Multitasking Learning with unreliable labels

Debjit Paul

Computer Science Department

Saarland University

Masters Thesis Proposal
Supervised by: Prof.Dr.Dietrich Klakow

Abstract

In recent times neural networks are producing impressive results due to the presence of large data sets. However, a common known problem in classification task is unreliable labels, due to artificial annotators, human annotation mistakes. With the increase in size of training data, neural network can outperform several classical Machine Learning algorithms. However, manual annotations of a large data set is expensive. In the field of Natural language Processing, artificial annotated labels are commonly used technique to annotate large data. However, the labels generated by the artificial tagger are unreliable. This thesis explores how to handle unreliable labels for large data set and also analyses the Noisy Label Neural Network (NLNN) algorithm in depth. NLNN algorithm learns not only the neural network parameters but also the noise distribution in the labels. In this thesis, I will investigate how artificial annotator works for Part of Speech tagging and how to clear the noisy labels produced by artificial annotator (Brill Tagger) using NLNN algorithm. In recent years, Multi-tasking Learning has been applied to various tasks. MTL attempts to improve multiple classification tasks simultaneously. Our goal will be to investigate not only how this noise-robust method (NLNN) efficiently works for Single-task learning but also for Multi-tasking learning.

Contents

List of Figures

Chapter 1

Introduction

Recently, the increase in size of data in the world has become overwhelming for scientist. The massive growth in the amount of data that can be accessed due to World Wide Web has partially fulfilled the dream of decreasing the gap made between computers and humans. The advancement in the field of machine learning helped to make progress in solving several tasks such as face-recognition and object-recognition, not only in photographs but also in live videos, detection of latent topics in natural language texts, fraud detection, online search and online recommendation system.

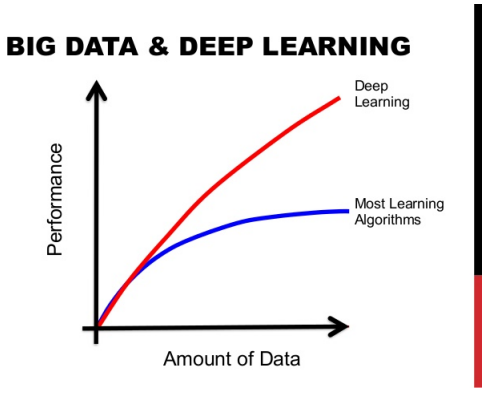

Figure 1.1: Big Data vs Deep Learning

Machine Learning emphasizes on automatic processing from general observation. There has been a lot of progress made in solving a lot of complex tasks in the field of machine learning. Classification technique is one of the paradigm of machine learning, which identifies the concept for new instances with the help of fixed given instances (observations) related to it. The observation plays a key role for classification task. These observations can be manual (human annotated) or artificially generated. The human annotated labels or observations are reliable and correct, nevertheless for a large data it is very expensive and time consuming. Comparatively, the artificial

annotation labels are less-expensive. But, artificial annotator introduces noise and the catch is that during training these labels are assumed to be reliable. Therefore, they may be misleading information that will subvert the model. In fact, noisy labels have been shown to be more harmful than noisy features [30]. Bekker & Goldberger [2] proposed one solution called Noisy Labels Neural Network which estimates the noisy distribution in the noise labels and neural network parameters in the training phase.
Lately, a term has been coined known as 'Multi-tasking Learning' (MTL) which involves in training two or more tasks simultaneously over shared representation [10]. This thesis aims to use and explore the effectiveness of NLNN for Part-of-speech tagging and Multi-tasking Learning. The application of NLNN for automatic POS tagged labels can arouse following research questions:

- Does NLNN produce positive results on a more complex linguistic task?
- Does NLNN produce positive results on noise that is artificially generated ?
- Does NLNN work for MTL?

