ABSTRACT

Semi-supervised Learning for Electrocardiography Signal Classification

Dedong Zhang, M.S.

Mentor: Liang Dong. Ph.D.

An electrocardiogram (ECG) is a cardiology test that provides information about the structure and function of the heart. The size of the ECG data collected from patients can be very large, and the data analysis is tedious. Inspired by human learning, in this thesis we propose a new semi-supervised training framework for deep neural network to classify ECG data. The idea is to reward the valid associations that belong to the same class after a round trip during cross-matching of supervised and unsupervised learning, while penalizing the incorrect associations. The implementation of our framework can be easily integrated with any existing training setup. With data preprocessing, the detection of heart disease is improved.

Semi-supervised Learning for Electrocardiography Signal Classification

by

Dedong Zhang, B.S.

A Thesis

Approved by the Department of Electrical and Computer Engineering

_____

Kwang Y. Lee, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of

Master of Science in Electrical and Computer Engineering

Approved by the Thesis Committee

_____

Liang Dong, Ph.D., Chairperson

_____

Keith Evan Schubert, Ph.D.

_____

Carolyn Skurla, Ph.D.

Accepted by the Graduate School

May 2018

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES viii

ACKNOWLEDGMENTS ix

DEDICATION x

CHAPTER ONE 1
Introduction 1
Project aims and contribution 1
Thesis outline 2

CHAPTER TWO 3
Background 3
ECG data distribution 5
Semi-supervised training 8
Unsupervised training 9

CHAPTER THREE 11
Machine Learning 11
MNIST dataset 13
Convolutional Neural Network 13
Learning by association 16

CHAPTER FOUR 20
Experimental results 20
Experimental setup 20
Datasets 20
Splitting the data 21
Feature extraction 21
Results 23

CHAPTER FIVE 33
Summary 33

APPENDIX 34
MIT-BIH beat type records 35

BIBLIOGRAPHY 40

LIST OF FIGURES

Figure 2.1 One period of an ECG signal ……4

Figure 2.2 Distribution of ECG in different data spaces…….6

Figure 2.3 Comparison of real ECG data…….....……...8

Figure 3.1 Semi-supervised learning.…….……....…….11

Figure 3.2 Partial MNIST database……...11

Figure 3.3 The full connected NN and locally connected NN are compared…….....13

Figure 3.4 The locally connected NN and convolutional net are compared……..…14

Figure 3.5 LeNet-5 network structure……15

Figure 4.1 Data split for both ECG database and MNIST database……..22

Figure 4.2 Confusion matrix with mistakes that were made…….26

Figure 4.3 The patient 118 raw data from the MIT-BIH dataset, red point shows the found R-wave position……...……...27

Figure 4.4 Figure on the top shows the heartbeat wave of patient 118 before wavelet transform. And the following figure shows the same patient after prepossessing……28

Figure 4.5 Learning rate for semi-supervised learning algorithm training ECG labeled and unlabeled signal from MIT-BIH Database…….……...……29

Figure 4.6 The output layer's corresponding handwriting data distribution visualizing...31

Figure 4.7 The output layer's corresponding ECG data distribution visualizing.....32

Figure 4.8. Using the supervised learning method to classify the MIT-BIH database without extraction preprocessing……..……...33

Figure 4.9 Generating Receiver Operating Characteristic(ROC) Curves for supervised learning method to classify the MIT-BIH database without extraction preprocessing……...34

Figure A.1 Schematic diagram of semi-supervised training network……...……39

Figure A.2 Schematic diagram of supervised learning and unsupervised learning in side of Figure A.1……...40

LIST OF TABLES

Table 4.1 Summary of MIT-BIH dataset for different arrhythmia classes……...18

Table 4.2 Summary of handwriting data in the MNIST dataset…….…….....19

Table 4.3 Error (%) results on the test set of MNIST (lower is better) ...…….24

Table 4.4 Results on MIT-BIH. Accuracy (%) on the test set (higher is better)……...24

Table A.1 Table of beat types for MIT-BIH database...……30

Table A.2 Symbols used in MIT-BIH database plots……32

ACKNOWLEDGMENTS

I owe my deepest gratitude to my advisor Dr. Liang Dong for his support and advice. Without his help, this thesis would be impossible. I also would like to express my sincere appreciation to those who have guided and supported me to reach this point in my life: Thanks to Dr. Carolyn Skurla and Dr. Keith Schubert for their invaluable guidance in the field of biomedical. Thank Philip Haeusser for providing me with an idea of how to cross-matching supervised and unsupervised learning. Thank Yuan Xing and Yuchen Qian for advising on preconditioning optimization. Thank Dr. Jie Tian from the Chinese Academy of Sciences Institute of Automation for inspiring me to this research direction. Finally, I would like to thank the School of Engineering and Computer Science of Baylor University, as well as Baylor Research and Innovation Collaborative for letting me work on this project using their research facilities.

DEDICATION

To Dr. Yongsheng Zhang a good father a rigorous and knowledgeable professor

CHAPTER ONE

Introduction

Project Aims and Contribution

Machine learning has been increasingly used in various real-world applications. In such scenarios, however, labeled data is not always available due to cost and other constraints. For example, in medical classification area, due to the need of patients' privacy protection, medical institutions and clinics often limit the disclosure of patient information. This is a big limitation for medical detection applying to training deep neural networks. Even with sensitive information taken off, without of the professional physician's label, the data do not actually improve the performance of supervised learning. To address this problem, semi-supervised learning leverages a large volume of unlabled data and a small set of labled data. The aim of this thesis is to investigate the feasibility of using partially labeled ECG signals and the improved semi-supervised learning model to achieve the goal of ECG classification. Inspired by humans' association learning, in this thesis we propose a new semi-supervised training framework for deep neural network to classify ECG data. "Association" is resulted from embedding labeled data to unlabeled one and back. The proposed method rewards the valid associations that belong to the same class after a round trip, while penalizing the incorrect associations. We demonstrate our proposed framework on MNIST dataset and MIT-BIH dataset, showing that the semi-supervised learning can significantly improve the performance of ECG classification by using the unlabeled data. In particular, for ECG

database with fewer labeled data, our approach can achieve the same performance of training with more labeled data.

Thesis Outline

This thesis has the following structure:

Chapter one briefly introduces the basic concepts of the project and describes the contribution of the project.

Chapter two discusses the background of the ECG classification, including ECG signal distribution, the principle of ECG feature extraction, as well as classification. It also discusses the state-of-the-art for semi-supervised learning and unsupervised learning.

Chapter three introduces image feature extraction and convolutional neural network principles. Meanwhile, an optimized semi-supervised learning model is introduced and mathematical formulas are derived accordingly.

Chapter four describes a practical implementation of semi-supervised learning in handwriting recognition and ECG signal classification. Experimental results obtained on several data sets and feature collection are also shown.

Chapter five concludes the thesis and points out future research directions.

CHAPTER TWO

Background

Electrocardiogram or ECG is a tool used to visualize the electricity that flow though the heart. The electrodes placed on the skin detect the tiny electrical changes that arise from the heart muscle's electrophysiologic pattern of depolarizing during each heartbeat. The electrodes are placed in different parts of the human body and connected to the positive and negative electrodes of the electrocardiograph ammeter by lead wires. This method of recording the electrocardiogram circuit is called electrocardiogram lead. Different electrode placements and connection methods can be used to form different leads. In the long-term practice of clinical electrocardiography, the standard 12-lead system became the wildly used international standard. It is a very commonly performed cardiology test. ECG can be used to measure the rate and rhythm of the heartbeats, the size and location of the ventricles, the presence of any damage to cardiac muscle cells or the conduction system, the effects of cardiac drugs, and the function of implanted pacemakers. The ECG signal is usually composed of the following waveforms: P, QRS and T, as shown in Figure 2.1.

In recent years, computer-assisted ECG analysis has made great progress [1]. Some existing ECG devices have widely used these methods, include locating the main wave (R wave) of ECG, and the identification of the start and end positions for each wave. If the ECG waveform features can be accurately extracted, a valid judgment can be made according to the rules of disease classification. However, even for the best R wave

extraction method so far [2], the method was only tested on the standard database MIT-BIH [3]. Their performance degrades quite a lot on the actual clinical data set. For P and T wave extraction, the accuracy of wave extraction for MIT-BIH is not stable, not to mention the raw clinical data.

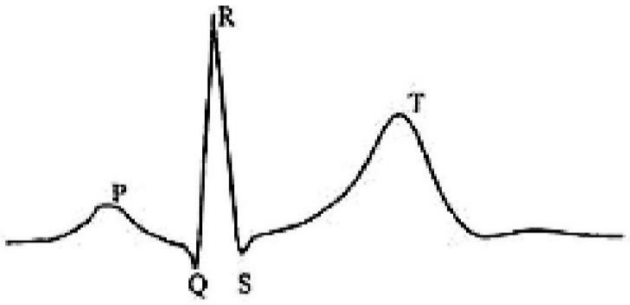

Figure 2.1 One period of an ECG signal

For this reason, various mathematical features such as wavelet features [4], higher-order statistics [5], power spectral features [6], Lyapunov coefficients [7], Hermit coefficients [8], Shannon entropy [9], Hermite polynomial coefficients [10] and linear prediction error features [11] were proposed. Combined with methods such as time-domain features, independent components analysis (ICA), waveform features, principal component analysis (PCA), and linear discriminant analysis (LDA), they are used in ECG classification and recognition. It is undeniable that these methods are valid for the given dataset, but according to the published literature, many of them are just conclusions drawn from a standard dataset or from a subset of them (even the entire MIT-BIH dataset only contains 47 patients). Thus the generalization of the algorithm is difficult to guarantee. There is also an important conclusion of the heart rate classification: the positive abnormal recognition rate in patients (ECG samples with both training and test

sets all contain the same individual) is much higher than the inter-patient (there is no intersection of individuals in the ECG samples of the training set and the test set), and the recognition rate of a certain disease should be between positive intra-patient and patient's abnormal recognition rate [12].

