Extrapolative Data Augmentation of Scarce Abnormal Medical Images by Using Normal Images and Domain Knowledge
正常画像とドメイン知識を利用した異常医用画像の外挿的データ拡張

by

Takahiro Suzuki
鈴木陽大

A Senior Thesis
卒業論文

Submitted to
the Department of Information Science
the Faculty of Science, the University of Tokyo
on January 31, 2020
in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Science

Thesis Supervisor: Issei Sato　佐藤一誠
Lecturer of Information Science

ABSTRACT

Data augmentation has been widely used for improving the accuracy of classification. In medical image data analysis, abnormal image augmentation is particularly important because the number of abnormal images is often much less than that of normal images. However, previous work on data augmentation is mainly based on a large number of data for each class. Therefore, abnormal medical image augmentation still remains challenging when there are not enough abnormal data. In this thesis, we propose a novel scheme for generating abnormal images by using normal images and domain knowledge, without explicitly using abnormal images. We analyze the performance of our method by generating lung opacity images from normal lung images that are provided by National Institutes of Health.

論文要旨

分類手法の正確性を上げるために、データの増強は広く用いられている。医用画像データ分析において、異常画像のデータ数は正常画像の数に比べて一般に少ないため、異常画像の増強は特に重要である。しかしながら、既存研究は主に、大量の同じクラスのデータを用いたデータ増強に着目している。したがって異常データが充分にないときは異常画像の増強は難しいままである。この論文で我々は、異常画像を直接的に使うことなく、正常画像とドメイン知識を活用して異常画像を生成する革新的な手法を提案する。我々はアメリカ国立衛生研究所が提供する正常な肺画像をもとに肺炎の画像を生成することで、我々の手法の性能を分析する。

Acknowledgements

I would like to say a big thank you to my supervisor, Lecturer Issei Sato for his generous support and guidance. I am also indebted to Assistant Professor Shouhei Hanaoka and Project Lecturer Yukihiro Nomura for their technequal advices. I am also really grateful to Professor Sugiyama for revising the thesis.

Contents

Chapter 1

Introduction

Deep neural networks have achieved high performances in several computer vision tasks such as image classification, object recognition/detection, semantic/instance segmentation, image captioning. For example, Residual Network (ResNet) [5] is a well-known architecture that achieved state-of-the-art performances in image classification tasks [2, 5]. These discriminative models tend to work well when trained on a large amount of labeled data, and data augmentation is a commonly-used technique to increase the number of training data. Especially in medical images, the number of abnormal images with lesions are often much less than that of normal images, and it is not so rare that the disease rate is under 1% for scarce diseases. In addition, imbalance between normal and abnormal images makes it difficult to train the discriminative model. Therefore, for medical images, the augmentation of scarce abnormal images has a great significance.

Previous studies about data augmentation methods can be generally classified as follows: data preprocessing approach and model-based approach. The former approach includes primitive and simple operations such as rotating and flipping. Mixing up several images to generate new images [11, 16] is also this kind of approach. In contrast, the latter approach uses generative models such as Generative Adversarial Networks (GANs) [4] or Variational Auto Encoders (VAEs) [9]. In the subsequent paragraphs, we give an overview and limitations of these previous approaches.

First think about the data preprocessing approach. Primitive operations such as rotation and flips is considered to have only limited effect, because it can do only naive data expansion. Not only that, medical images are usually taken under the condition that a subject is fixed to some imaging and medical equipments, and thus medical images are relatively standardized. Therefore, primitive operations that do work well in ordinal images do have only limited effect in medical images; when there are almost no image in which a subject is titled, why rotation seems effective?

Another example of data preprocessing approach is mixup [11, 16]. The idea is mixing several images and their labels to generate new images. Although this method is innovative in term of the data augmentation beyond classes, it does not have any countermeasures against imbalance in the number of data between different classes. Namely, in order to generate data in the target class, it needs to use some data in the same class. So, when the number of abnormal images is limited, it is challenging to augment abnormal image through this method. This is why we want to augment the scarce data in the target class from a large number of data in another class, i.e., generate abnormal images from normal images.

Next think about the model-based approach. A famous example is Generative Adversarial Networks (GANs) [4] or Variational Auto Encoders (VAEs) [9]. GAN was firstly proposed by Goodfellow et al. [4], and it has attracted a great deal of attentions from researchers. GAN is based on the game theory where a generator try to make realistic images to deceive a discriminator, while the discriminator try to distinguish between real images and the images generated by the generator. The generator and discriminator are trained alternately, and it is expected that a well-trained generator is able to generate realistic images that look similar to real images.

Although GAN has achieved better results in many researches including medical images [3], there is a limitation. A large amount of data from the target class are basically needed for training the generator and discriminator. This implies that abnormal medical data augmentation through this approach [4] is difficult under the condition that there are not a large amount of abnormal medical images.