Chapter 2

Background

In the last decade the use of machine learning has increased drastically throughout computer science and beyond. Machine learning automatically learn programs from observations (training data). The inference of human being in machine learning is intend to provide training data. Preferably, with as little involvement as possible. The idea is to provide enough information (training data) to learn so that it can used to categorize test data.

Generally, the training data are assumed to be accurate and reliable, provided by human experts. Due to the lack of time and knowledge the expert human annotation becomes impractical. Therefore, the researcher often resort to artificial annotators which proves to be inexpensive yet provides unreliable labels.

2.0.1 Noise-robust techniques

In recent times, there has been lot of approaches proposed to make deep learning robust to noise. There has been different approaches to predict noise, one of them was by Larsen et al [4]. They assumed the generation of noise is independent of class labels and the features. They proposed noise as outlier probability and learned using probabilistic modelling and back propagation.

Mnih & Hinton [6] proposed an advanced noise-robust model which uses robust loss functions. This method can handle missing class problems. A similar approach was proposed by Lee [5] using minimum entropy regularization. They used pseudo-labels as training data for unlabeled data. However, in both the approaches they considered binary classification. Reed et al. [8] introduced a simple approach using bootstrapping scheme to handle noise and incomplete labels. In this work they labeled the unlabeled based on a small list of seeds (labels).

A similar approach to NLNN [2] was proposed by Sukhbaatar and Fergusk [9] to model noise with linear layer. They introduce an extra linear layer whose weight matrix has the shape of a noise distribution. However, the additional linear layer does not constitute a model of a noisy channel.

There are differences between the earlier works and the Noisy Labels Neural Network approach which has been discussed in the next chapter. One the main significant distention in the previous works is that they have not explicitly modeled noise distribution.

Chapter 3

Noisy Labels Neutral Network

Bekker & Goldberger [2] introduced Noise Labels Neural Network algorithm which address the problem of training neural network on noisy labels. They proposed an idea to add an extra layer of noise channel with the neural network architecture. Similar idea was proposed by Sukhbaatar and Fergus's [9] of introducing an extra linear layer after the soft-max layer, they have shown the linear layer as the transformation matrix between true labels and noise labels with some strong assumption. However, Bekker & Goldberger [2] modeled the noise generation by a parameter. They [2] assumed that the input features are independent of noise generation in observed labels.During the training phase, NLNN model learns the neural network parameter and the noise distribution of the noise labels using the probabilistic modelling and backpropagation. This chapter illustrates its training process. The equations and variable names in this chapter were kept similar to their original forms in [2].

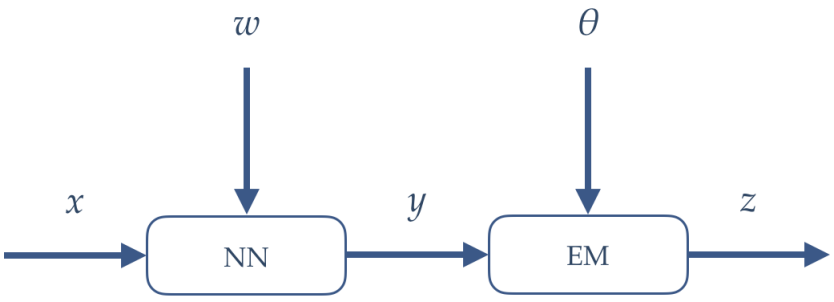

Figure 3.1: NLNN model

In [2] the above NLNN model was proposed. In this approach the noise generation

is modelled as θ which is a confusion matrix between the hidden true labels and the observed noisy labels:

$$\theta(i,j) = p(z = j|y = i) \tag{3.1}$$

The idea is to consider that we are given n feature vectors $x_1, ..., x_n$ with the corresponding noisy labels $z_1, ..., z_n$ which can be seen as noisy versions of the hidden true labels $y_1, ..., y_n$. The probability of observing noise will be

$$p(z = j|x; w, \theta) = \sum_{i=1}^{k} p(z = j|y = i; \theta)p(y = i|x; w) \tag{3.2}$$

