Classification on medical images using two-cycle semi-supervised learning and mix-up algorithm

Chiang, Chia-Cheng[106062214] and Pu, Ching[106062107]

National Tsing Hua University, Hsinchu City 300, Taiwan

Abstract. Deep learning has achieved considerable results upon varieties of specific field.However, large amount of high-quality dataset is required for training such a deep learning model.Collecting and labeling both require considerable cost.For medical image analysis, labelling requires further expert domain knowledge. Thus, we are trying to improve the performance of using less labeled data but making further improvement by unlabeled data.We apply a semi-supervised learning technique called two-cycle learning [4], trying to have better feature extraction.We also apply a data augmentation technique called mix-up [5], to make the limited data more general. Finally, we compare the experiments we made, and get considerable results.

Keywords: Two-cycle learning · Mix-up · Semi-supervise learning

1 Introduction

Nowadays, more and more artificial intelligence techniques have been applied on medical images analysis. For example, magnetic resonance imaging(MRI) [7], computed tomography (CT) [8], have greatly increased knowledge of normal and diseased anatomy for medical research and are a critical component in diagnosis and treatment planning. With applying an algorithm called image data segmentation, radiologists can obtain extra assistance on delineation of anatomical structures.

Currently, one bottleneck on such approach is that the ground-truth data with expert domain knowledge is too expensive. We can hardly collect a valuable dataset that is huge enough. Even though some united hospital have the ability to collect such dataset, the task may cost over ten years. Anther issue on medical image analysis is that the collected dataset is mostly specific on certain body parts or disease, which makes the existed dataset can't be shared through different tasks.

Trying to solve such restriction, we would like to improve the performance on limited amount of ground-truth data to decrease the dependency on data-collection.There are kinds of approaches, e.g. data augmentation using GAN [2], or semi-supervised learning.We assume that GAN is too costly on computing and still has many restriction like stability.Thus, we are going to apply another data-augmentation approach called mix-up, which is a simple but effective method to train a general model even using little amount of data. Another method we apply

is semi-supervised learning.Semi-supervised learning is to use extra unlabeled data to provide further information.One method is to give the unlabeled data so-called pseudo label.But the policy to determine the pseudo label is another question.We here according to the assumption that the data with the same label should be close on the projected feature map use a two-cycle learning to obtain a more tightly-bounded feature map.More details can be seen on following sections.Since we believe that if we can have better results on classification, we can apply our results to other fields, we here use normality classification as our benchmark.However, we believe that our research can be general upon different application.

2 Methods

2.1 Two-cycle learning

Two-cycle learning is a semi-supervised learning approach.It can be separated into two part:1. First cycle(Clustering) 2. Second cycle(Pseudo-labeling).Based on the assumption that the data that has the same label may be close on the projected feature map, we first use an unsupervised clustering method to assist the feature extractor to learn the ability to separate the data into different clusters.Secondly, we'll make use of the feature extractor to apply the K-NN algorithm to get the pseudo label of the unlabeled data.After doing the pipeline of training, we then obtain a reliable model.

First-cycle learning

In the first cycle, we use a pretrained neural network to project the raw data into a feature map, applying K-means algorithm on all data regardless of labeled or unlabeled.By doing so, each data will be given a cluster id.In high-level presentation, it can be interpreted as the discrete expert domain knowledge that the radiologist can make use of.The distance of two data thus implies the difference between them.Then, we'll use a fully-connected layer to predict the cluster-id among these data.Through the training going on, the network will be equipped with the ability to perform the ability of clustering, which is believed useful on the following training.

Second-cycle learning

In the second cycle, after we derive a network that is equipped with the ability of separating data into different clusters from first cycle, we then apply K-NN algorithm to give the unlabeled data a pseudo label.Since the feature map we apply the K-NN algorithm can provide a more precise standard on how to determine the pseudo label, we then apply supervised learning approach to train a anomaly classification model.With the extra unlabeled data, we can conquer

the restriction on limited data and over-fitting issue. Figure 1 is an visualization of two cycle learning.