ECG Data Distribution

In order to simulate the thinking process of the doctor when diagnosing, we should start with the classification of a certain disease. First, when the doctors determine between a positive anomaly and healthy, in fact, they also consistently match patient data with the disease template. If there is no match, in most case the doctor will consider the patient has healthy ECG signal. For a particular disease, medical experts have a set of diagnostic criteria (except for incurable diseases), such as ECG "right bundle branch block" (Figure 2.3) the diagnostic rules:

(1) P-wave appears in front of each QRS wave;

(2) PR interval: 0.12 ~ 0.20s;

(3) QRS wave: > 0.12s;

(4) V1 or V2 lead is rSR', RSR' and Rsr';

(5) Dull S-wave in V5, V6, and I lead.

Assuming that x is the ECG data of a randomly picked patient, various medical features $f_m(x)$ are obtained through the feature extraction function, i.e., the indicators in the diagnostic rule. Obviously, the ECG data points of the same disease are clustered in the medical feature space, which satisfies the clustering assumptions in pattern recognition. The left graph in Figure 2.2 shows roughly the distribution of ECG data in

the medical feature space. The black dots are abnormal ECGs, and the blank areas are normal ECGs. Notice that, here is only a schematic of the relative distribution of various types of ECG data. Individual data points could be staggered together (corresponding to rare complex disease).

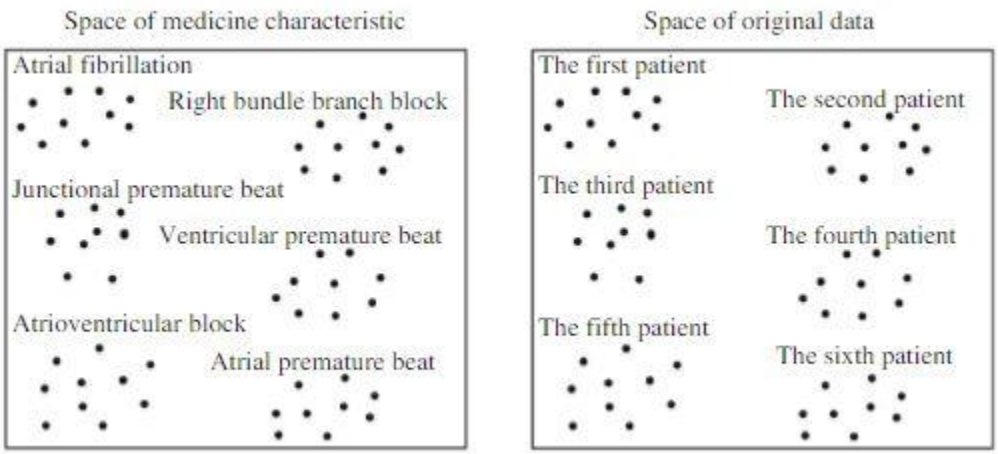

Figure 2.2 Distribution of ECG in different data spaces.

Therefore, it can be seen that the medical feature space does not need a very complex classifier to meet the requirements (e.g., Support Vector Machine (SVM) can handle this classification task well), provided that $f_m(x)$ can not only handle all kinds of noise interference, but also cope with the data diversity due to physiological differences between different people. Obviously, $f_m(x)$ is a highly complex nonlinear function. The key to successful design an automated ECG analysis algorithm to meet the clinical application requirements positive anomalies between patients, diagnosis of specific diseases is the construction of the medical feature extraction function $f_m(x)$.

In the following, we analyze the distribution of ECG data based on positive anomalies in patients. In order to obtain ECG heartbeat samples from MIT-BIH or other databases, most existing work took the approach to find the location of R wave using the main wave extraction algorithm. Then they took R as the center and pushed forward and backward to form a heartbeat sample. After that, they divided the dataset randomly or by other means. Wubbeler et al. [13] took the peak of R wave as the center and extracted the waveform data of 100ms before and after the center, respectively. They then calculated the Euclidean distance of each feature and combined with the threshold method. We can achieve a highly efficient identification algorithm, 74 individual ECG data in the PTB database [3] as the experimental sample, and the recognition accuracy is up to 99%. Therefore, the same person's ECG raw data tend to belong to a same cluster. The right part of Figure 2.2 shows roughly the distribution of the ECG data in the raw data space. The ECG data from the same person gather together, and the data of different people disperse in different regions. The two sides of Figure 2.2 are very similar. The only difference is that one is in the medical characteristics space, while the other is in the original data space. Therefore, putting an inside patient classification algorithm to the clinical practice, even a very efficient one, the classification performance will be significantly reduced. As for the classification of patients in the medical feature space, only one category needs to be considered. The positive anomaly classification should consider multiple categories, many of which are interlacing rare complex disease in the area. To some extent, the classification of diseases among patients is easier than the classification of positive anomalies.

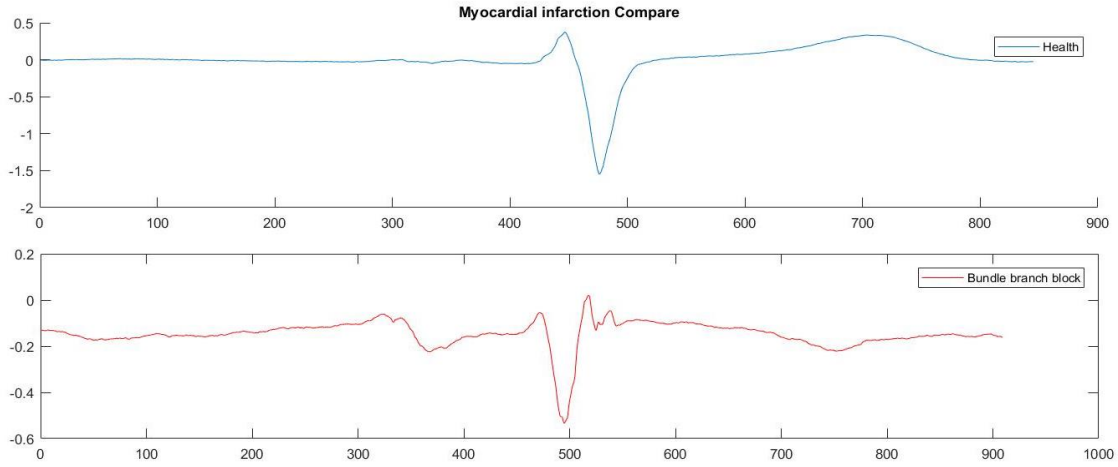

Figure 2.3 Comparison of real ECG data. The above figure is one lead V1 heartbeat of healthy person, and the other figure is one heartbeat of Bundle Branch block patient. Both X axis in the plot represents the time in millisecond. And Both Y axis represents the voltage in mV.[3]

Semi-supervised Training

The semi-supervised learning has become more and more popular in recent years. It falls between supervised learning and unsupervised learning. By incorporating a small set of labeled data into the large volume of unlabeled data, the semi-supervised training can improve the learning accuracy significantly. In neural network research, the semi-supervised training has been used in SVM [14] to detect image boundaries, where a set of labeled samples were used as extra regulators.

One approach that can considerably improve the prediction accuracy of neural networks is to bootstrap the model with labeled samples. Such samples can be available from the model itself. For example, [22] used the class of data that has maximum predicted probability as the labels for unlabeled data. Then both set of data are trained at the same time. The paper also adopted noise cancellation and other techniques to achieve better performance on MNIST.

Another way to improve the accuracy is to apply an auto-encoder to the existing network ([27] [37] [39]), which leads to more efficient representations. M. Sajjadi et al. [30] also studied the mutual exclusivity loss for semi-supervised deep learning. In their work, an extra regulating term was used for the unlabeled data. [31] discussed the problem of regularization with stochastic transformations and perturbations for deep semi-supervised learning.

Unsupervised Training

Unsupervised training is to infer the inner structure by using unlabeled data. In unsupervised training, there is no evaluation on the accuracy of the output of certain algorithms since no labels are available. Though unsupervised training often finds more applications than its counterpart, it is necessary to differentiate their purposes. Semi-supervised training leverages a small set of labels as the guidance for learning, while the evaluation of unsupervised learning algorithms mainly depends on the choice of cost function. The authors of [12] proposed to use ECG morphology and heartbeat interval features for automatic classification of heartbeats. The basic idea is to apply Restricted Boltzmann Machine [33] to do pre-training on a network. [11] proposed a two-stage method for ECG classification using Gaussian mixture model. [19] built high-level features using large scale unsupervised learning. An auto-encoder was used as a regulator. Wubbeler et al. [13] used the electrocardiogram to verify humans, where clustering was used. Ubeyli E D [7] evaluated the diagnostic accuracy of the recurrent neural networks (RNN) on the ECG signals. The RNN has composite features such as

wavelet coefficients and Lyapunov exponents. In [6], the relative position of two randomly picked samples in the image is predicted.

CHAPTER THREE

Machine Learning

Due to the nature of the ECG data distribution, the clinical application of ECG classification algorithm should be patient-oriented. This is challenging because of the following reasons. First, part of the lead off and some of the QRS complexes are not obvious. Second, noise interference is ubiquitous, and the target test set is open. Third, due to the lack of ECG data, the decision function is a highly complex nonlinear function, which needs to be fitted in the end. The solution of those challenges is to use the deep learning network. Martinez et al. [17], for the first time, applied deep learning to the classification and identification of physiological signals, achieving only 70% -75% accuracy. Due to the limited ECG data resources, that is not an advantageous result for ECG data classification. On the other hand, the application of handwritten recognition of deep learning has reached a general accuracy of more than 98%. So, to verify the reliability of the algorithm we proposed, it is important to verify using MNIST datasets with large data sources. Our key insight is that data from the same class will share much similarity. By properly embedding labeled and unlabeled data, better performance can be achieved for a Convolutional neural network (CNN). Two vectors A and B are formed respectively from labeled and unlabeled raw data (A_{RAW} and B_{RAW}) that fed into the CNN. The optimization method we proposed is to do cross matching between supervised and unsupervised learning CNN. Imagine someone walks from A to B and then comes back. The walk is based on the mutual similarity between A and B. It is only valid if the

walker turns out belong to the same class as he is at the very beginning. This can be shown in Figure 3.1.