Therefore, the log-likelihood of the model parameters is:

$$L(w, \theta) = \sum_{t=1}^{n} log(\sum_{i=1}^{k} p(z_t|y_t = i; \theta)p(y_t = i|x_t; w)) \tag{3.3}$$

Our goal is to find the noise distribution θ and the neural network parameter w which maximizes the above likelihood function. In order to learn about the optimal parameters set we use EM algorithm. The EM algorithm is an iterative process for finding the parameters that maximizes the likelihood. The EM algorithm is explained as follows:

In **E-step**:

We estimate $y_1, ...y_n$ true hidden labels from the noisy labels and current parameters (θ,w).

$$c_{ti} = p(y_t = i|x_t, z_t; w_0, \theta_0) = \frac{p(z_t|y_t = i; \theta_0)p(y_t = i|x_t; w_0)}{\sum_t p(z_t|y_t = i; \theta_0)p(y_t = i|x_t; w_0)} \tag{3.4}$$

We update the confusion matrix θ based on the hidden true labels .

$$\theta(i,j) = \frac{\sum_t 1_{z_t=j} p(y_t = i|x_t; w)}{\sum_t p(y_t = i|x_t; w)} \tag{3.5}$$

In **M-step**: We update the parameters of neural network w using standard back-propagation in order to maximize the log likelihood.

However, EM algorithm has one drawback that it tends to converge to a local optimum. The above consequence can be overcome with proper initialization. A potential solution to the challenge of proper initialization can be using well-trained and tuned neural network. Based on the output of the neural network we can set the parameters of the EM algorithm.

The NLNN algorithm is described in the following table :

Input: Data points x and corresponding labels z

Initialization:
1.Pre-training
2.Output parameter
3.Initialise θ based on predictions y and original labels z

E-step:
Estimate true labels c based on parameters w and initialized θ

M-step:
1.Re-compute θ based on c
2.Re-train NN to maximize likelihood based on c
Iterate Repeat **E-step** and **M-step** until likelihood converges
Output Predictions y, NN parameter w and noise parameter θ

Table 3.1: NLNN Algorithm

Chapter 4

Experiments

In Bekker & Goldberger [2] performed NLNNs on two tasks: a handwritten recognition task and phoneme classification tasks. They generate noisy labels from clean labels by injecting noise using permutation and uniform distribution. In our experiments we used artificial annotator to annotate labels, hereby injecting noise.

The goal of the section is to demonstrate the following statements :

- The effectiveness of the NLNN model on POS tagging task using artificial annotator on mixing two different corpus
- The effectiveness of the NLNN model on POS tagging task using artificial annotator on same corpus

4.1 Corpus

Penn Treebank corpus is popular corpus to test methods for part-of-speech tagging task. Penn Treebank corpus contains approximately 7 million words of part-of-speech tagged text. I divided the data set into two sections 0-22 for training, 23-24 for testing. The North American News Text corpus is composed of news text that has been formatted using TIPSTER-style SGML markup.

4.2 Experiment 1

4.2.1 Noise Generation

Brill Tagger was used for part-of-speech tagging to artificially annote the North American News Corpus. The brill tagger labelled the North American Data set with unreliable labels. In order to generate the training data I mixed this unreliable data with

the manually annotated Penn Tree-bank section 0-20 data set. I used 100K training data and 25K test data in our first experiment.

4.2.2 Method

I trained NLNN on data mixed with Penn tree bank data (clean labels) and North American News Text Corpus(noisy labels) in the following way. First, network parameters were initialized with a 30 epochs of standard back propagation (BiLSTMs), where the input was noisy data. The pre-training was performed to initialize the parameters for EM algorithm. The EM algorithm was then initiated and left to run until convergence of the likelihood and each iteration also involved a re-training of the neural network for 15 epochs.

