M-Protain Diagnostics Using Generative Active Learning

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. In order to figure out the best representation of this dataset, 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. With the desire to fill this gap, our paper combines active learning and generative model to achieve relatively good results on small datasets.

For instance, a large dataset of medical images regarding M-protain diagnostics is accessible. Nonetheless, each image comes with a diagnostics report without the trainable label that we desire. 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 provide an indicating result, then time can be significantly saved. 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.

Due to the lack of explicit label, this is a classic semi-supervised learning task. 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 same amount labeled training instances if it is allowed to choose the data to be labeled from an unlabeled dataset (Settles 2009). 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. 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 also tries to utilize the combination of active learning and generative models to gain a task learner, but we made specific augmentations specifically 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

Several techniques were widely used by researchers to tackle with small datasets and partially labeled datasets. For example, active learning(Settles 2009) tries to pick the most informative data to label so that models can learn better when the total number of labeled data is limited; generative methods(Kingma et al. 2014)(Springenberg 2015) wish to benifit target classifier using knowledge of data distribution learnt by generative models; data augmentation(Tanner and Wong 1987) enriches small dataset by generating new synthetic data; domain transfer(Pan and Yang 2009) utilizes large, easily acquired datasets in different tasks or different settings to help the training of target learner. In this paper we mainly focus on active learning and generative methods.

Active Learning

In general, active learning has three main settings: membership query synthesis, stream-based selective sampling, and pool-based active learning(Pan and Yang 2009). Among them the last one is mostly refered to. Pool-based active learning gather a large pool of unlabeled data and ranks their informativeness according to an acquisition function that may coevolve with the task learner in the training process. Various acquisition functions have been developed by researchers. For example, some acquisition function can be derived by maximising the uncertainty of the current learner(i.e. finding the most uncertain data in the pool). Another acquisition function maximises the Bayesian Active Learning by Disagreement(BALD)(Houlsby et al. 2011). It can choose pool points that are expected to maximise the information gained about the model parameters through maximising the mutual information between predictions and model posterior. Nonetheless, active learning framework was specifically designed for datasets with rather small-scaled labeled data, and even after the iterative process of oracle labeling, the total number of trainable data is still significantly less than classic deep learning scenarios. Under such constraint, the task learner still suffer from overfitting problem when dealing with high-dimenisonal data. Only few researchers appled active learning on high-dimensional data, and to our best knowledge, almost all of them proposed special techniques(Gal, Islam, and Ghahramani 2017).

Generative Methods

Firstly proposed by Goodfellow *et al.* in 2014, generative adversarial nets(GANs) (Goodfellow et al. 2014) 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 various kinds of tasks. Some of the most well-known variations of GAN includes WGAN(Arjovsky, Chintala, and Bottou 2017) which stablizes training process and provide solution to mode collapse problem, CGAN(Mirza and Osindero 2014) that firstly introduced models that can generate specific classes according to the label fed as part of the input, BiGAN(Donahue, Krähenbühl, and Darrell 2016) that enables bidirectional transformation, VAEGAN(Larsen et al. 2015) that connects GAN and VAE using a conjuncted generator/decoder, 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).It tries to capture the most important latent variables that determine how real data are formed. Although more robust to small pertubations in latent space, VAE tends to lose more graphic details(more blurry) when reconstructing the input data compared to GAN due to its pixel-wise reconstruction loss.

Generative Active Learning

Conventional pool-based active learning method draws data to be labeled from the large unlabeled data pool. However, data in the unlabeled dataset are spread all over the data space. Hence, after exhausting the most informative data, the acquisition function will begin to query less informative data later on in the training process. However, if queried data are all synthesized(i.e. generated), its informativeness will not degrade over time. Moreover, given that the generative model fully captured the true distribution of data, it can generate informative data whose mode is not fully covered in the original dataset(Zhu and Bento 2017). Also, with promising conditional generative model like ACGAN, researchers came up with several models that free active learning from low dimensional data since unlimited labeled data can be generated(Kong et al. 2019). This great advantage comes with high risks though. Active learning demands data with sufficient information. In other words, generated data should yield high uncertainty. Thus, automatically labeled generative models suffer from the trade off between informativeness and safety of data. More common methods are to do active learning process in the latent variable space of the generative model. Huijser and Gemert proposed a novel boundary annotation approach to train a classifier in the latent space, and every chosen data point in the latent space can be seen as a new generated data using generative models(Huijser and van Gemert 2017). In another work, Tran *et al.* modified iterative Bayesian data augmentation(BDA) framework using active learning and proposed VAE ACGAN model to tackle with the problem of BDA that it tends to generate less informative augmented data(Tran et al. 2019).