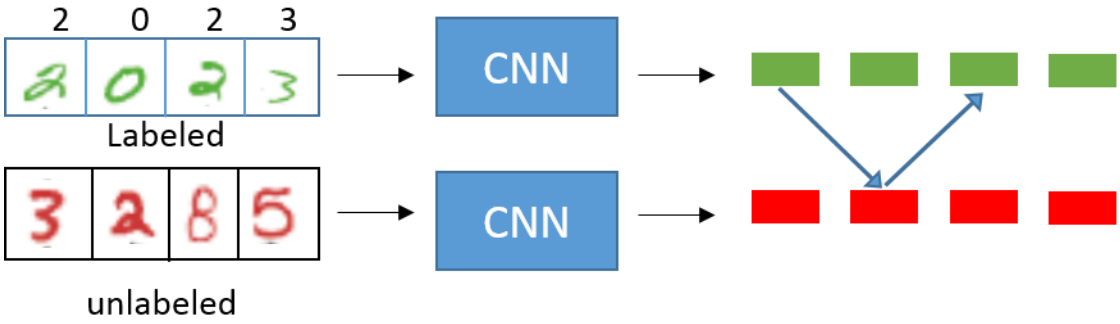

Figure 3.1 Semi-supervised learning. Labeled and unlabeled data are fed into the CNN to produce embeddings (blue). A red arrow goes from labeled data to unlabeled data and back is called an association cycle.

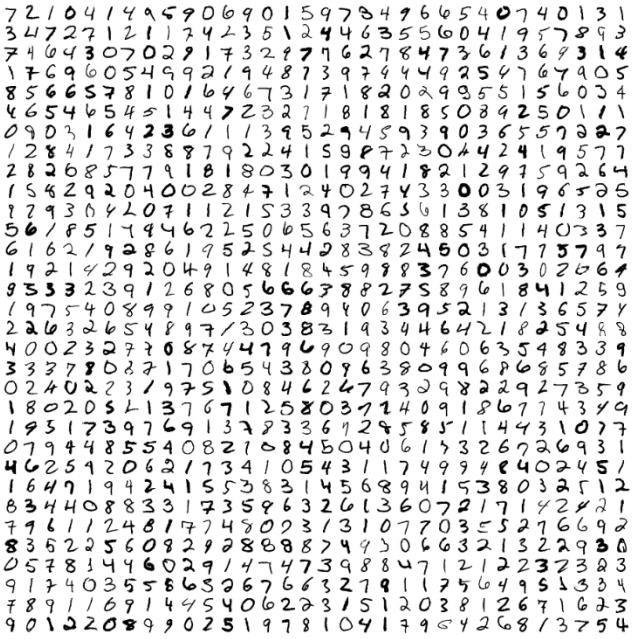

Figure 3.2 Partial MNIST database.[34]

MNIST Dataset

The MNIST dataset is a large modified handwritten database from the National Institute of Standards and Technology (NIST). The training set consists of numbers written by 250 different people, 50% of whom are high school students and 50% from the staff of the Census Bureau. The test set is also the same proportion as the handwritten training set. Figure 3.2 shows what the MNIST data look like. There are 60,000 training samples and 10,000 test samples. Each is an anti-aliasing grayscale with a 28 x 28 pixel size and a normalized 20 x 20 size digital portion, centered on the image, maintaining its original shape. Notice that the MNIST database also maintains the correspondence between handwritten numbers and identities. This provides a reference for us to choose the appropriate number of people and the appropriate amount of data.

Convolutional Neural Network

The Convolutional Neural Network (CNN) has been widely used in deep learning and been very successful in image processing. CNN is the foundation for many successful models that use international standard ImageNet dataset to train. Compared to the traditional image processing algorithms, CNN does not require complex pre-processing of the image. The original image can be directly used without feature extracting. For an ordinary one-dimensional signal, CNN does not seem to be a valid training model. It is because of the high reproducibility of ECG signals, and the fact that the shapes of different ECG classes or ECG annotation information are relatively similar, which makes ECG classification different from other one-dimensional signal

classification. Because CNN can extract deeper features through multiple layers, the differences between different ECG signal can be easier to discover.

In image processing, the image is often viewed as one or more 2D vectors. For example, the previously mentioned MNIST handwritten image can be viewed as 28 × 28 2D vectors, which only has one-color channel (black and white). If the RGB color picture has three color channels, it can be represented as three 2D vectors. The traditional neural networks often introduce a huge number of parameters thus computation overhead since the input and the next hidden layer are fully connected. By contrast, the CNN is locally connected so that the computation overhead can be significantly reduced.

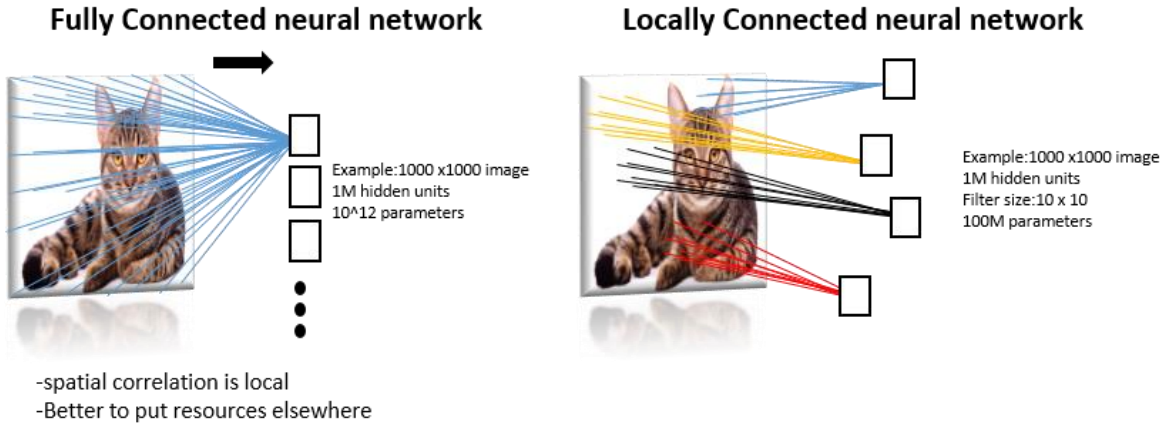

Figure 3.3 The full connected neural network and locally connected neural network are compared, with the left-hand side showing the fully connected schematic and the right-hand locally connected.

For example in Figure 3.3 shows a 1000 × 1000 input image, if the number of neurons in the next hidden layer is 10^6, there are $1000 \times 1000 \times 10^6 = 10^{12}$ weight parameters for the full connection, which is a huge number. These

parameters are difficult to train. In contrast, using local connections, each neuron in the hidden layer is only connected to a 10 × 10 partial images in the image, so the number of weight parameters at this time is $10 \times 10 \times 10^6 = 10^8$, which reduces 4 orders of magnitude. Although the number of parameters is significantly reduced compared to the full connected NN, it is still very high. Using the method of weight sharing can further reduce the number of parameters. Specifically, in the local connection, each neuron of the hidden layer is connected to a 10 × 10 partial image. Therefore, there are 10 × 10 weight parameters, and they are shared to the rest. That means the weight parameters of 10^6 neurons in the hidden layer are the same. No matter what the number of hidden layer neurons is, the number of weight parameters to be trained is 10 x 10, which is called the size of the convolution kernel or size of the filter.

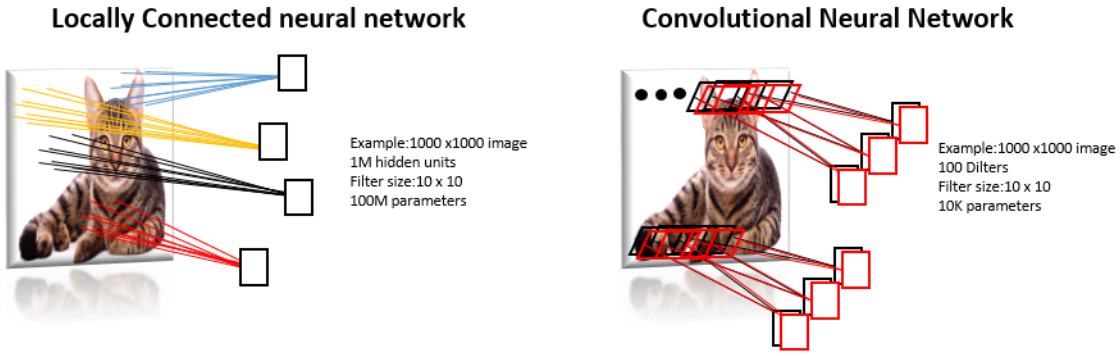

Figure 3.4 The locally connected NN and convolutional net are compared, with the left-hand side showing the locally connected schematic and the right-hand convolutional net schematic.

In this way, only one feature of the image is extracted. If more features are to be extracted, multiple convolution kernels can be added. Different convolution kernels can obtain features of different mappings of the image, which are called Feature Maps. As the

Figure 3.4 shows, if there are 100 convolution kernels, the number of weight parameters is only $100 \times 100 = 10^4$. In addition, the offset parameters are also shared, sharing the same filter. Figure 3.5 is a classic CNN LeNet-5 network. As the figure shows, there are mainly two types of network layers in the CNN: the convolution layer and the pooling/sampling layer (i.e., Pooling). The role of the convolution layer is to extract various features of the image, while the role of the pool layer is to abstract the original feature signal, thereby greatly reducing the training parameters as well as model overfitting.