Similarly, VAE [9] trained in a supervised manner is not appropriate for the same reason: a large number of abnormal images are required for training the model. As an unsupervised learning method, VAE can be used in one-class classification [8], yet there seems to be few studies about that. A possible reason is that the training of VAE is intrinsically difficult, because it cannot optimize its loss function directly, instead it maximizes its variational lower bound [9]. Thus, augmentation of scarce abnormal images remains challenging by previous data augmentation methods.

In this thesis, we propose a novel method for generating abnormal medical images by using normal medical images and domain knowledge. We does not count on any actual abnormal images; thus, our method is applicable and usable even when there are not sufficient number of abnormal images. In more detail, artificial lesions are created by making use of domain expertise and superimposed on normal medical images. In order to cope with the imbalance between normal and abnormal images, we give normal images as input to the proposed simulator incorporating domain knowledge and get abnormal images as output. Furthermore, we also suggest a human-in-the-loop framework for augmenting scarce abnormal images in an extrapolative manner: select high-quality pseudo abnormal images among those generated by the above optimal model, add them to the existing scarce abnormal images used for validation, and begin again to find optimal hyperparameters' values in the proposed simulator.

The concept of the proposed method is to do the data augmentation of scarce abnormal medical images only from normal images and domain knowledge in order to make the classifier that can properly predict whether an input is normal or abnormal. We also want to do the data augmentation in an extrapolative way in order to effectively increase not only the amount but also the variety of the scarce abnormal images.

In order to realise the purpose, we use the proposed simulator incorporating domain knowledge to get pseudo abnormal images from normal images. Both the generated images and the normal images are used to train a classifier, and the validation data that consists of normal images and actual scarce abnormal images is used to measure the generalization performance of the trained classifier. The proposed simulator has some hyperparameters, and optimization methods are effectively used to find the optimal values of the hyperparameters with which the simulator shows high generalization performances. We also use the proposed human-in-the-loop framework for extrapolative data augmentation: among the pseudo abnormal images generated by the optimal simulator, select highly qualified pseudo abnormal images and add them to the actual abnormal images. After this big loop is over, one can run the small loop again, and then continue this cycle to expand the abnormal images more and more.

The proposed data augmentation methods are novel in terms of the ability of coping with the data imbalance and the incorporation of extrapolation; yet, there are some challenging points that derived from the difficult problem settings. First, since the simulator uses domain knowledge to compensate for the lack of data, it is necessary to use a different simulator for each disease case. In this thesis we propose the simulator that work well on X-ray images of pneumonia, but for another rare disease case, one need to make or find another simulator. Second, it is not easy to find good hyperparameters of the simulator that makes a plenty of high-quality pseudo abnormal images. One reason is that, in this paper, we used only AUC to find and decide the optimal simulator; yet it is not always an excellent evaluation criteria, because sometimes AUC becomes high with the generated images whose artificial lesions are emphasized than natural lesions. So we may need more evaluation standards in addition to AUC. The second reason is that, the relation between hyperparameters and AUC is a black box function, so it is intrinsically difficult to efficiently search even though a searching algorithm is used. We actually want to set up and optimize a certain loss function directly in a differential way.

Table 1.1: Comparison of the proposed data augmentation method with the previous methods. The proposed method is simulation-based and HITL.

	preprocessing	model-based	proposed
required data amount	little	many	little
need to learn	-	✓	-
deal with data imbalance	-	-	✓
Existence of extrapolation framework	-	-	✓

Table 1.1 shows the comparison of the previous data augmentation methods with the proposed method, i.e., simulation-based and human-in-the-loop method. First, since it does just simple operations such as rotating, fliping, and interpolation [11, 16], the data preprocessing method does not require a large amount of data. However, because of the simplicity of the operations, it cannot cope with the data imbalance; data augmentation of the scarce abnormal images remains challenging. Second, as to the model-based method, in the first place, training data is necessary for training the generative models. In the case where the number of abnormal images is rare and small, the abnormal images are insufficient to be used as the training data for the generative models, and thus the training of the models will not work well. In short, the model-based method cannot deal with the data imbalance neither. On the other hand, the proposed method, simulation-based approach, can cope with the data imbalance by using the simulator incorporating domain knowledge, and thus it is possible to generate scarce abnormal images even when the number of the abnorml images is small. Moreover, to the best of our knowledge, we are the first explorer to do the data augmentation in an extrapolative manner by using the proposed human-in-the-loop framework.

Chapter 2

Related Work

2.1 One-class Classification

One-class classification [8], which is a possible method for our task, refers to making a negative/positive classifier that is trained only by using negative data. The primitive method of one-class classification is to use an autoencoder or VAE. The methods based on an autoencoder were recently studied in anomaly detection for medical images [15, 7]. In detail, an autoencoder is trained by decreasing the difference between an input and its reconstruction. Hence, an autoencoder trained by negative data can reconstruct negative inputs with small error, while it is expected that the autoencoder cannot reconstruct positive inputs well because of the first occurrence of positive data. Therefore, it may be possible to distinguish whether the input is negative or positive by using its reconstruction error.

