Variational Adversarial Active Learning - Extended Abstract

Syna Ebrahimi *
UC Berkeley

Samarth Sinha *
University of Toronto

Trevor Darrell
UC Berkeley

Abstract

Active learning aims to develop label-efficient algorithms by sampling the most representative queries to be labeled by an oracle. We describe a pool-based semi-supervised active learning algorithm that implicitly learns this sampling mechanism in an adversarial manner. Unlike conventional active learning algorithms, our approach is task agnostic, i.e., it does not depend on the performance of the task for which we are trying to acquire labeled data. Our method learns a latent space using a variational autoencoder (VAE) and an adversarial network trained to discriminate between unlabeled and labeled data. The minimax game between the VAE and the adversarial network is played such that while the VAE tries to trick the adversarial network into predicting that all data points are from the labeled pool, the adversarial network learns how to discriminate between dissimilarities in the latent space. We evaluate our method on various image classification and semantic segmentation benchmark datasets and establish a new state of the art on ImageNet and BDD100K. Our code and complete version of this work [9] are available at https://github.com/sinhasam/vaal.

1. Introduction

The recent success of learning-based computer vision methods relies heavily on abundant annotated training examples, which may be prohibitively costly to label or impossible to obtain at large scale in online and continual learning [5, 9, 4]. In order to mitigate this drawback, active learning [2] algorithms aim to incrementally select samples for annotation that result in high classification performance with low labeling cost. This paper introduces a pool-based active learning strategy which learns a low dimensional latent space from labeled and unlabeled data using Variational Autoencoders (VAEs). VAEs have been well-studied and valued for both their generative properties as well as their ability to learn rich latent spaces. Our method, Variational Adversarial Active Learning (VAAL), selects instances for labeling from the unlabeled pool that are *sufficiently* different in the latent space learned by the VAE to maximize the performance of the representation learned on the newly labeled data. Sample selection in our method is performed by an adversarial network which classifies which pool the instances belong to (labeled or unlabeled) and does not depend on the task or tasks for which are trying to collect labels.

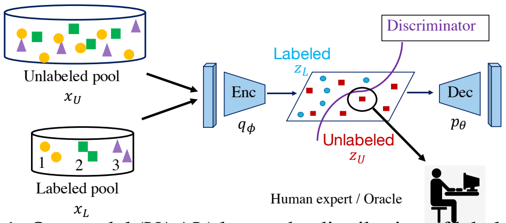

Figure 1: Our model (VAAL) learns the distribution of labeled data in a latent space using a VAE optimized using both reconstruction and adversarial losses. A binary adversarial classifier (discriminator) predicts unlabeled examples and sends them to an oracle for annotations. The VAE is trained to fool the adversarial network to believe that all the examples are from the labeled data while the discriminator is trained to differentiate labeled from unlabeled samples. Sample selection is entirely separate from the main-stream task for which we are labeling data inputs, making our method to be *task-agnostic*

2. Method

Let (x_L, y_L) be a sample pair belonging to the pool of labeled data (X_L, Y_L). X_U denotes a much larger pool of unlabeled samples (x_U). The goal of the active learner is to train the most label-efficient model by iteratively querying a fixed sampling *budget*, b number of the most informative samples from the unlabeled pool $(x_U \sim X_U)$, using a sampling strategy to be annotated by the oracle such that the expected loss is minimized. We use a VAE for representation learning in which the encoder learns a low dimensional space using a Gaussian prior and the decoder reconstructs the input data to minimize the following objective function.

$$\begin{aligned}\mathcal{L}_{\text{VAE}}^{trd} = \quad & \mathbb{E}[\log p_\theta(x_L|z_L)] - \mathrm{D}_{\text{KL}}(q_\phi(z_L|x_L)||p(z)) \\ & + \mathbb{E}[\log p_\theta(x_U|z_U)] - \mathrm{D}_{\text{KL}}(q_\phi(z_U|x_U)||p(z)) \end{aligned} \tag{1}$$

[*] Equal contribution.

where q_ϕ and p_θ are the encoder and decoder parametrized by ϕ and θ, respectively. $p(z)$ is the prior chosen as a unit Gaussian. We also train an adversarial network to map the latent representation of $z_L \cup z_U$ to a binary label which is 1 if the sample belongs to X_L and is 0, otherwise. The key to our approach is that the VAE and the discriminator are learned together in an adversarial fashion. While the VAE maps the labeled and unlabeled data into the same latent space with similar probability distribution $q_\phi(z_L|x_L)$ and $q_\phi(z_U|x_U)$, it fools the discriminator to classify all the inputs as labeled. On the other hand, the discriminator attempts to effectively estimate the probability that the data comes from the unlabeled data. The objective function for the adversarial role of the VAE and the loss function for discriminator are given as:

$$\mathcal{L}_{\text{VAE}}^{adv} = -\mathbb{E}[\log(D(q_\phi(z_L|x_L)))] - \mathbb{E}[\log(D(q_\phi(z_U|x_U)))] \quad (2)$$

$$\mathcal{L}_D = -\mathbb{E}[\log(D(q_\phi(z_L|x_L)))] - \mathbb{E}[\log(1 - D(q_\phi(z_U|x_U)))] \quad (3)$$

By combining Eq. (1) and Eq. (2) we obtain the full objective function for the VAE in VAAL as below

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{VAE}}^{trd} + \mathcal{L}_{\text{VAE}}^{adv} \quad (4)$$

For sampling strategy in VAAL use the probability associated with the discriminator's predictions as a score to collect b number of samples in every batch predicted as "unlabeled" with the lowest confidence to be sent to the oracle. Note that the closer the probability is to zero, the more likely it is that it comes from the unlabeled pool.