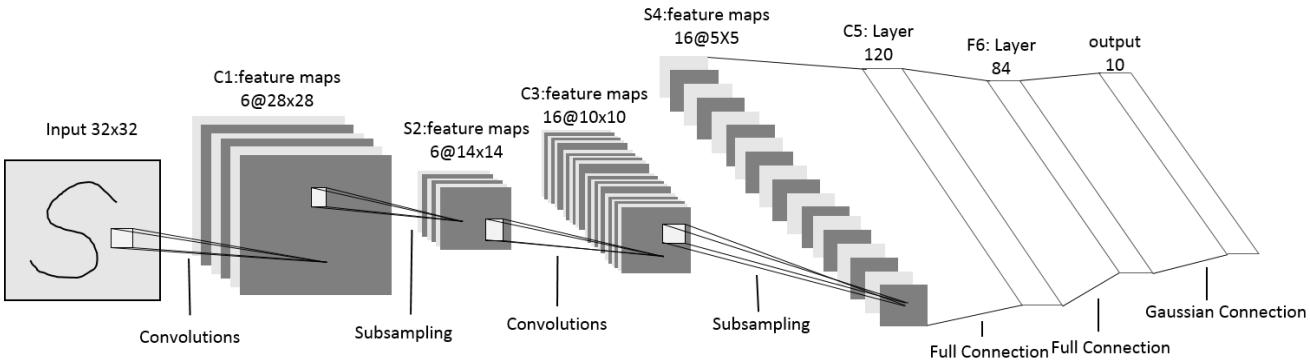

Figure 3.5 LeNet-5 network structure.

Learning by Association

The logic behind our semi-supervised training is to maximize the probability that after a round-trip between A and B, the walker remains in the same annotation class, where rows in matrices A and B index the samples in the data. The similarity between embeddings A and B is defined as

$$M_{ij} := A_i \cdot B_j$$

where dot product is used. Note, however, any metrics other than dot product can be applied, e.g., Euclidean distance. The reason why we pick dot product is because we found it has the best convergence performance by experiments. The similarities are then transformed into the transition probabilities from A to B by the following operation:

$$P_{ij}^{ab} = P(B_j | A_i) := (\text{softmax}_{\text{cols}}(M))_{ij}$$

$$= \exp(M_{ij}) / \sum_{j'} \exp(M_{ij'})$$

Similarly, by replacing M with M^T, we can obtain the transition probability from B to A, i.e., P^{ba}. Combing them together, the probability of the round-trip between A and B can be calculated as

$$P_{ij}^{aba} := \left(P^{ab}\, P^{ba}\right)_{ij}$$

$$= \sum_k \left(P_{ik}^{ab} P_{kj}^{ba}\right)$$

Thus, we have the following probability:

$$P(\text{correct walk}) = \frac{1}{|A|} \sum_{i \sim j} P_{ij}^{aba}$$

$$\text{where } i \sim j \Leftrightarrow \text{class}(A_i) = \text{class}(A_j)$$

Now define L_{walker}, L_{visit}, and $L_{\text{classification}}$ as the loss of walk, visit, and classification respectively. Then the total loss is:

$$L_{\text{total}} = L_{\text{walker}} + L_{\text{visit}} + L_{\text{classification}}$$

Walker loss. First, we define walker loss aiming to keep consistency for association cycles. A walk is said to be consistent if it falls into the same class after the round trip. The intuition is to reward the probability that walks to the correct class while penalize wrong walks. The reason why we use a uniform probability instead of requiring

an exact sample is because the walk may end up at a different sample after the round trip, but it still belongs to the same class with the start. In this case, using a probability makes more sense. We further define the loss as follows:

$$L_{\text{walker}} := \text{H}(\text{T}, P^{aba})$$

where H is cross-entropy, T is the uniform target distribution of correct round trip, and P^{aba} is the probability of the trip respectively. The target uniform distribution is as follows:

$$T_{ij} := \begin{cases} 1 \,/\, \text{class}(\text{A}_i), & class(\text{A}_i) = class(\text{A}_j) \\ 0, & else \end{cases}$$

where class(A_i) denotes how many times class(A_i) happens in A.

Visit loss. It is always beneficial to use all of the unlabeled data in semi-supervised learning, instead of just associating samples that are easily available. By incorporating all data, the embedding can be more general. However, some unlabeled samples tend to be hard to use due to various reasons such as bad drawing. Thus, we define a loss function for "visiting" such samples as a cross-entropy H:

$$L_{\text{visit}} := \text{H}(\text{V}, P^{visit})$$

where V is the uniform target distribution and P^{visit} are the visit probabilities. The visit probabilities for samples in B and the uniform target distribution are:

$$P_j^{visit} := \langle \text{P}_{ij}^{ab} \rangle_i$$

$$V_j := 1 \,/\, |\text{B}|$$

Classification loss. We have successfully solved the problem of creating embedding so far. Then next question is how to map the embedding into the classes. By

carefully study, we found that we can address this if we add a fully connected layer with softmax and a cross-entropy loss on top of the network. Note that such mapping is not necessary for network to be able to converge. However, it is critical for network performance evaluation. We define the classification loss as $L_{\text{classification}}$. Using the method proposed by Adam [16], the total loss can be minimized. We also applied random data augmentation, which is discussed in Chapter Four. We implemented the semi-supervised training using Google TensorFlow library.

CHAPTER FOUR

Experimental Results

Experimental Setup

Datasets

Two datasets were used in the process of experimental validation of the algorithms introduced in this thesis. Table 4.1 shows the summary of MIT-BIH dataset and Table 4.2 shows the summary of MNIST dataset.

Table 4.1 Summary of MIT-BIH dataset for different arrhythmia annotation classes.

heartbeat classes	#Data point	Symbol
Normal beat	15,065,352	N
Left bundle branch block beat	1,621,266	L
Right bundle branch block beat	1,456,446	R
Atrial premature beat	511,143	A
Aberrated atrial premature beat	30,150	a
Nodal (junctional) premature beat	16,683	J
Supraventricular premature beat	402	S
Premature ventricular contraction	1,432,125	V
Fusion of ventricular and normal beat	161,202	F
Atrial escape beat	3,216	e
Nodal (junctional) escape beat	46,029	j
Ventricular escape beat	21,306	E
Fusion of paced and normal beat	197,382	f
Non-conducted P-wave (blocked APB)	38,793	x
Unclassifiable beat	6,633	Q

Table 4.2 Summary of handwriting data in the MNIST dataset.

Name	Description	#Example	Size
train-images-idx3-ubyte	Training set images	60,000	9,681KB
train-labels-idx1-ubyte	Training set labels	60,000	29KB
t10k-images-idx3-ubyte	Test set images	10,000	1611KB
t10k-labels-idx1-ubyte	Test set labels	10,000	5KB

Splitting the data

Every data set was split into four parts (Figure 4.1)

- Data to train the supervised learning
- Data to train the unsupervised learning
- Data to tune the parameters
- Data to test the learning result

we ran the nets on 10 randomly chosen subsets of the data and report median and standard deviation.

Figure 4.1. Data split for both ECG database and MNIST database.

Feature Extraction

For handwritten images, feature extraction of images is conducted in convolutional neural networks. Because CNN is created by simulating the process of

visual recognition of images, its network structure is specially designed for 2D images. To deal with 1D ECG signals, it is necessary to optimize the convolutional neural network to adapt the characteristics of the electrocardiogram signal. In addition, because there are too many useless fragments in the signal of ECG itself, it is necessary to perform feature extraction on the raw data before training in order to obtain better performance and reduce training time. The design of the ECG classification algorithm is actually to fit a nonlinear decision function F(x) = g(f(x)), where f(x) is the function fitted by the feature extraction algorithm. For example, it can be a medical or mathematical feature extraction function, or a hybrid feature extraction function. In this experiment, we use the R-wave position extraction and wavelet transform method to determine the valuable information in the ECG signal. Wavelet transform has achieved very good results in many areas. In this thesis, we evaluate its performance in terms of feature extraction, which can be expressed as:

$$WT_f(a,\tau) = \frac{1}{\sqrt{a}} \int_0^R f(t)\psi\left(\frac{t-\psi}{a}\right)$$

where $\psi_{a,\tau}(t)$ is a wavelet basis function, with two parameters of scale a and translation τ. We used the multi-scale characteristics of the basis function to expand the signal at different scales and extract useful information.

The process of the experiment consists of three steps. The first step is to test the effectiveness of the semi-supervised training method through the MNIST dataset. At the same time, a large amount of training data shows the characteristics of the mixed training compared to other methods. The second step is to optimize the ECG data by feature

extraction to avoid partial lead off, unobvious partial QRS complex, and noise interference. The last step is to train the ECG data from MIT-BIH database after feature extraction and analyze the results.

Results

A number of experiments were executed using this setup:

1. In order to test the learning performance for ECG signal annotation classification. We use 117 heartbeat sample from MIT-BIH database train in a pure supervised learning algorithm. The totally accuracy is 83.3%. Figure 4.8 shows the Confusion matrix of the output accuracy result. And Figure 4.9 shows the Generating Receiver Operating Characteristic (ROC) Curves of the output accuracy result.
2. Use semi-supervised learning algorithm to train handwriting image from MNIST Database. Table 4.3 shows the result of 10,000 label data training on supervised side before cross-matching of supervised and unsupervised learning. Figure 4.2 shows the Confusion matrix of MNIST data after 20,000 steps of training and cross-matching. The test error is 0.97%. Figure 4.6 is the 2D distribution of the output layer's corresponding handwriting data
3. Find the R-wave of the raw ECG signal and use wavelet transform to extract feature information while reducing the impact of interference information on training. Figure 4.3 shows one patient's raw data and the R

peak found on the heartbeat wave. Figure 4.4 shows the same patient's data change after preprocessing.

4. Using semi-supervised learning algorithm train, ECG labeled and the unlabeled signal from MIT-BIH Database. Figure 4.5 shows the learning rate of the training process. Table 4.4 shows the result of accuracy for different annotation classes. Figure 4.7 is the 2D distribution of the output layer's corresponding ECG data.

