論文の要旨

題　目　　Structural Optimization of Deep Belief Network and Its Application for Speech Recognition
（Deep Belief Network の構造最適化と音声認識への応用研究）

氏　名　　Murman Dwi Prasetio

Abstract

Since the first time computers were created, humans have thought about how to make computers learn from experience. The term machine learning is basically a computer process for learning from data. Therefore, it continues to interact with data. All machines learning knowledge will definitely involve data. Data can be the same, but the algorithms and approaches vary to get optimal solutions. Machine learning is a branch of the artificial intelligence discipline which discusses the development of systems based on data. Many things are learned, but basically there are 4 main techniques learned for machine learning.

The first thing preparation for supervised learning is data. Data will usually be divided into 2 groups, namely training data and training data. Training data is used to train algorithms to find suitable models, while testing data is used to test and determine the performance of the model obtained on the testing stage. From the model obtained the predictions of data that are divided into two types depending on the type of output. If the prediction results are discrete. For examples gender classification took from speech (male and female output) is the classification process. While if the thread is continuous, it is called regression process. For examples of applying machine learning in life is detecting a person's disease from the existing symptoms. Another example is detecting heart disease from an electrocardiogram recording. In the case of information retrievers, language translation using a computer, by converting voice into text and spam email filters it is used for detecting and classifying sounds. Natural Language Processing (NLP) is the ability of computers to understand both written text and human speech. NLP techniques require capturing the meaning of an unstructured text from documents or communication from the users. Therefore, NLP is the primary way that systems can interpret text and spoken language.

Computer generated communication by voice has existed for a while now. Automatically understanding speech allows for use in various applications. An utterance is our most natural form of interaction when working with people, but still cannot be reached as a reliable interface between humans and machines. Although they are catching up, even the production quality systems like Google, Apple’s Siri or Amazon Alexa are far off from what a human-generated speech would sound like recently. Building a System of Understanding oral Language (SLU) requires solving several sub-problems, each of which presents significant challenges in itself.

A similar neural network results in a model for molecular activity prediction substantially more effective than production systems used in the pharmaceutical industry. Even though training assays in drug discovery are not typically very large, it is still possible to train very large models by leveraging data from multiple assays in the same model and by using effective regularization schemes. In the area of natural language processing, we describes restricted Boltzmann machine training algorithm suitable for voice data. Then, we introduces a new neural network generative model of parsed sentences capable of generating reasonable samples and demonstrate a performance advantage for deeper variants of the model.

Machine learning approaches, in particular of neural network to emphasized high capacity, scalable models that learn distributed representations of their input. The neural network consists of a number of units that have simple nonlinear transfer functions and approximate capabilities for a number of complex types of problems in comparison to a small number of calculations. Therefore, neural networks are applied for data analysis, data mining and data classification. Adequate learning cannot be done if the size of the network is too small. Conversely, over-fitting occurs in the learning data and loses the ability to generalize if the size is too large. Therefore, the appropriate neural network structure needs to be determined for each target problem for higher neural network performance. Traditionally, neural network structure was determined through a trial and error procedure based on the experience of a neural network designer. However, a very large computational time is required by the determination process.

This dissertation demonstrates the efficiency and generality of structural optimization method of a Deep Belief Network (DBN) which consists of multiple layers Restricted Boltzmann Machines (RBMs) using several kinds of evolutionary computation methods and modularization. DBN has succeeded in acquiring higher data analysis capability by effectively incorporating a feature extraction process which is conventionally performed by trial and error. In DBN, multiple RBMs were incorporated into the learning process as feature extractors. There were two kinds of experimental design to approach the method throughout these studies, data classification or prediction and speech recognition.

The experimental design in data classification or prediction using Caltech101 as benchmark of image classification in many related literature. The image data of the Caltech 101 are gray-scale 30x30 grids images, and each grid is scaled in the range of [0,1]. Images are classified into 4 categories "airplane", "cat", "face", "dolphin". There are 65 images per a category. Here, in the structure of RBM has a weakness such as repeating learning could be consuming the time of structure of evaluation and leads to inefficiency in the structure optimization, so by the modularization structure of optimization is performed efficiently by shortening calculation time. Also, the number of hidden layers and the number of each layer units are optimized, but optimizing them simultaneously optimizes the number of layers and optimizes the number of units first because the search space is too wide.

Neural network, although the learning degree increases as learning is done, it becomes excessive learning and loses generalization ability. Therefore, it is desirable that a network with both a learning error for learning data and a verification error for data unused for learning are low. First, although the target data is divided, simply dividing it into one for learning and verification excessively conforms to these two data, and for the other verification data. In this study, the division method was set to 1: 4: 5. We evaluate the structure by learning error / generalization ability verification error / verification error at that time. Speech recognition presents great advantages to human-computer interaction. It is easy to obtain speech data, and it does not require special skills like using keyboard, entering data via clicking the buttons on the GUI programs, and so on. Transferring text data into electronically media using speech is about 8-10 times faster than hand writing, and about 4-5 times faster than using keyboard by the most skilled typist. Moreover, the user can continue entering text while moving or doing any work that requires her to use her hands. Since a microphone or a telephone can be used, it is more economical to enter data, and it is possible to enter data from a remote point via telephone.

Speech is non-stationary signal where properties change quite rapidly over time. This is fully natural and nice thing but makes the use of DFT or auto-correlation as such impossible. For most phonemes the properties of the speech remain invariant for a short period of time (5-100 ms). Thus for a short window of time, traditional signal processing methods can be applied relatively successfully. Most of speech processing in fact is done in this way: by taking short windows (overlapping possibly) and processing them. The short window of signal like this is called frame. In implementation view the windowing corresponds to what is understood in filter design as window-method: a long signal (of speech for instance or ideal impulse response) is multiplied with a window function of finite length, giving finite length weighted (usually) version of the original signal. In the area of speech recognition, it develops a more accurate acoustic model using a deep belief network. First step we are conducted how to build the simplest data-set in voices, its will be extracted into a single data matrix, and a label vector with the correct label for each data file is created. Once the data turned into an input matrix, the next step is to extract features from the raw data, as is done in many other machine learning pipelines. In this experiment, we are used the simple frequency peak detection. The technique to found the peak of signal, it is called the Short Time Fourier Transform (STFT). The Fast Fourier Transform (FFT) is applied over chunks of the input data, resulting in a 2D FFT "image", usually called the spectrogram.

Acoustic-phonetic approach has been studied in great depth for before centuries. This approach is based upon theory of acoustic phonetics and postulates. The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. Three aspects of speech processing are investigated: acoustic parameterization, recognition algorithms and acoustic modeling. DBN performs feature extraction with unsupervised learning called Pre-training and supervised learning called Fine-tuning is performed based on the extracted features. The structural characteristics of DBNs, there are exists a great relationship between the structure of DBM, the number of hidden layers and units constituting each layer, and the performance in data classification or prediction. Performance improvement is expected by giving an appropriate structure corresponding to data. This model, which uses rectified linear units and dropout, improves the accuracy of classification to predict human voices (male or female) with the accuracy 90%.