Preliminaries

The entire defination of the task can be illustrated as below. Every IFE image consists of five columns of one dimensional signal with length of 300. Among them the first three column represents the heavy chain G, A, M, and the last two

is the light chain κ, and λ. According to an IFE image, a 12 dimensional 0/1 vector will be computed as output. The last five of this vector is whether the five columns is normal(0) or abnormal(1, contains obvious sharp monoclonal band). The first one is 0 if and only if all the last five are zero. The six variables in the middle represents the 3×2 matching between the heavy chain and light chain, it will be abnormal(1) if corresponding columns have matching monoclonal band. Manually set rules using GMM model is able to find the peak position and do the matching automatically but its accuracy of judging if one column is abnormal needs improving. This is where machine learning based model kicks in. If a filter of normal columns(i.e. a binary classifier) can only take all the abnormal columns in to consideration, than we can skip the bad performance part of manual rules and replace it with a reliable machine learning algorithm. To sum up, the task learner should in the end be able to output binary results on a 300 long one dimension signal. We can denote this data space by $\mathcal{X} \subset \mathbb{R}^{300}$, and every $x_i \in \mathcal{X}$ is a point in data space. Accordingly, latent space data point will be denoted by $z_i \in \mathcal{Z} \subset \mathbb{R}^n$, $\mathcal{Z}$ is the latent space and n is a fixed hyperparameter which should be significantly smaller than 300. A classifier $g(x_i)$ maps a data point to a binary label $y_i \in \{0, 1\}$. Because of the constraint on the amount of labeled data and consquently on the dimension of trained data, the classifier that is actually going to be trained is $f(z_i)$.

Proposed Method

Our framework consists of two steps: the first one is to train of a bidirectional generative model; and next we iteratively conduct the active learning process where we sample data points in latent space according to the acquisition function, transform it to 300 dimension signal for human experts to label, then put the labeled latent space data back to training set. Generative model will be fixed after the first step. All the data picking process are conducted in the latent variable space.

Choice of Generative Models

Generative models in most cases are mapping functions from low-dimension prior of latent variables to high-dimension data space whose parameters maximises the likelihood of existing data(Goodfellow 2016). After training with adequate amount of data, we will have enough knowledge of all the decisive latent variables and we know exactly how these variables influence the corresponding point in data space. However, just generative is not enough. In testing state we need to encode all the test data to latent space. Hence, only bidirectional generative model are considered. Here, not only did we tried out classic bidirectional generative models including BiGAN and VAE, but we also hypothesized a GMM generative model using prior knowledge of the formation of IFE image. We can denote the decoder(generator) and encoder of the bidirectional model respectively as $D(z_i)$ and $E(x_i)$ so that we have $g(x_i) = f(E(x_i))$. Of course we expect that $D(E(x_i))$ to be as close as it can to x_i, or classifing z_i will not provide any useful information.

GMM Choosing Gaussian mixture model to represent the formation of IFE image is not only natural, but also reasonable. For monoclonal band, the target protain are all the same. Therefore if the electrophoresis procedure introduce no systematic error, all the monoclonal protain should be in the same spot since the position that any type of protain appears on the IFE image depends solely on its molecular properties such as weight and the electric charge it carries. In reality, noise is unavoidable, and we can suppose that the accumulation of small noises is under normal distribution. On the other hand, the polyclonal band can be perceived as a normal distribution whose mean (and variance?) also obeys normal distribution becuase it contains many different kinds of protain which have different weight and electric proterty. Every kind of protain forms a normal distribution due to the noise of electrophoresis. Thus, it is still a normal distribution(can be proved, need to find prove or prove myself). From above, we can derive that the one dimension signal is the empirical distrubution of the summation of several gaussian components. Empirically, in real data the number of gaussian components will not exceed 3.

In practice though, we use a variation of GMM that do not need to suppose the number of gaussian components.

Active Learning Process

After we completes the first stage of training process, we now have a low dimension embedding of the real data(300 dimension signal). The second stage is to actively train a classifier in this latent variable space. Each iteration of active learning contains the training of classifier, sampling data to be labeled, expert labeling and updating the training set. The main difference between generative active learning and traditional pool based active learning is the way of sampling. Here we can sample any point near the decision hyperplane because we can transform any point in latent space to a point in data space. According to the classifier, different sampling strategy can be used.

Experiments

Different Representations(Generative Models) and Classifiers

Different approaches to find the mapping between data space and latent space should be measured in several ways. For now, we compared all the combinations of different generative model and weak classifier in the latent space on 40 training data with label, 2000 synthesized data without label, and 20 testing data. We did not enable active learning in this comparative stage, so basically it is just a standard framework with feature extraction and classification. Plus, we constructed a comparative end to end neural network consists of VAE and MLP that can only utilizes the small labeled dataset. We aim to evaluate the performance of four representations: GMM, VAE, AE(auto-encoder), PCA(principle component analysis), four weak classifier: linear SVM, rbf SVM, decision tree, random forest. In addition, hand crafted rules including GMM rule and derivative filter are tested, and also the end to end neural network we mentioned.

Firstly, reconstruction error is the most important property of a representation. In ideal case, the mapping should be injective. If not, the classifier in latent space will not be able to provide valid judgement in data space. Partly because of the lack of large unlabeled real dataset, GMM achieved overwhelming advantage compared to other representations.

Secondly, the test accuracy of different combinations showes that

Because of the active learning process, the embedding methods should satisfy these properties below.

Discussion

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.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1183–1192. JMLR.org.

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.

Goodfellow, I. 2016. Nips 2016 tutorial: Generative adversarial networks.

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

Huijser, M., and van Gemert, J. C. 2017. Active decision boundary annotation with deep generative models. In *Proceedings of the IEEE international conference on computer vision*, 5286–5295.

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

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, 3581–3589.

Kong, Q.; Tong, B.; Klinkigt, M.; Watanabe, Y.; Akira, N.; and Murakami, T. 2019. Active generative adversarial network for image classification.

Larsen, A. B. L.; Sønderby, 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.

Springenberg, J. T. 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*.

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.

Tran, T.; Do, T.-T.; Reid, I.; and Carneiro, G. 2019. Bayesian generative active deep learning.

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