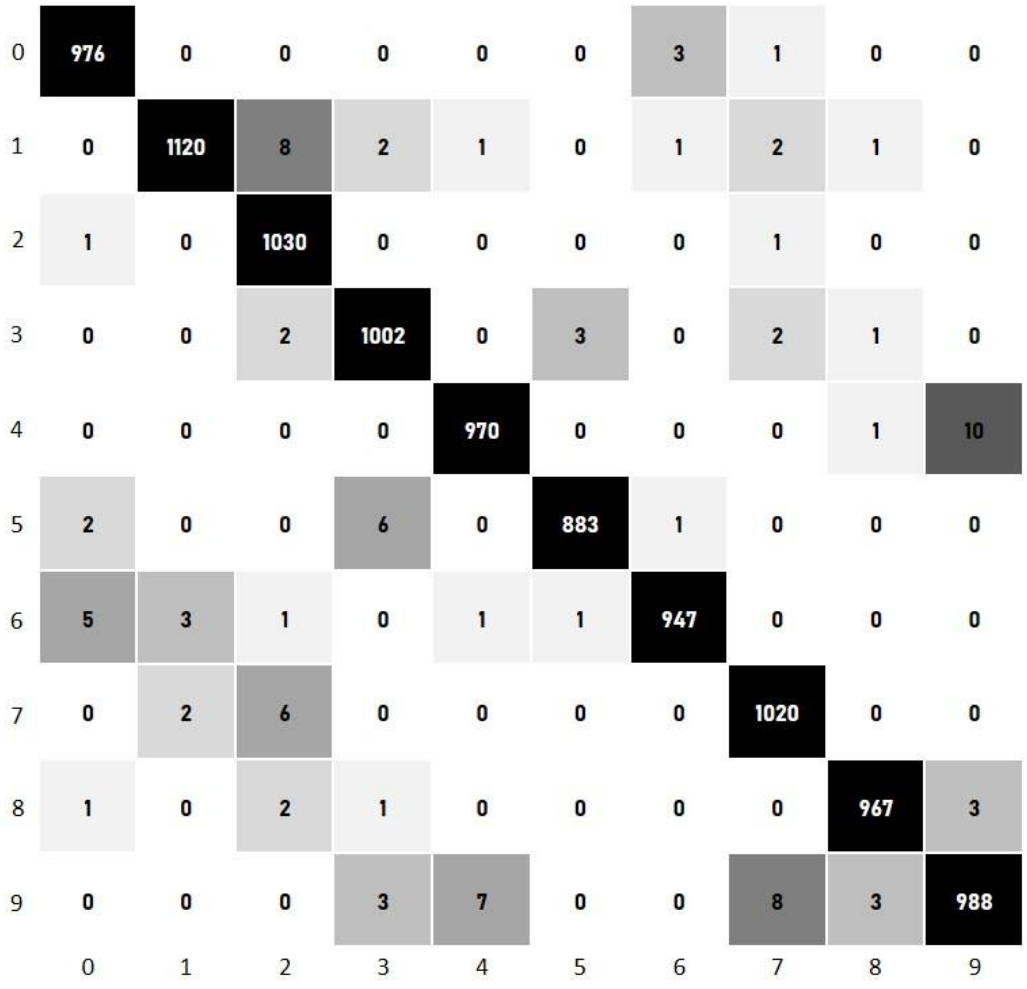

Figure 4.2 Confusion matrix with mistakes that were made. The x-axis is the true-label results. Y-axis is the results of the output from the learning structure.

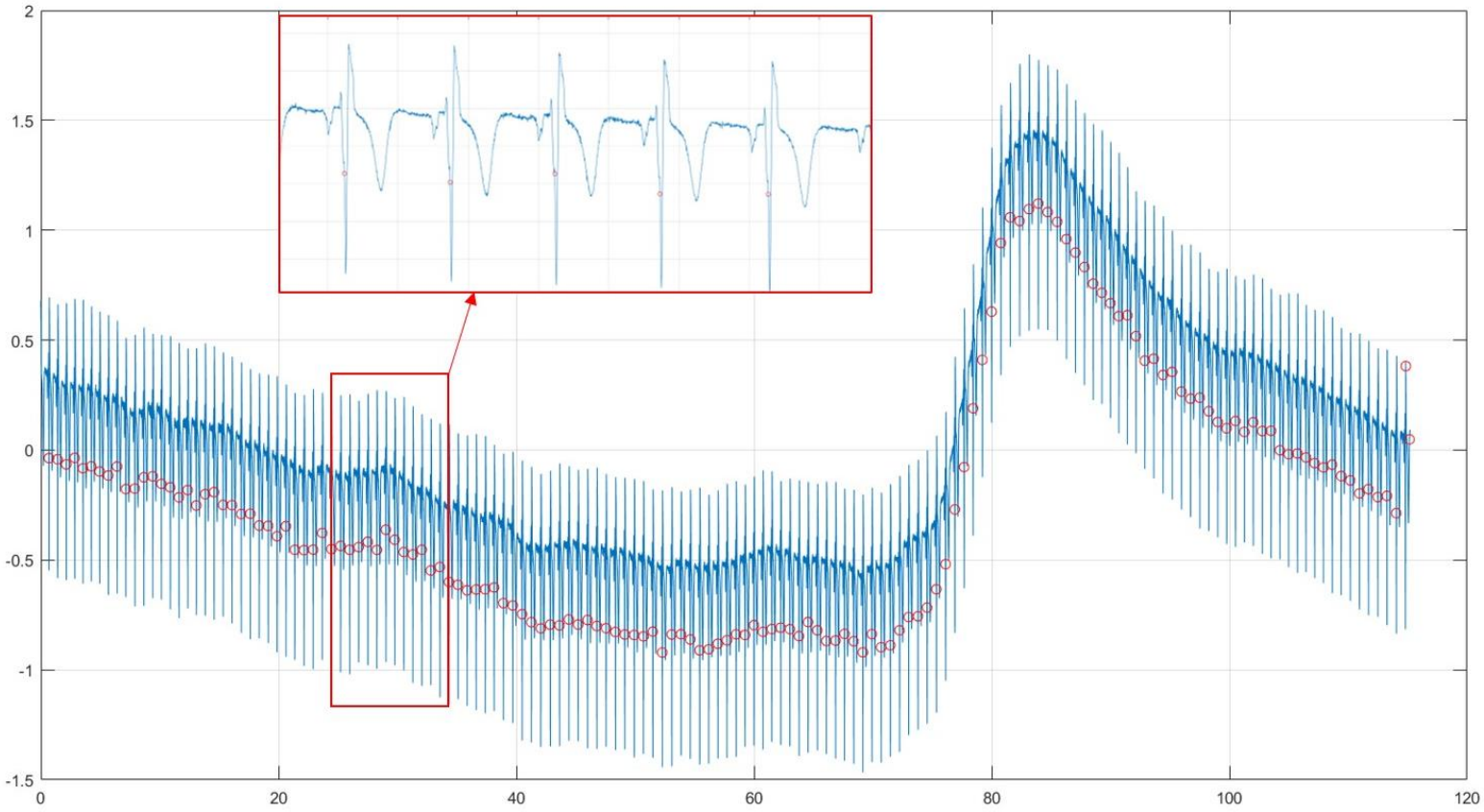

Figure 4.3 The patient 118 lead V1 raw data from the MIT-BIH dataset, red point shows the found R-wave position. X axis in the plot represents the time in second, Y axis represents the voltage in mV.

