c) SVHN (d) Fashion-MNIST

Figure 3: Performance comparison on balanced datasets with state-of-the-art methods.



(a) imbalanced CIFAR-10 (b) CIFAR-10 (c) CIFAR-100

Figure 4: Performance comparison on imbalanced datasets and ablation studies on proposed methods.

Baselines. We compare our work with the following schemes: Random sampling, TA-VAAL (Kim et al., 2021), CoreSet (Sener & Savarese, 2017), LL4AL (Yoo & Kweon, 2019), MC-Dropout (Gal et al., 2017), VAAL (Sinha et al., 2019) and CoreGCN (Caramalau et al., 2021). For a fair comparison, we train five independent networks with different seeds and report the mean performance.

Results on balanced dataset. In Fig. 3, we compare the performance of our method with various baselines on balanced datasets. It can be seen that the proposed algorithm performs the best on all datasets. Also, it is noteworthy that the proposed method shows a particular excellence in the early stages wherein other methods show lower performance because of overfitting to the small-sized labeled set. The result shows that the proposed methods such as distribution-aware distribution alignment and distribution-based acquisition function helps the DNN to prevent overfitting.

Results on imbalanced dataset. Fig. 4a compares the performance on imbalanced CIFAR-10 whose imbalance ratio is set to 10 (i.e., 5 classes have 10 times more samples than the remaining 5 classes). As shown, the proposed method exhibits a margin over baseline methods. We suppose that I_{Like} in (Eq. 3) reflects the imbalance in the labeled set. This is because π in GMM parameters ψ reflects the number of per-class samples so that the likelihood-based metric (Eq. 3) guides to select unlabeled samples which resolve the category imbalance between the labeled and unlabeled sets.

Ablation studies. Fig. 4b and Fig. 4c show ablation studies on our proposed methods. Note that, for brevity, we denote the distribution-aware regularization in Section 2.2 as *Reg*; distribution alignment in Section 2.3 as *DA*; active sample selection in Section 2.4 as *AS*. In every cycle, it can be seen that proposed methods consistently improve the performance. Specifically, *Reg* and *DA* contribute more to performance gain in early cycles while AS serves prominent gain at late cycles. It is confirmed that *Reg* and *DA* resolve distribution discrepancy better when the size of a labeled set is small and the model is vulnerable to overfitting. On the other hand, *AS* steadily improves performance even in late cycles by utilizing both likelihood-based metrics and model-based metrics.

4 Conclusion

In this work, we propose a unified framework for active learning which makes use of Gaussian mixture models. We fit GMM to labeled set and unlabeled set to characterize distributional information and assist the active learning in versatile ways including distribution-aware regularization, distribution alignment, likelihood-based informativeness metric. We validate the superiority of proposed methods through extensive comparisons with baselines and ablation studies. In the future, we plan to extend our work on various computer vision tasks (e.g., object detection (Haussmann et al., 2020)) or various settings such as open set (Park et al., 2022), model evaluation (Kossen et al., 2022)

4.1 ACKNOWLEDGEMENT

This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD). Also, this work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) funds from MSIT of Korea (No. 2020-0-00626).

REFERENCES

Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pp. 137–153. Springer, 2020.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9583–9592, 2021.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M Alvarez. Scalable active learning for object detection. In *2020 IEEE intelligent vehicles symposium (iv)*, pp. 1430–1435. IEEE, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8166–8175, 2021.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. *arXiv preprint arXiv:2202.06881*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9274–9283, 2021.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Dongmin Park, Yooju Shin, Jihwan Bang, Youngjun Lee, Hwanjun Song, and Jae-Gil Lee. Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning. *arXiv preprint arXiv:2210.07805*, 2022.

Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12237–12246, 2022.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

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

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 93–102, 2019.