3. Experiments: Image Classification and Segmentation

For image classification We experimented with ImageNet [3] with more than 1.2M images of 1000 classes. For semantic segmentation, we evaluate our method on BDD100K [10] datasets with 19 classes which is a diverse dataset with 10K images with full-frame instance segmentation labels. Fig. 2 shows that we improve the state-of-the-art by 100% increase in the gap between the accuracy achieved by the previous state-of-the-art methods (Core-set and Ensemble) and random sampling. As can be seen in Fig. 2, this improvement can be also viewed in the number of samples required to achieve a specific accuracy. For instance, the accuracy of 48.61% is achieved by VAAL using 256K number of images whereas Core-set and Ensembles w. VarR should be provided with almost 32K more labeled images to obtain the same performance. Random sampling remains as a competitive baseline as both DBAL and MC-Dropout perform below that.

On BDD100K VAAL is able to achieve %mIoU of 42.3 using only 40% labeled data while the maximum mIoU we obtained using 100% of these datasetes is 44.95. In terms of required labels by each method, in order to reach 41% of mIoU, other baselines need 5% − 10% more annotations than VAAL requires only 30%.

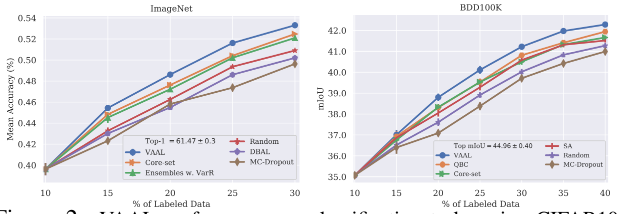

Figure 2: VAAL performance on classification tasks using CIFAR10, CIFAR100, Caltech-256, and ImageNet compared to Core-set [8], Ensembles w. VarR [1], MC-Dropout [6], DBAL [7], and Random Sampling. Best visible in color. Data and code required to reproduce are provided in our code repository

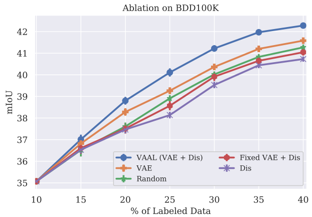

Figure 3: Ablation results on analyzing the effect of the VAE and the discriminator denoted as Dis here. Data and code required to reproduce are provided in our code repository

Ablation Study: Figure 3 presents our ablation study including: 1) eliminating VAE, 2) Frozen VAE with D, 3) eliminating D. In the first ablation, we explore the role of the VAE by having only a discriminator trained on the image space to discriminate between labeled and unlabeled pool. As shown in Fig. 3, this setting results in the discriminator to only memorize the data and yields the lowest performance. In the second ablation scenario, we add a VAE to the previous setting to encode-decode a lower dimensional space for training D. However, here we avoid training the VAE and hence merely explore its role as an autoencoder. This setting performs better than having only the D trained in a high dimensional space, but yet performs similar or worse than random sampling suggesting that discriminator failed at learning *representativeness* of the samples in the unlabeled pool. In the last ablation, we explore the role of the discriminator by training only a VAE that uses 2-Wasserstein distance from the cluster-centroid of the labeled dataset as a heuristic to explicitly measure uncertainty. In this setting, we see an improvement over random sampling which shows the effect of explicitly measuring the uncertainty in the learned latent space. However, VAAL appears to outperform all these scenarios by implicitly learning the uncertainty over the adversarial game between the discriminator and the VAE.

4. Conclusion

In this paper we proposed a new batch mode task-agnostic active learning algorithm, VAAL, that learns a latent representation on both labeled and unlabeled data in an adversarial game between a VAE and a discriminator, and implicitly learns the uncertainty for the samples deemed to be from the unlabeled pool. We demonstrate state-of-the-art results on large-scale image classification and segmentation datasets.

References

[1] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 2

[2] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996. 1

[3] Jia Deng, Wei Dong, Richard Socher, Lie-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 2

[4] Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, and Marcus Rohrbach. Uncertainty-guided continual learning with bayesian neural networks. In *International Conference on Learning Representations*, 2020. 1

[5] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial continual learning. *arXiv preprint arXiv:2003.09553*, 2020. 1

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 2

[7] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*, 2017. 2

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

[9] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5972–5981, 2019. 1

[10] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. 2