Although this method seems to work well in the previous studies [15, 7], there seems to be some drawbacks of the method for the NIH pneumonia data. Firstly, it is not easy to find an appropriate architecture of the model for the data. Secondly, it just makes an classifier and cannot generate and augment positive data, which might be useful for improving the accuracy of classification. Lastly, the NIH data [14] is different from those of previous studies [15, 7], so there is no guarantee that the model also works well. Indeed, our preliminary experiments showed that the model did not work well for the NIH pneumonia data.

2.2 Other Data Augmentation Strategies

Recently, AutoAugment [1] opened up a new research field of searching data augmentation policies automatically from data. In detail, it sets the search space for policies that consist of simple operations such as rotation or shearing and of some hyperparameters related to the operations. After setting the search space, optimal policies were found by using the validation accuracy of each augmentation and search algorithms by reinforcement learning, which achieved state-of-the-art accuracy on many standard datasets. There are many variants of AutoAugment, and Fast AutoAugment [6] used Bayesian optimization instead of reinforcement learning as a search algorithm and proposed more efficient strategies for shortening the computation time for the search. However, these methods focus on data augmentation within the target class, so augmenting positive medical images is still impossible without using themselves. Yet, the way of thinking is applicable to our methods, i.e, if we are able to establish the good model that only has a few hyperparameters, Bayesian optimization can be used to find the optimal values of the hyperparameters.

We are more interested in data augmentation beyond different classes, and there have been some related researches. [16, 11] proposed the idea of mixing two images and their labels. However, they just considered the interpolation of the two data points. What we want to do in this thesis is to establish the model that generates positive medical images by using many existing normal images. So our interest is “one-way” or extrapolative data augmentation, i.e., the generation of data in the target class from the data in another class. This is much more difficult than the previous problem settings, and to the best of our knowledge, we are the first explorer of the extrapolative data augmentation.

2.3 X-ray simulation

There are several previous studies about the simulation of generating x-ray images from objects [13, 12]. The aim of the simulation is to produce a realistic X-ray images as fast as possible. In order to calculate x-ray attenuation in real time, some techniques were used such as using physical laws or implementing on the Graphical Processing Unit (GPU). For example, [12] used the Beer-Lambert law [13] to calculate x-ray attenuation. Namely, the intensity of an X-ray is expressed as follows:

$$I(x, y) = I_0(x, y) \cdot \exp(-\int_z \mu(x, y, z)dz), \tag{2.1}$$

where $I(x, y)$ represents the intensity at point (x, y) after passing through a subject, $I_0(x, y)$ is an initial intensity of an X-ray at the same point (x, y), $\mu(x, y, z)$ is an attenuation coefficient at spatial coordinates (x, y, z). Note that X is the horizontal axis, Y is the vertical axis, Z is the axis parallel to the X-ray. X, Y, Z form a three-dimensional Cartesian coordinate system, and each of x, y, z represents an element of the corresponding axis.

This law can be used in our approach. In our case, the input is not a virtual object but an image, so conceptually, we need to reconstruct the a physical object from the image. It is challenging in principle, so some tricks and presumption are needed.

Chapter 3

Proposed method

In this chapter, we explain the basic principles of the positive data augmentation in the field of pneumonia. The principles are based on domain knowledge of doctors and are naturally introduced to the proposed model. In addition, we explain the heuristic algorithm of detecting lungs, which was used for implementation. Furthermore, we describe the whole human-in-the-loop architecture that might make it possible to augment positive data in a more reliable way. The model described in this chapter is effectively used in the overall cycle architecture.

3.1 Intuition about the proposed method

Our task is to improve the accuracy of normal/abnormal prediction. Our approach is data augmentation, but the way of generating images is completely different from previous studies. Unlike previous methods, we use normal images and domain knowledge to generate abnormal images, without any use of actual abnormal images. An artificial lesion is created by making use of domain knowledge of experts, and it is superimposed on normal lungs.

The intuition about the proposed method is simple: the difference between abnormal and normal is whether there is a lesion or not, so it may be possible to generate anomaly by superimposing a lesion on a normal image.

3.2 Domain knowledge and formulation about pneumonia

In the experiment, we focused on detecting pneumonia by using the Chest X-ray dataset from National Institutes of Health (NIH) [14]. One expertise on pneumonia is as follows: pathogens enter a certain point in the lungs, from which inflammation spreads isotropically. Therefore, the typical shape of a lesion is considered as a sphere. Another expertise is the Beer-Lambert law [13], which was already mentioned in the previous chapter.