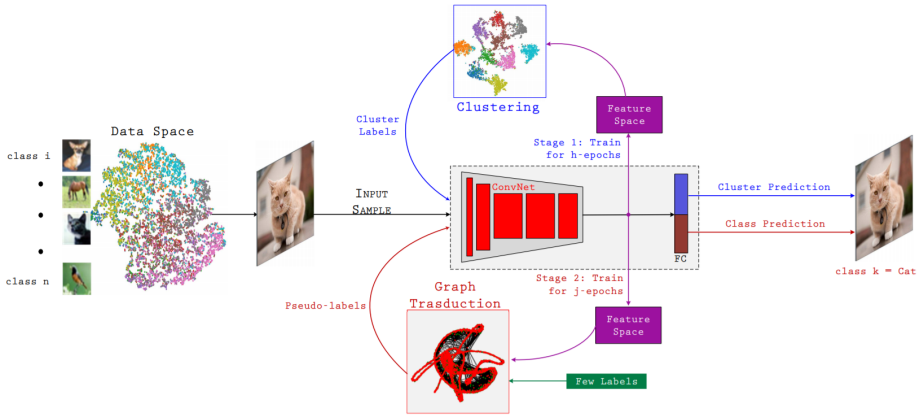


Fig. 1: Two cycle learning visualization

2.2 Mixup

The data augmentation method we chose is mixup [5]. Mixup is a data augmentation strategy which use interpolation technique to mix two different images. First, it will make an image paired with another image in the same batch when training. Then these paired images will be mixed according to a parameterλ $(0 < \lambda < 1)$ which means the new image contains λ first image and $1 - \lambda$ second image. Note that λ is not a fixed number, it is sampled from an arbitrary probability distribution instead. When computing the loss of the mixed image, it will use the same method to mix the loss up, which means it'll compute two loss using the mixed image and the two original labels, then use λ and $1 - \lambda$ to mix them up. Just like the equations shown below.

$$
\begin{aligned}
x' =& \lambda x_i + (1 - \lambda) x_j \\
y' =& \lambda y_i + (1 - \lambda) y_j \\
loss =& \lambda criterion(y', y_i) + (1 - \lambda) criterion(y', y_j) \\
where\ \lambda \in& [0, 1]\ is\ a\ random\ number
\end{aligned}
$$

The advantage of using mixup as the data augmentation strategy is the model can have a clearer information of what it looks like between any two labels, and thus it can make a more precise decision boundary.

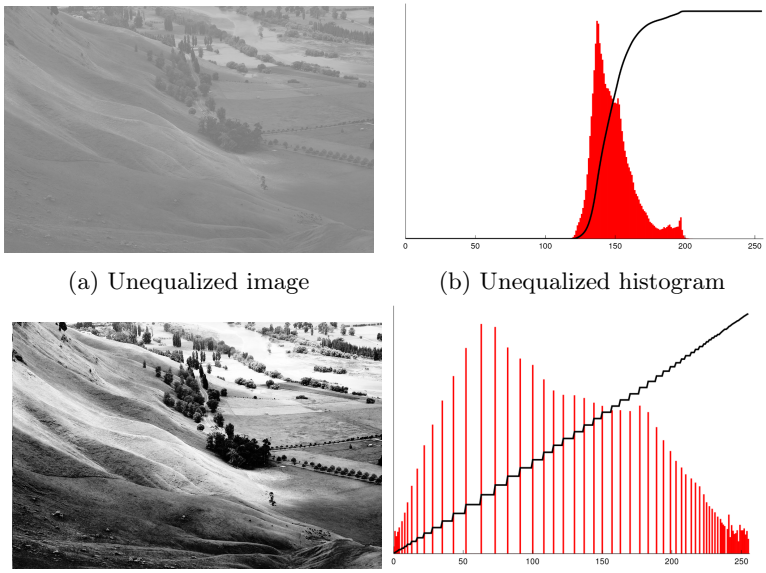
3 Experiments

3.1 Data set

We use MURA (musculoskeletal radiographs) as the experiment dataset. MURA contains large amount of bone X-ray and each of them is labeled as normal or abnormal. To simulate the data imbalance situation (The unlabeled data is much more than labeled data), we use a parameter $r = \frac{all\ data\ number}{labeled\ data\ number}$ to denote the ratio between labeled data and unlabeled data. Then label in the dataset are randomly deleted to satisfy r.

3.2 Preprocess

In the MURA dataset, images are in different size and direction (direction means the image's long side is horizontal or vertical), thus we will first do rotating to every image to make their long side be horizontal. Then we will do scaling to make each image in the same size. In the end, we will do histogram equalization [1] to gain the contrast, which make the x-ray's detail and boundary clearer. Like Figure 2a is the image before histogram equalization, Figure 2b is the histogram before histogram. After histogram equalization, the distribution of histogram will be more distributed like Figure 2d, which result in the new image like Figure 2c .