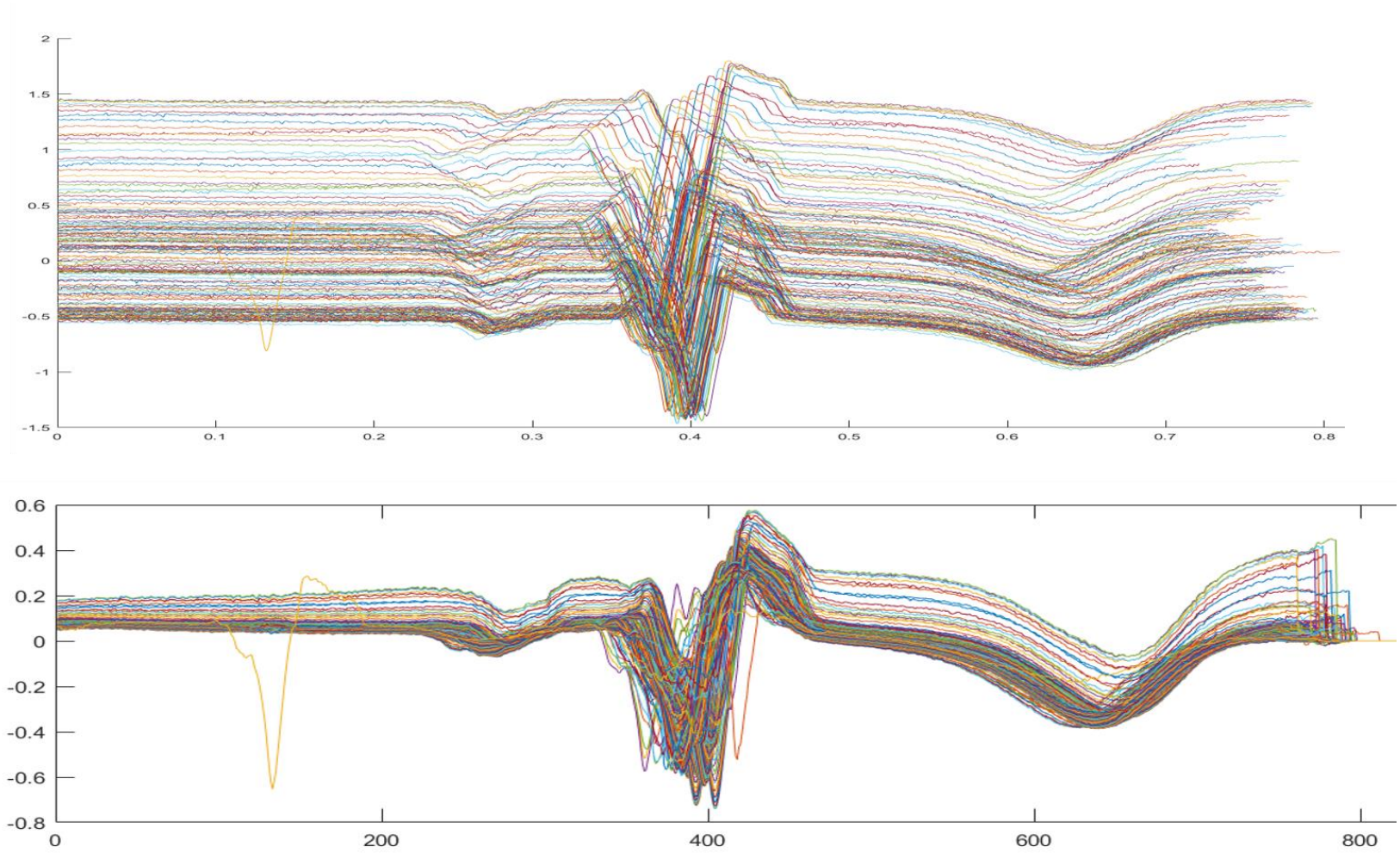

Figure 4.4 Figure on the top shows the heartbeat wave of patient 118 lead V1 ECG signalbefore wavelet transform. And the following figure shows the same patient after prepossessing. The X axis in the top plot represents the time in second and the X axis in the bottom plot represents the time in millisecond. Both Y axis represents the voltage in mV.

Figure 4.5 Learning rate for semi-supervised learning algorithm training ECG labeled and unlabeled signal from MIT-BIH Database. X axis represents the steps of the learning process. Y axis represents the learning rate of the algorithm.

Table 4.3 Error (%) results on the test set of MNIST (lower is better). The standard deviation in parentheses.

	# labeled samples (%)		
Method	100	1000	ALL
Ladder, conv small Γ	0.98(0.50)	-	-
Improved GAN	0.93(0.07)	-	-
Mutual Exclusivity + Transform	0.55(0.16)	-	0.27(0.02)
Ours	0.91(0.09)	0.74(0.03)	0.36(0.03)

Table 4.4 Results on MIT-BIH. Accuracy (%) on the test set (higher is better). The standard deviation in parentheses.

			Classification result (%)				
Method	#Supervised training data	#Unsupervised training data	N	S	V	F	Q
Kiranyaz &Gabbouj	419,377	0	0.98	0.64	0.93	0.76	0
our	138,690	416,070	0.92 (0.1)	0.47 (0.13)	0.88 (0.1)	0.87 (0.19)	0.28 (0.21)

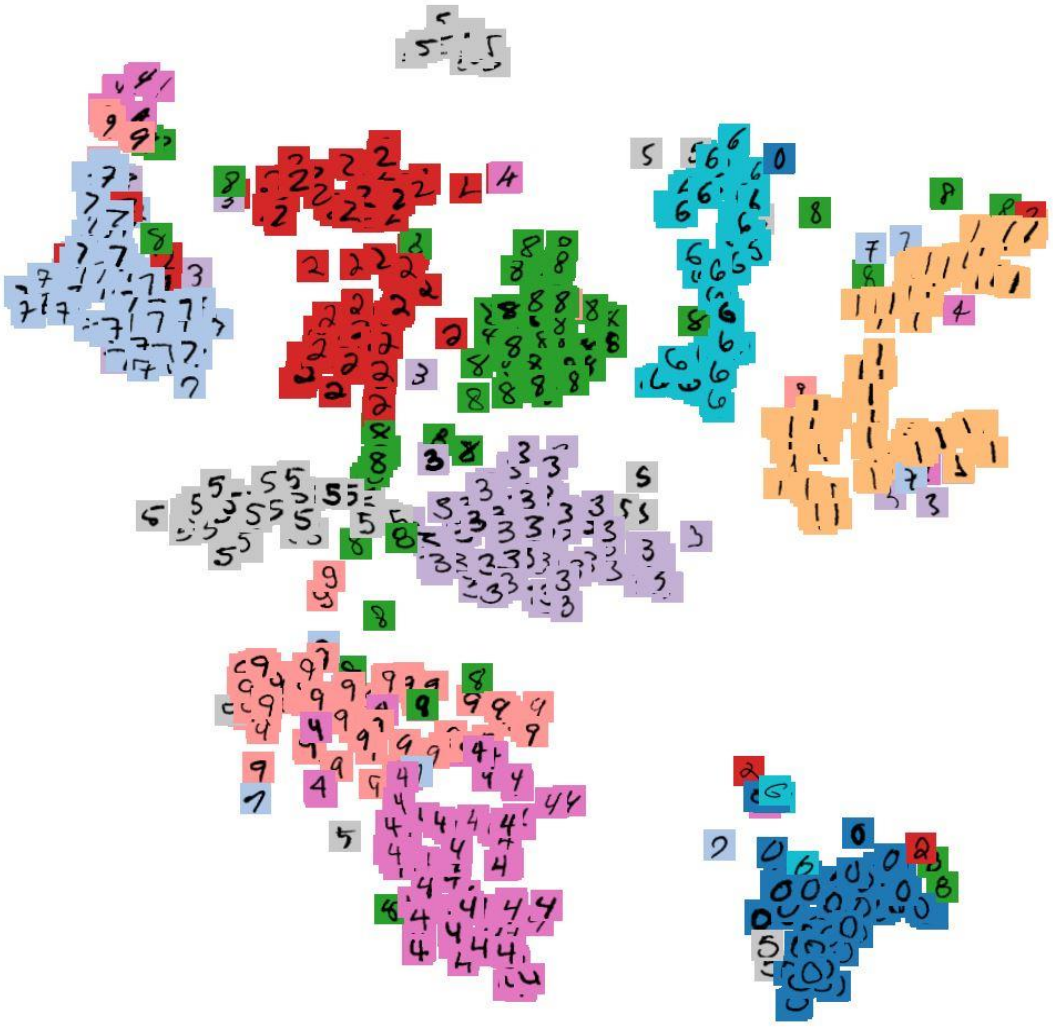

Figure 4.6 The output layer's corresponding handwriting data distribution visualizing. Note that handwriting images of different colors represent different classes sample points.

Figure 4.7 The output layer's corresponding heartbeats distribution visualizing. Note that each number represents one annotation heartbeat data from MIT-BIH data set.

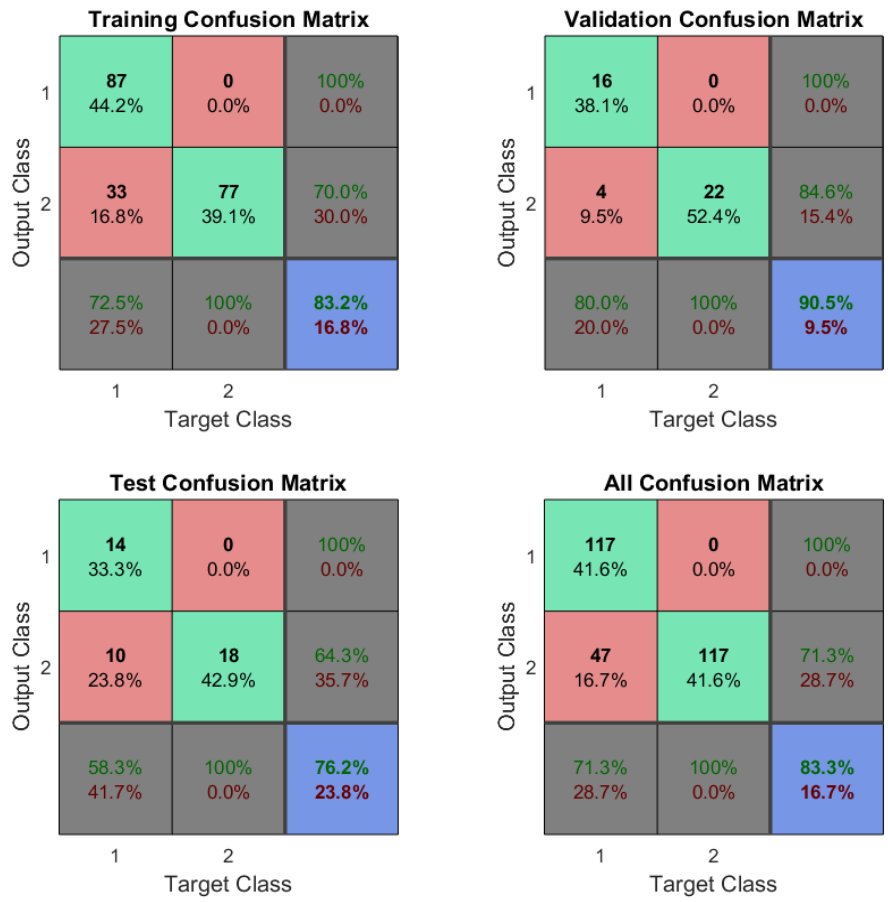

Figure 4.8 Using the supervised learning method to classify the MIT-BIH database without extraction preprocessing. label 1 for normal annotations, label 2 for abnormal annotations. The x-axis is the true-label results. Y-axis is the results of the output from the learning structure.