In order to use the law of physics effectively, we want to calculate the intensity from the pixel value at the corresponding position. However, according to a radiologist Shouhei Hanaoka, the relationship between a pixel value and an intensity cannot know exactly, because it is influenced by the specification of used sensor, corrections, and other factors. Thus, without knowing details of the above details, it cannot tell the relationship between them in principle.

Therefore, we need some presumption to calculate the intensity from the pixel value at the corresponding position. Go back to the equation (2.1) again. In the equation (2.1), μ is not negative by definition, so $I(x, y) \leq I_0(x, y)$ holds from the above expression, which implies that the intensity after passing through a subject is weaker than the previous intensity before passing. When the intensity is weakened, the pixel value at the corresponding position (x, y) becomes bigger and the corresponding color gets more white in a grayscale image. In other words, a pixel value is negatively correlated with the intensity at the corresponding position. Therefore, we presume that the relationship between them can be expressed by a linear function with a negative slope.

Now let us think about a linear function f passing through (0,255) and (I_0, 0) in a plane coordinate $I \times V$, where I represents intensity and V is a pixel value. In the following, we explain how to calculate the pixel values of an image with a sphere artifact, after receiving the normal image as an input.

First, $v(x, y)$ denotes the pixel value at a position (x, y) of a given normal image. Then, the intensity at the corresponding point is expressed as $f^{-1}(v)$, which is actually equal to $(1 - \frac{v}{255})I_0$. Next, by the above equation (2.1), the intensity after passing through a sphere artifact S is calculated by

$$I(x, y) = f^{-1}(v(x, y)) \cdot \exp\left(-\int_S \mu(x, y, z)dz\right), \tag{3.1}$$

where $\int_S \mu(x, y, z)dz$ stands for the integration along Z axis inside the sphere. Hence, the pixel value of a generated abnormal image is computed by

$$f(I(x, y)) = \exp(-\int_S \mu(x, y, z)dz) \cdot v(x, y) + (1 - \exp(-\int_S \mu(x, y, z)dz)) \cdot 255. \tag{3.2}$$

Furthermore, in order to calculate the integration in the above expression efficiently, assume that μ is uniform and constant inside the sphere. Then it is expressed as

$$\int_S \mu(x, y, z)dz = \mu \cdot l(x, y), \tag{3.3}$$

where (3.3) $l(x, y)$ is the length of the line $(x, y, \cdot)$ inside the sphere. $\cdot$ means that the value of z is arbitrary. Specifically, when the distance from the center of the sphere to the line is r, and the radius of the sphere is R, the following equation holds:

$$l(x, y) = \begin{cases} 2\sqrt{R^2 - r^2} & (0 \leq r \leq R), \\ 0 & \text{(otherwise)}. \end{cases} \tag{3.4}$$

Therefore, the above expression (3.2) can be expressed simply as

$$f(I(x, y)) = \exp(-\mu \cdot l(x, y)) \cdot v(x, y) + (1 - \exp(-\mu \cdot l(x, y))) \cdot 255, \tag{3.5}$$

which is able to calculate once the hyperprameters μ and R are given. In our experiments, several values about these hyperprameters are tried and results are compared. Details are discussed in subsequent chapters.

3.3 Another shape — Ellipsoid

In the above section, we introduced the sphere shape of an artifact, which looks like a circle in 2D images. Although you can check that only this scheme will work well and make a difference, we have a motivation to introduce more shape varieties since actual pneumonia is not only a circle-shape.

As we mentioned already, inflammation basically spreads isotropically in a lung. However, it sometimes spreads anisotropically because of some reasons

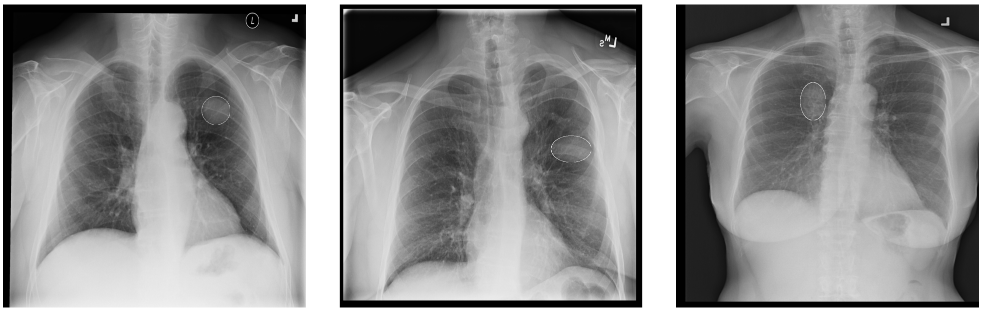

Figure 3.1: Examples of generated abnormal images. The left image has a spherical lesion, and the middle and right images have different shapes of ellipsoidal lesions respectively. Just for clarity, the lesion boundaries are emphasized by white lines.