(a) Unequalized image (b) Unequalized histogram

(c) Equalized image (d) Equalized histogram

Fig. 2: Histogram equalization comparison

3.3 Framework

In our implementation, we chose resnet101 [3] as our backbone. To obtain an elementary, it will be trained on labeled data for 40 epochs. Then first cycle training will be carried out, the first cycle training will first use the model's convolution layer to obtain the feature vectors of all data. Next, k-means clustering will be done on these feature vectors to obtain a cluster number for each data. Lastly, we will train the model on all the data with cluster number as their label. Note that the model's fully connected layer will first be reset, and the output layer dimension is the cluster number, when training the model with cluster number, the convolution layer is freeze for $\frac{m}{3}$ epochs, then we will unfreeze the convolution layer and train the model for m epochs. All the training process above will repeat for n times, which means do the first cycle training for n cycle. The second cycle training is like first cycle training. The only difference is after we obtain all data's feature vector, we will do KNN on them to predict unlabeled data's pseudo label with labeled data. Then reset model's fully connected layer, and its output dimension is equal to data's true label number. The rest training detail is the same as first cycle learning. Note that the second cycle learning will also do n times.

3.4 Evaluation

We have done a series of experiments and use accuracy, precision, recall and f1-score to evaluate the methods' performance. In Table 1, baseline denotes the model only train on labeled data for 80 epochs. All those experiments with mixup denotes the training data have been mixed up. Those experiments with(n, m) prefix denotes that it is a two cycle learning experiment, and first cycle and second cycle are repeated for n times respectively, in each cycle training, the epoch parameter is m.

	unbalance ratio:10						unbalance ratio:50			
	baseline	baseline mixup	(2,20)	(2,20) mixup	(5,10)	(5,10) mixup	baseline	baseline mixup	(2,10)	(2,10) mixup
accuracy	0.589	0.668	0.627	0.681	0.642	0.660	0.577	0.659	0.599	0.616
precision	0.594	0.664	0.663	0.692	0.658	0.694	0.593	0.645	0.617	0.635
recall	0.590	0.676	0.628	0.686	0.634	0.650	0.597	0.668	0.613	0.635
f1-score	0.592	0.670	0.644	0.689	0.646	0.671	0.595	0.656	0.615	0.635

Table 1: Experiment result table

3.5 Discussion

According to our experiments, we showed that both two-cycle learning and mix-up algorithm can achieve respectable improvement under restricted condition.One question is that the ratio between the number of cycle and the number of epochs is not relevant to the performance of networks. We are not sure about it is due to the limited experiment setting or other environment factors, but we treat it as a temporary conclusion.Another interesting result is that under extremely limited data, mix-up is more effective than two-cycle learning.We assume that since mix-up can provide continuous information instead of discrete information, which is a critical factor that can prevent training from over-fitting.

4 Conclusion

Since the limitation of data collection, deep-learning-based approach on medical image analysis is still not generally applied.However, if the reduction of ground-truth training data is possible, it'll be the shortcut of further realization for models to understand expert domain knowledge.Our work contributes to show that it is truly possible to achieve such goal under tough condition.Our future work will be to make realization on the realization under limited situation and try to prove that our work can be applied on different task, e.g. image segmentation.

5 Contribution

Each member's contribution to this project is the same.

References

1. Computer Vision, Graphics, and Image Processing, 1987. Adaptive histogram equalization and its variations. 38(1), p.99.
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2020. Generative Adversarial Networks. [online] arXiv.org. Available at: <https://arxiv.org/abs/1406.2661> .
3. He, K., Zhang, X., Ren, S. and Sun, J., 2020. Deep Residual Learning For Image Recognition. [online] arXiv.org. Available at: <https://arxiv.org/abs/1512.03385> .
4. Sellars, P., Aviles-Rivero, A. and Schönlieb, C., 2020. Two Cycle Learning: Clustering Based Regularisation For Deep Semi-Supervised Classification. [online] arXiv.org. Available at: <https://arxiv.org/abs/2001.05317> .
5. Zhang, H., Cisse, M., Dauphin, Y. and Lopez-Paz, D., 2020. Mixup: Beyond Empirical Risk Minimization. [online] arXiv.org. Available at: <https://arxiv.org/abs/1710.09412> .
6. Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R., Langlotz, C., Shpanskaya, K., Lungren, M. and Ng, A., 2020. MURA: Large Dataset For Abnormality Detection In Musculoskeletal Radiographs. [online] arXiv.org. Available at: <https://arxiv.org/abs/1712.06957> .

7. Lundervold, A., amp; Lundervold, A. (2018, December 13). An overview of deep learning in medical imaging focusing on MRI. Retrieved June 25, 2020, from https://www.sciencedirect.com/science/article/pii/S0939388918301181
8. Cheng, J., Ni, D., Chou, Y., Qin, J., Tiu, C., Chang, Y., . . . Chen, C. (2016, April 15). Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Retrieved June 25, 2020, from https://www.nature.com/articles/srep24454