M-Protain Diagnostics Using Generative Active Learning

Hanyu Li
Department of Physics
Tsinghua University
l-hy16@mails.tsinghua.edu.cn

Junsong Yuan
Department of Computer Science and Engineering
University at Buffalo
jsyuan@buffalo.edu

Abstract

With comparatively less labeled data and high labeling cost, most of the medical involved tasks can not be directly tackled by state of art machine learning approaches for their lack of large carefully labeled datasets. Our paper is based on the a dataset of immunofixation electrophoresis(IFE) images used in the M-protain diagnostics that has no annotation. In order to make the diagnostics process more efficient, our paper try to train a binary classifier(normal or not) with only few data instance labeled by human experts and all other unlabeled data. We do the semi-supervised training by combining active learning with generative models. In our proposed method, we do these things iteratively: first we find the most uncertain data instances in the latent space of the generative model using the classifier; then we generate synthetic IFE images for human oracle to annotate; afterwards we add these labeled data back in the training set of the classifier. In addition, according to prior knowledge of the IFE images, we propose a specific explainable generative model based on Gaussian mixture model(GMM) that is only effective in this dataset, and compare the result of it with universal effective generative model like GAN and VAE. We conduct extensive experiments to demonstrate the difference between applied generative models, evaluate the effect they make on active learning quantitavely, and explore the reason behind the results.

Introduction

As deep models achieve astonishing results in almost every machine learning tasks, some unavoidable problems such as the need for large carefully labeled dataset has troubled researchers from the start. Part of the reason behind the tremendous success in deep learning is the availability of large-scale labeled data(Sun et al. 2017). Although data labeling companies and platforms claim that they can provide inexpensive yet high quality data(Buhrmester, Kwang, and Gosling 2011), achieving such datasets can be extremely costly or even unrealistic in the scenarios where labeling requires high professionality. For instance, some medical image tasks can not be labeled by people without systematic training. However, the small number of these experts has determined that large-scale dataset is difficult to biuld. Plus, they are probably already preoccupied.

M-protain stands for Myeloma protein, which can be identified by applying immunofixation electrophoresis(IFE) because its sharp monoclonal band in the image. Different categories of results may indicate MGUS, smouldering myeloma(sMM), or multiple myeloma(MM). It usually takes three doctors to examine the IFE image and reach the final conclusion. Therefore the process is highly time consuming. In real world scenarios, more than half of the electrophoresis results are obviously normal which do not need further concern. Although final decision should be made by doctors, if a classifier can give an indicating result, then time can be greatly saved. Every IFE image consists of five columns of one dimensional signal. According to these columns, a 12 dimensional 0/1 vector will be computed as output. Part of the proposed method can be implemented by manually constructed rules, so in the end, machine learning involved section is narrowed down to a binary classification(normal or abnormal) of one dimensional signal.

A large dataset of IFE images is accessible, but each of them comes with a diagnostics report instead of the binary output of each column. Thus, due to the lack of explicit label, this is a classic semi-supervised situation. Our paper tries to utilize the combination of active learning and generative model to tackle with this unlabeled dataset. Active learning is that a machine learning algorithm that can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the training data from an unlabeled dataset(Settles 2009). When using same amount of labeled data, active learning tends to achieve better results. Thus, active learning techniques significantly reduce the amount of labor required compared to manually label all existing data. Deep generative models including GAN and VAE are currently purveiling in a variety of applications. Firstly proposed by Goodfellow *et al.* in 2014, generative adversarial nets(GANs) has drawn a lot attention in the field of computer vision, natual language processing, and etc. Previous researchers inspired by its adversarial structure developed a large amount of variations that can be applied to a variety of tasks. Some of the most well-known models are WGAN(Arjovsky, Chintala, and Bottou 2017), CGAN(Mirza and Osindero 2014), BiGAN(Donahue, Krhenbhl, and Darrell 2016), VAEGAN(Larsen et al. 2015), and so on. Variation auto encoder(VAE) on the other hand, tries to develop a generative model in a variation inference kind of way(Kingma and

Welling 2013). Although more robust to small pertubations in latent space, VAE tends to lose more graphic details when reconstructing the input data. Zhu and Bento made the first attempt(Zhu and Bento 2017) to generate data for active learning process. Since then, a number of research have been conducted to find the most effective way to boost the active learning performance using generative models.

Our paper try to utilize the combination of active learning and generative models to gain a task learner for IFE column binary classification. By generating(decoding) synthetic IFE images based on latent space of the generative models, the model can query the least certain generated instances in the latent space, and thus improve the classifier in the latent space iteratively. In addition, according to prior knowledge of the IFE images, we propose a specific explainable generative model based on Gaussian mixture model(GMM) that is only effective in this dataset, and compare the result of it with universal effective generative model like GAN and VAE. We conduct extensive experiments to demonstrate the difference between applied generative models, evaluate the effect they make on active learning quantitavely, and try to find the reason behind it.

Related Work

Multiple techniques were used to tackle with small datasets and partially labeled datasets including active learning(Settles 2009), generative models(Goodfellow et al. 2014)(Kingma and Welling 2013), data augmentation(Tanner and Wong 1987), and domain transfer(Pan and Yang 2009) etc. (Zhu and Bento 2017)

Preliminaries

The problem is defined as below

Proposed Method

Our method make use of

Experiments

I did such experiments

Discussion

After I did these experiments

Conclusion

To sum up

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6(1):3–5.

Donahue, J.; Krhenbhl, P.; and Darrell, T. 2016. Adversarial feature learning.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Larsen, A. B. L.; Snderby, S. K.; Larochelle, H.; and Winther, O. 2015. Autoencoding beyond pixels using a learned similarity metric.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets.

Pan, S. J., and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10):1345–1359.

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

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 843–852.

Tanner, M. A., and Wong, W. H. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association* 82(398):528–540.

Zhu, J.-J., and Bento, J. 2017. Generative adversarial active learning. *ArXiv* abs/1702.07956.