such as a barrier of ribs, which makes the shape distorted. Thus, we introduce an ellipsoid as another form of lesions. Generally, an ellipsoid is expressed as

$$\frac{(x-x_0)^2}{a^2}+\frac{(y-y_0)^2}{b^2}+\frac{(z-z_0)^2}{c^2}=1, \tag{3.6}$$

where (x_0, y_0, z_0) is the center point of the ellipsoid. Here, we want to know the length of the line $(x, y, \cdot)$ inside the ellipsoid. By the expression (3.7), we can have the length by

$$l(x,y)=\begin{cases}2c\sqrt{1-\frac{(x-x_0)^2}{a^2}-\frac{(y-y_0)^2}{b^2}} & (\frac{(x-x_0)^2}{a^2}+\frac{(y-y_0)^2}{b^2}\leq 1),\\ 0 & \text{(otherwise)}.\end{cases} \tag{3.7}$$

We note that (3.8) is the generalized expression from (3.5). Actually, when $a = b = c = R$ is assumed, (3.8) is equal to (3.5) by using $r=\sqrt{(x-x_0)^2+(y-y_0)^2}$. Therefore, using (3.8) instead of (3.5) makes it possible to generate lesions with a more variety of shapes, and the validation accuracy is expected to be improved. Actually, we got the expected results in our experiments. See subsequent chapters.

One can see and compare the images generated by these approaches in Figure (3.3).

3.4 The heuristic algorithm to detect lungs

The artifacts generated by the proposed method should be superimposed on lungs in the negative images, so that the generated images look like realistic positive medical images. However, it is not so obviously easy to recognize the area of lungs in an image by using a certain algorithm. Instead, we applied the alternative strategy that are based on the knowledge that the area of lungs are almost full of air and the pixel values there tend to be lower than other parts in the image. In more detail, we employed the following algorithm: (i) once a normal image is given, a point (x, y) is randomly chosen from the area designated by the hyperparameter M. (ii) calculate the average pixel value around the chosen point (x, y). (iii) If the value is lower than a threshold, decide to use the point (x, y) as the center of the artifact. Otherwise, restart from (i).

A practical problem here is the difficulty of deciding the threshold value. When the value is too high, the artifact is more likely to be put outside of lungs and the image looks unnatural. On the other hand, when the value is set too

Algorithm 1 Lung Detection

```
Input: $(H, M, T, N, k, l, u)$
$a, b, c \sim U(l, u)$
$n \leftarrow 0$
while True do
    $x, y \sim U(M, H - M)$
    $v \leftarrow 0$
    for $i \in [x - a, x + a]$ do
        for $j \in [y - a, y + a]$ do
            $v \leftarrow v+$ PixelValues$(i, j)$
        end for
    end for
    if $v \leq 2a \cdot 2b \cdot T$ then
        return$(x, y)$
    else
        $n \leftarrow n + 1$
        if $N \leq n$ then
            $T \leftarrow T + k$
            $n \leftarrow 0$
        end if
    end if
end while
```

low, an infinite loop can happen in some cases because of (iii). Considering the situation that some X-ray images look brighter and others look darker, and the average pixel value differs from image to image, it is not wise and unnecessary to fix the value. For this reason, we applied the adaptive algorithm where the threshold value is changed dynamically if the loop (iii) continues for a large number of times. Namely, we firstly set the threshold T, basically as low, and if trials of (iii) continued more than a hyperparameter N, the threshold value was incremented so that the loop could be avoided. See the algorithm 1 for details.

In the algorithm 1, H means the height and width of the image. In our case, for instance, the size of each picture is 1024 × 1024, so H is set to 1024. M represents the margins that should be removed for lung detection. N and T are the thresholds explained above. k is the value that is added to T. l and u is respectively the lower bound and the upper bound of the artifact' size that was already mentioned in the previous chapter. M, T, N, k, l, u, and the attenuation coefficient μ that was also mentioned in the previous chapter are the hyperparameters in our experiments.

The values of these hyperparameters can be decided naturally by using domain and common knowledge. So, firstly, we did the quick experiments in which several values of the hyperparameteres are set based on some reasoning beforhand. Through the procedure, we can see the effectiveness of the proposed model and approach quickly.

Furthermore, by using searching algorithm such as Bayesian optimization, it is possible to find the optimal values of these parameters and the optimal data augmentation method in the proposed model. Through this procedure, we also did the experiments to find the best data augmentation method and collect some high-quality positive data that were generated by the augmentation method.

3.5 The whole architecture for extrapolative data augmentation

Let us rethink the motivation and situation again when there are only a small number of positive data. In the situation, we want to increase the amount and variety of positive data. However, when the number of positive data is too small to cover the whole positive data domain, we think that the model to generate positive data should be independent from the existing positive data. The reason is that we want to increase the variation of positive data by generating those that have not yet appeared in and are rather different from the existing positive data. That is why we call it extrapolative data augmentation. Furthermore, in order to generate data that belong to outer domain of the existing data, we think that humans' annotations and assistance are necessary, important and useful.