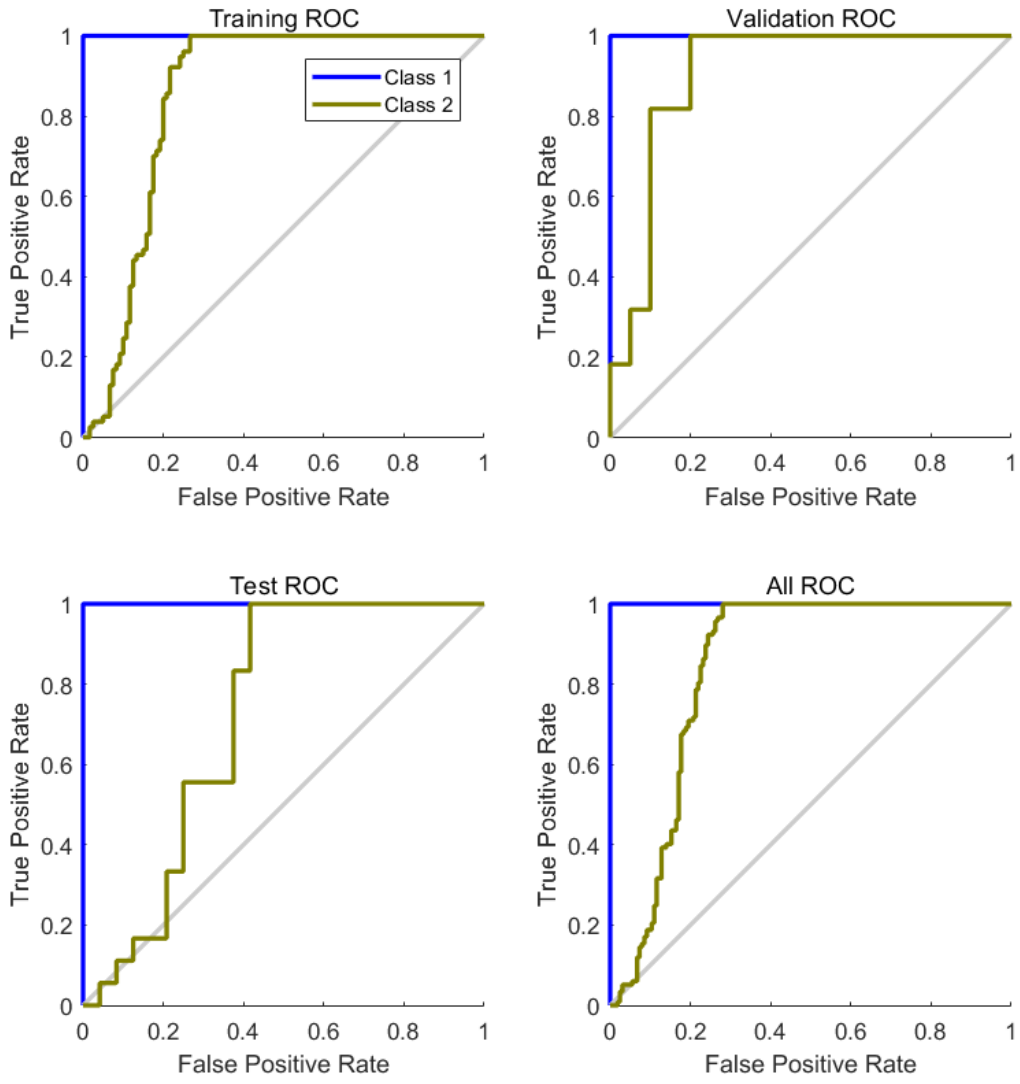

Figure 4.9 Generating Receiver Operating Characteristic(ROC) Curves for supervised learning method to classify the MIT-BIH database without extraction preprocessing. Each point on the ROC curve reflects the sensitivity to the same signal stimulus.

CHAPTER FIVE

Summary

In this thesis, we propose a semi-supervised learning model for ECG signal classification. The model can be easily applied to both 1D and 2D signal training. According to the distribution characteristics of the ECG data, the feature extraction method avoids exacting useless information by using preprocessing. Although the effect of a small amount of ECG training data is not as obvious as the data collection, there are some important conclusions that can be drawn. First, with a small amount of label data, it is easier to use our proposed optimization to achieve higher accuracy as good as pure supervised training. Second, because this training method is based on a common convolutional neural network, it has good adaptability for either one-dimensional time series signals or two-dimensional images.

Some future research directions in this area include: (1) Extend this method of monitoring networks and non-supervised networks to other models, such as Long short-term memory neural network (LSTM). (2) Continue to optimize convolutional neural networks. One possible direction could be update the CNN to Capsule Network to make more specific feature comparisons.

APPENDIX

MIT-BIH beat type records

Table A.1 Table of beat types for MIT-BIH database.

Record	**Annotation**																
	N							**V**	**F**	**O**	**N**		**E**	**P**	**F**	**O**	**Q**
	.	**L**	**R**	**A**	**a**	**J**	**S**	**V**	**F**	**!**	**e**	**j**	**E**	**P**	**f**	**p**	**Q**
100	2239	-	-	33	-	-	-	1	-	-	-	-	-	-	-	-	-
101	1860	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	2
102	99	-	-	-	-	-	-	4	-	-	-	-	-	2028	56	-	-
103	2082	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-
104	163	-	-	-	-	-	-	2	-	-	-	-	-	1380	666	-	18
105	2526	-	-	-	-	-	-	41	-	-	-	-	-	-	-	-	5
106	1507	-	-	-	-	-	-	520	-	-	-	-	-	-	-	-	-
107	-	-	-	-	-	-	-	59	-	-	-	-	-	2078	-	-	-
108	1739	-	-	4	-	-	-	17	2	-	-	1	-	-	-	11	-
109	-	2492	-	-	-	-	-	38	2	-	-	-	-	-	-	-	-
111	-	2123	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
112	2537	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-
113	1789	-	-	-	6	-	-	-	-	-	-	-	-	-	-	-	-
114	1820	-	-	10	-	2	-	43	4	-	-	-	-	-	-	-	-
115	1953	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
116	2302	-	-	1	-	-	-	109	-	-	-	-	-	-	-	-	-
117	1534	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-
118	-	-	2166	96	-	-	-	16	-	-	-	-	-	-	-	10	-

Record	.	L	R	A	a	J	S	V	F	!	e	j	E	P	f	p	Q
119	1543	-	-	-	-	-	-	444	-	-	-	-	-	-	-	-	-
121	1861	-	-	1	-	-	-	1	-	-	-	-	-	-	-	-	-
122	2476	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
123	1515	-	-	-	-	-	-	3	-	-	-	-	-	-	-	-	-
124	-	-	1531	2	-	29	-	47	5	-	-	5	-	-	-	-	-
200	1743	-	-	30	-	-	-	826	2	-	-	-	-	-	-	-	-
201	1625	-	-	30	97	1	-	198	2	-	-	10	-	-	-	37	-
202	2061	-	-	36	19	-	-	19	1	-	-	-	-	-	-	-	-
203	2529	-	-	-	2	-	-	444	1	-	-	-	-	-	-	-	4
205	2571	-	-	3	-	-	-	71	11	-	-	-	-	-	-	-	-
207	-	1457	86	107	-	-	-	105	-	472	-	-	105	-	-	-	-
208	1586	-	-	-	-	-	2	992	373	-	-	-	-	-	-	-	2
209	2621	-	-	383	-	-	-	1	-	-	-	-	-	-	-	-	-
210	2423	-	-	-	22	-	-	194	10	-	-	-	1	-	-	-	-
212	923	-	1825	-	-	-	-	-	-	-	-	-	-	-	-	-	-
213	2641	-	-	25	3	-	-	220	362	-	-	-	-	-	-	-	-
214	-	2003	-	-	-	-	-	256	1	-	-	-	-	-	-	-	2
215	3195	-	-	3	-	-	-	164	1	-	-	-	-	-	-	-	-
217	244	-	-	-	-	-	-	162	-	-	-	-	-	1542	260	-	-
219	2082	-	-	7	-	-	-	64	1	-	-	-	-	-	-	133	-
220	1954	-	-	94	-	-	-	-	-	-	-	-	-	-	-	-	-
221	2031	-	-	-	-	-	-	396	-	-	-	-	-	-	-	-	-
222	2062	-	-	208	-	1	-	-	-	-	-	212	-	-	-	-	-
223	2029	-	-	72	1	-	-	473	14	-	16	-	-	-	-	-	-
228	1688	-	-	3	-	-	-	362	-	-	-	-	-	-	-	-	-
230	2255	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
231	314	-	1254	1	-	-	-	2	-	-	-	-	-	-	-	2	-
232	-	-	397	1382	-	-	-	-	-	-	-	1	-	-	-	-	-
233	2230	-	-	7	-	-	-	831	11	-	-	-	-	-	-	-	-
234	2700	-	-	-	-	50	-	3	-	-	-	-	-	-	-	-	-

Table A.2 Symbols used in MIT-BIH database plots.