From such a background, we propose the human-in-the-loop architecture for extrapolative augmentation of positive data. The details of this architecture are as follows. Firstly, define the model that generate positive images from negative images. In our case, we established the model by using the formulations and heuristic algorithms based on domain knowledge as already mentioned. Secondly, by using searching algorithms such as Bayesian optimization, find the optimal hyperparameters of the model that achieves the highest validation score. In our experiments, we used Area-Under-Curve (AUC) of Receiver-Operating-Characteristic-curve (ROC) as the score, because the validation dataset is imbalanced. Note that this is the first time for the real positive images to be utilized, under the presumption that the number of the positive images is not enough for training the model directly but enough for deciding some hyperparameters of the model. Thirdly, with radiologists and experts, make the careful selections of realistic positive images generated by the best data augmentation. Compared with other augmentations, it is expected that a larger number of images generated by the optimal data augmentation seize the feature of and look like real positive images. The selected data, good pseudo positive data, play the role of extractively augmented data. Furthermore, by adding the data to the existing positive data, you can restart this cycle again, find another optimal data augmentation, and get other good pseudo positive data in the same way.

The excellent point of this human-in-the-loop architecture is that it is comprised of the components that are independent from each other. Therefore, each component are replaceable. For instance, if you find another good model for generating positive data, you can use the model instead of the proposed one. Not only that, but also, once good pseudo positive data are gathered enough, generative models such as GAN or VAE can be used effectively in that part. In this way, the whole architecture can be customized as you like.

Chapter 4

Experiments

In this chapter, we empirically investigate the performance of the proposed methods. First, by using domain and common knowledge, we fixed the hyperparameters of the proposed model that generate positive images from negative images. Our results show the effectiveness of the model. Next, we searched the best hyperparameters by means of Bayesian optimization and collected highly qualified pseudo positive images. We also confirmed the effectiveness of the best augmentation method by investigating the performance on the test dataset. Furthermore, we added the selected images to the positive data in the valid set, and turned the second lap in the whole architecture. In this case, both the valid and test AUC were much improved than the first lap, which might indicate the success of the extrapolative data augmentation.

4.1 Dataset

We firstly used the Chest X-ray dataset from National Institutes of Health (NIH) [14], which is open dataset and publicly available to use. However, the disease labels in the dataset are given from radiological reports automatically by means of natural language processing, which sometimes causes the labels to be inaccurate. Therefore, we decided to use the data which is labeled by human experts in order to see the performance of our methods correctly. Actually, thanks to recent enormous efforts [10], about 30000 data in the dataset are relabeled by skilled radiologists from the Radiological Society of North America and Society of Thoracic Radiology. More specifically, each chest radiograph contains pneumonia label, age, view position, and other information such as gender and color mode. The pneumonia label includes " Normal ", " Lung Opacity " or " No Lung Opacity / Not Normal ", where " Lung Opacity " means pneumonia in this context.

Furthermore, in order to see the effectiveness of the proposed methods in a limited number of data, we randomly selected the 2000 normal data and 10 abnormal data from the above 30000 subdataset. As a definition, the normal data necessarily satisfies the following three conditions: (i) label = " Normal " (ii) view position = " PA " (iii) age ≥ 18. Similarly, the abnormal data meet the above condition (ii), (iii), and another condition (iv) label = " Lung Opacity " . We separated the 2000 normal data into 1000 and 1000 data, where the former data were used for generating abnormal images and for training, while the latter data and the selected 10 abnormal data were used for validation. The reason why the validation data consists of 1000 normal images and 10 abnormal images is that we are thinking about the scarce disease such that the disease rate is under 1 percent. Note that the size of each image is 1024 × 1024 pixels.

4.2 Implementations

By using the proposed model, we generated 1000 images from the same number of existing 1000 normal images. We considered the generated images as abnormal images and made the training set that consists of the 1000 normal images and the same number of the generated abnormal images. Every image in both training set and validation set was downsampled to a size of 224 × 224 pixels, because that is the expected input size for ResNet34, a classifier that predicts whether the input is normal or abnormal. After that, we trained the classifier on the training set and investigated the generalization performance on the validation set. We saw the validation performance every time the training had finished in one epoch.

For training, Adam was used as an optimizer, and the learning rate was set to 0.0005. Binary-Cross-Entropy was used as a loss function, and the batch size was 16, so each epoch was 125 iterations. We trained the discriminator about 15 epochs, since the classification accuracy on the training set generally plateaus around this point.

On the other hand, when validation, we used Area-Under-Curve (AUC) of Receiver-Operating-Characteristic-curve (ROC) as the evaluation criteria. The reason is that the validation dataset consists of 1000 normal images and 10 abnormal images, and in such an imbalanced situation, AUC is widely used to calculate the performance correctly and objectively. In more details, the trained classifier not only predicts whether an input is normal or abnormal, but also outputs the degree of confidence that the input is abnormal. Based on the scores and true labels of the all 1010 images in the validation set, AUC is calculated.

4.3 Experiments

We did two kinds of experiments. Firstly, we fixed the hyperparameters of the proposed model that generate positive images. To some extent, the values can be set reasonably, based on the domain knowledge such as the tumors' sizes and the common sense such as whether generated abnormal images look natural and realistic. In this way, we set the values of the hyperparameters without validation data, and investigated the general performances of the proposed methods by testing on the validation data.

Secondly, we used the Bayesian optimization to find the best hyperparameters such that the classifier trained by the best data augmentation method outputs the heighest validation AUC. In this process, the validation data was used to find the best values, so we investigated the general performance of the best data augmentation method by testing on the test data. In our experiments, 25 candidates were searched by the Bayesian optimization in total. The search space was as follows: $\mu \in [0.0001, 0.0050], M \in [90, 400], N \in [1, 100], T \in [0, 200], k \in [1, 20], l \in [10, 50], u \in [50, 90]$, where μ represents the attenuation coefficient of the artifacts, M is the margin that is ignored when deciding the center point of the artifacts, N and T are the thresholds and k is the value that are used in the Algorithm 1 mentioned in the previous chapter. l is the lower bound and u is the upper bound of the artifact' size respectively. Namely, a, b, c , the size of the ellipsoid introduced in the expression (3.6), follow the uniform distribution $\mathrm{U}(l, u)$ independently. The unit of M, l, u is a pixel, and each value of the three was decided based on the input's size, i.e., 1024 × 1024 pixels in our case.

Furthermore, Among the images generated by the best data augmentation method in the 1st lap, 30 images were randomly selected. By a radiologist, 6 images out of 30 were considered as highly qualified pseudo abnormal images.

These collected abnormal data were effectively used in the next lap. Namely, the selected images were added to the abnormal data in the valid set in order to increase the amount and variety of abnormal data. If the abnormal data in the validation data are augmented effectively, it is expected that we will have more chances to find better parameters that show better performances on the test set. Therefore, we searched turned the second lap in the whole architecture, i,e., tried to find better hyperparameters again. By testing on the test set, we investigated the performance of the best data augmentation method in the 2nd lap, which showed better result.

4.4 Results and Discussion

Table 4.1: The validation AUC when the hyperparameters' values are varied

hyperparameters	validation AUC
$\mu = 0.0025, M = 200, N = 30, T = 100, k = 10, l = 30, u = 70$	0.708
$\mu = 0.0015, M = 200, N = 30, T = 100, k = 10, l = 30, u = 70$	0.736
$\mu = 0.0010, M = 200, N = 30, T = 100, k = 10, l = 30, u = 70$	0.803

Table 4.2: The validation and test AUCs of the best data augmentation method in the 1st and 2nd lap. The first one consists of $\mu = 0.00195, M = 364, N = 78, T = 5, k = 6, l = 36, u = 57$. and the second one consists of $\mu = 0.00199, M = 269, N = 42, T = 9, k = 11, l = 31, u = 87$

lap	validation AUC	test AUC
1st lap	0.880	0.847
2nd lap	0.901	0.940

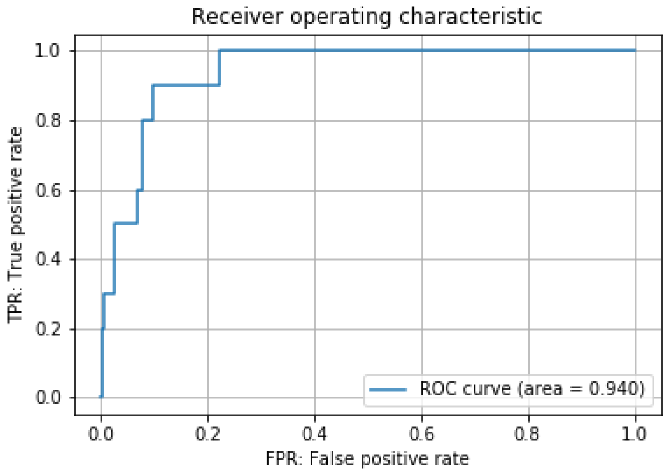

Figure 4.1: The ROC curve of the test set that consists of 1000 normal and 10 abnormal images. AUC = 0.940

Table 5.1 shows the validation AUCs compared by different values of the hyperparameters. In the first three trials in the Table 5.1, the hyperparameters except μ are the same, because the values can be decided naturally. For example, the common size of pneumonia stays in the range between 30 pixels and 70 pixels, and T is not so relevant unless the value is high. The value of μ is difficult to decide only from input images, so three values are tried for comparison.