Symbol	*Meaning*
· *or* N	Normal beat
L	Left bundle branch block beat
R	Right bundle branch block beat
A	Atrial premature beat
a	Aberrated atrial premature beat
J	Nodal (junctional) premature beat
S	Supraventricular premature beat
V	Premature ventricular contraction
F	Fusion of ventricular and normal beat
[	Start of ventricular flutter/fibrillation
!	Ventricular flutter wave
]	End of ventricular flutter/fibrillation
e	Atrial escape beat
j	Nodal (junctional) escape beat
E	Ventricular escape beat
/	Paced beat
f	Fusion of paced and normal beat
x	Non-conducted P-wave (blocked APB)
Q	Unclassifiable beat
|	Isolated QRS-like artifact

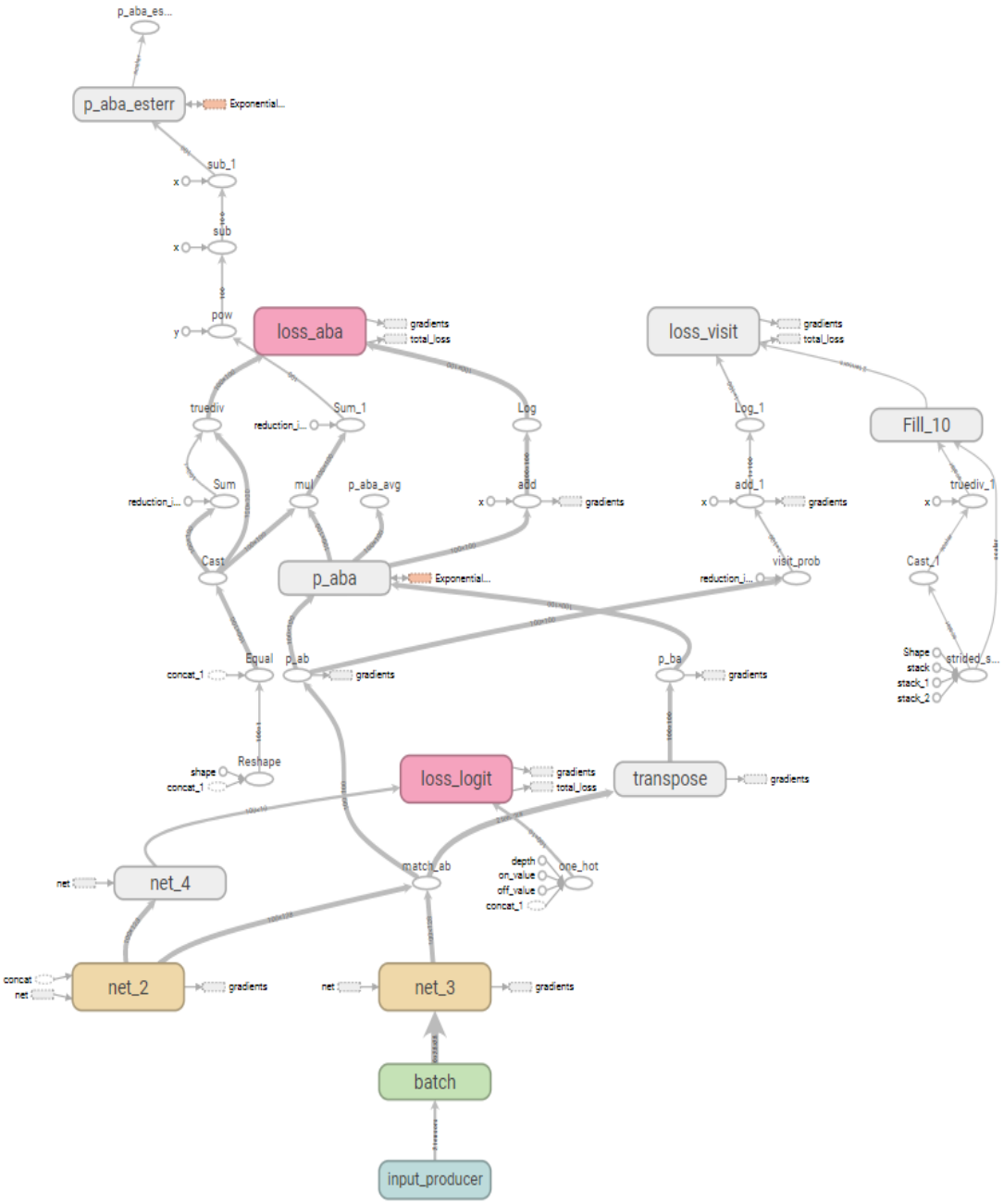

Figure A.1 Schematic diagram of semi-supervised training network.

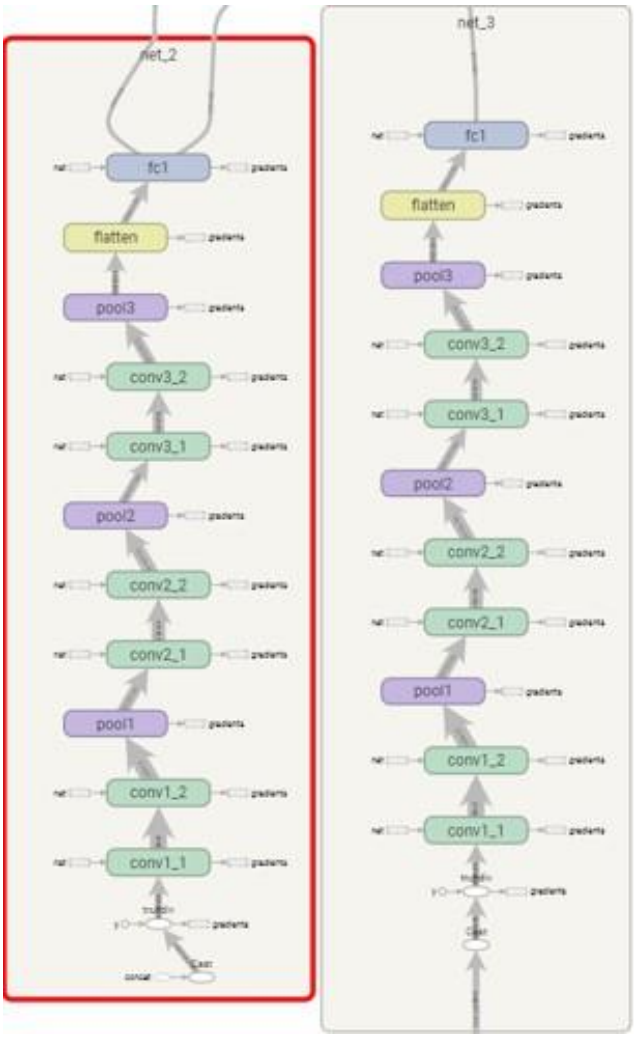

Figure A.2 Schematic diagram of supervised learning and unsupervised learning in side of Figure A.1.

BIBLIOGRAPHY

1. Clifford G D, Azuaje F, McSharry P E. Advanced Methods and Tools for ECG Data Analysis. London: Artech House, 2006

2. Zhu H H, Dong J. An R-peak detection method based on peaks of Shannon energy envelope. Biomed Signal Process Control, 2013, 5: 466–474

3. Goldberger A L, Amaral L A N, Glass L, et al. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation, 2000, 101: 215–220

4. Ye C, Kumar B V, Coimbra M T. Heartbeat classification using morphological and dynamic features of ECG signals.IEEE Trans Biomed Eng, 2012, 59: 2930–2941

5. Kutlu Y, Kuntalp D. A multi-stage automatic arrhythmia recognition and classification system. Comput Biol Med, 2011, 41: 37–45

6. Ubeyli E D. Combining recurrent neural networks with eigenvector methods for classification of ECG beats. Digit Signal Process, 2009, 19: 320–329

7. Ubeyli E D. Recurrent neural networks with composite features for detection of electrocardiographic changes in partial epileptic patients. Comput Biol Med, 2008, 38: 401–410

8. Lei W K, Li B N, Dong M C, et al. An application of morphological feature extraction and support vector machines incomputerized ECG interpretation. In: Proceedings of 6th Mexican International Conference on Artificial Intelligence, Aguascallentes, 2007. 82–90

9. Li X H, Shu L, Hu H L. Kernel-based nonlinear dimensionality reduction for electrocardiogram recognition. NeuralComput Appl, 2009, 18: 1013–1020

10. Doquire G, de Lannoy D, Francois D, et al. Feature Selection for Interpatient Supervised Heart Beat Classification. In: Proceedings of International Conference on Bio-inspired Systems and Signal Processing, New York, 2011. 67–73

11. Martis R J, Chakraborty C, Ray A K. A two-stage mechanism for registration and classification of ECG using Gaussian mixture model. Pattern Recogn, 2009, 11: 2979–2988

12. De Chazal P, O'Dwyer M, Reilly R B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng, 2004, 51: 1196–1206

13. Wubbeler G, Stavridis M, Kreiseler D, et al. Verification of humans using the electrocardiogram. Pattern Recogn Lett, 2007, 28: 1172–1175

14. T. Joachims. Transductive inference for text classification using support vector machines. In ICML, volume 99, pages 200–209, 1999. 2

15. Coghill, Anne M. and Lorrin R. Garson, eds. The ACS Style Guide: Effective Communication of Scientific Information. 3rd ed., Washington, D.C.: American Chemical Society, 2006.

16. Y. Kim. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014. 7

17. D. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 4

18. D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 3581–3589. Curran Associates, Inc., 2014. 7

19. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc., 2012. 1

20. Q. V. Le. Building high-level features using large scale unsupervised learning. In 2013 IEEE international conference on acoustics, speech and signal processing, pages 8595–8598. IEEE, 2013. 2

21. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. 1

22. Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, 1998. 4

23. D.-H. Lee. Pseudo-label: The simple and efficient semisupervised learning method for deep neural networks. In Workshop on Challenges in Representation Learning, ICML, volume 3, page 2, 2013. 1, 2

24. L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther. Auxiliary deep generative models. arXiv preprint arXiv:1602.05473, 2016. 7

25. T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii. Distributional smoothing by virtual adversarial examples. arXiv preprint arXiv:1507.00677, 2015. 7

26. Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In NIPS workshop on deep learning and unsupervised feature learning, volume 2011, page 4. Granada, Spain, 2011. 6

27. A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015. 3

28. M. Ranzato and M. Szummer. Semi-supervised learning of compact document representations with deep networks. In Proceedings of the 25th international conference on Machine learning, pages 792–799. ACM, 2008. 2

29. A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In Advances in Neural Information Processing Systems, pages 3546–3554, 2015. 5

30. K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In European conference on computer vision, pages 213–226. Springer, 2010. 7

31. M. Sajjadi, M. Javanmardi, and T. Tasdizen. Mutual exclusivity loss for semi-supervised deep learning. In 2016 IEEE International Conference on Image Processing (ICIP), pages 1908–1912. IEEE, 2016. 2

32. M. Sajjadi, M. Javanmardi, and T. Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. arXiv preprint arXiv:1606.04586, 2016. 2, 4, 5, 7

33. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. arXiv preprint arXiv:1606.03498, 2016. 3, 5, 7

34. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition.*Proceedings of the IEEE*, november 1998