Among the three trials, the validation AUC is the highest when $\mu = 0.0010$. The reason may be that generated images in this case tend to look more natural and realistic than the other two cases, in terms of the intensity of the generated artifacts. Actually, when μ=0.0025 or 0.0015, some small artifacts look too white and unnatural, which are generated when a and b are small, and c is large. Hence, thourh this approach, the value of the hyperparameter μ can also be set reasonably. Table 5.1 show that, with the reasonable values of hyperparameters and the proposed model for generating abnormal images, validation AUC 0.803 was achieved. This result suggests the effectiveness of the proposed model that enables to generate abnormal images only from normal images.

Table 5.2 shows the highest validation AUC and when the Bayesian optimization is used to find the best values of the hyperparameters, and its test AUC. In the 1st lap, among the 25 times of searching, 0.880 was the best score of AUC. Since the validation set was used for the searching process, we also investigated the performance of the best classifier on the test set, and its AUC was 0.847. From this result, it can be said that better hyperparameters and better generative model were found by using the searching algorithm effectively. Furthermore, In the 2nd lap, the validation AUC was 0.901 and the test AUC was 0.940, and much improved than the 1st lap. The figure 1 shows its ROC curve, showing the marvelous result. The reason why the results are much improved might be derived by the effective augmentation of abnormal data by using the 1st lap, i.e., adding highly qualified pseudo abnormal data. That might imply the success of the proposed architecture, extrapolative data augmentation. By repeating the same loop again and again, it is expected that the more highly qualified pseudo abnormal data will be collected, which is really valuable when there are not enough data.

Chapter 5

Conclusions

In this thesis we proposed a novel approach for generating pneumonia images by using domain knowledge and normal lung images on Chest X-ray. The approach does not rely on any actual pneumonia images and thus is useful when there are not sufficient number of abnormal medical images. Moreover, hyperaparameters can be set reasonably by domain knowledge, and even without hyperparameters' search, the approach is proven to work well with AUC achieving 0.803.

Furthermore, we also proposed a novel human-in-the-loop architecture that enables to augment abnormal images in an extrapolative way. The procedures are as follows. First, find the optimal hyperparameters' values by using searching algorithms such as Bayesian optimization, and select high-quality pseudo abnormal images generated by the optimal model. Next, add the selected images to the existing scarce abnormal images that are used as a validation, and start again from the beginning.

If the selected images effectively increase not only the amount but also the variety and diversity of the scarce abnormal images, the AUC in the second time is expected to be higher than the first time in the human-in-the-loop architecture. Through an empirical investigation, AUC achieved 0.940 in the second time, much superior to the first time with AUC 0.847. This result suggests that the selected, highly qualified pseudo abnormal images might successfully augmented the scarce abnormal images in an extrapolative manner.

In addition, each module is independent and thus replaceable in the human-in-the-loop architecture. For example, the proposed model, approach for generating abnormal images, can be replaced as one like. Actually, it seems to be effective to change the model sometimes a little bit to increase more varieties of abnormal images. Furthermore, one can apply the architecture to another scarce disease just changing the model part properly. In the near future, we will tackle the abnormal data augmentation of more scarce diseases by using the proposed methods here in this thesis.

References

[1] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018.

[2] Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. Imagenet: A large-scale hierarchical image database. volume 00, pages 248–255, 06 2009.

[3] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *CoRR*, abs/1803.01229, 2018.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[6] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *CoRR*, abs/1905.00397, 2019.

[7] Daisuke Sato M.D., Shouhei Hanaoka M.D., Yukihiro Nomura, Tomomi Takenaga, Soichiro Miki M.D., Takeharu Yoshikawa M.D., Naoto Hayashi M.D., and Osamu Abe M.D. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head CT volumes. In Nicholas Petrick and Kensaku Mori, editors, *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 388 – 393. International Society for Optics and Photonics, SPIE, 2018.

[8] Mary M. Moya and Don R. Hush. Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9:463–474, 1996.

[9] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2352–2360. Curran Associates, Inc., 2016.

[10] George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan

Vummidi, Kavitha Yaddanapudi, and Anouk Stein. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[11] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. *CoRR*, abs/1711.10284, 2017.

[12] Franck Vidal and Pierre-Frédéric Villard. Development and Validation of Real-time Simulation of X-ray Imaging with Respiratory Motion. *Computerized Medical Imaging and Graphics*, 49:15, April 2016.

[13] Franck P. Vidal, Manuel Garnier, Nicolas Freud, Jean Michel Létang, and Nigel W. John. Simulation of X-ray Attenuation on the GPU. In Wen Tang and John Collomosse, editors, *Theory and Practice of Computer Graphics*. The Eurographics Association, 2009.

[14] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315, 2017.

[15] Qi Wei, Yinhao Ren, Rui Hou, Bibo Shi, Joseph Y. Lo, and Lawrence Carin. Anomaly detection for medical images based on a one-class classification. In Nicholas Petrick and Kensaku Mori, editors, *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 375 – 380. International Society for Optics and Photonics, SPIE, 2018.

[16] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017.