4.2.3 Result

Figure 4.1 shows accuracy scores obtained with respect to NN and NLNN. The results show that the NLNN is able to improve performance, although the difference is very small. In order to get a more accurate evaluation for this task F1 -score, which is shown in 4.2.

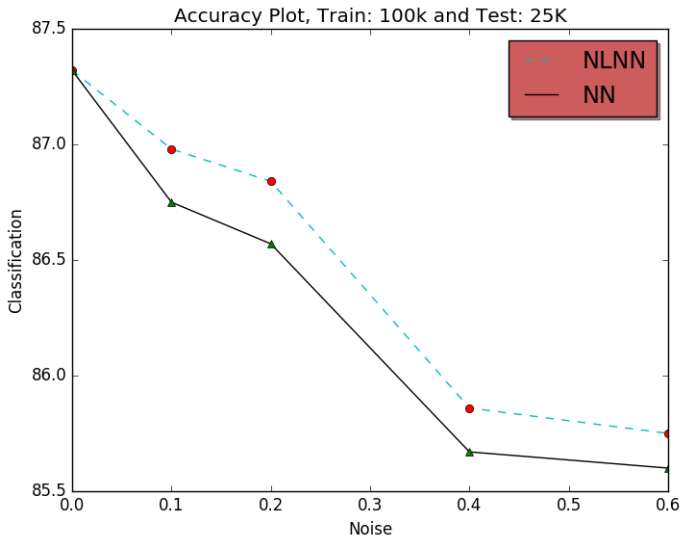

Figure 4.1: Accuracy of NLNN and NN on Penn Tree bank data set and North American News Corpus

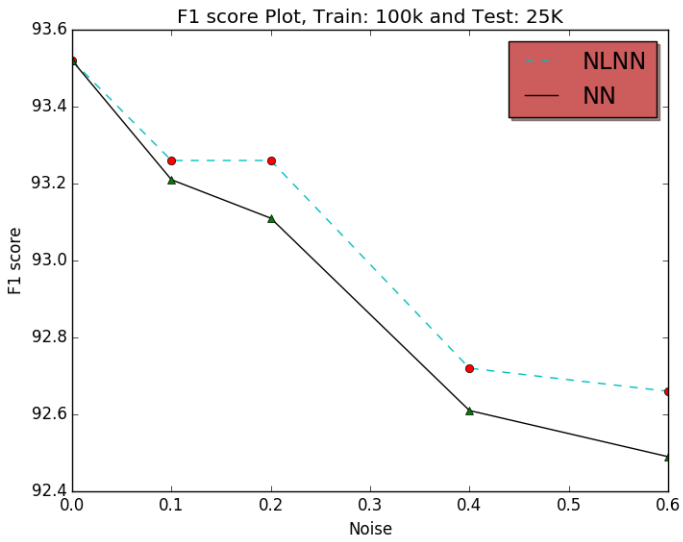

Figure 4.2: F1 score of NLNN and NN on Penn Tree bank data set and North American News Corpus

4.3 Experiment 2

4.3.1 Noise Generation

Brill Tagger was used for part-of-speech tagging to artificially annote the Penn-tree bank section 20-21 Data-set. The brill tagger labelled the section with unreliable labels. Added this unreliable data with the clean manually annotated Penn Tree-bank section 0-20 data set. I used 100K training data and 25K test data in our first experiment.

4.3.2 Method

I trained NLNN on Penn tree bank data in the following way. First, network parameters were initialized with a 30 epochs of standard back propagation (BiLSTMs), where the input was noisy data. The EM algorithm was then initiated. In each iteration I replaced the labels of the clean section with the original clean labels. This process continue to run until convergence of the likelihood and each iteration also involved a re-training of the neural network for 15 epochs.

4.3.3 Result

Figure 4.3 shows accuracy scores obtained with respect to NN and NLNN. The results show that the NLNN is able to improve performance, in compare to the earlier results.

Clearly, in 40% noise fraction we can observe a clear improvement of around 5% in accuracy. In order to get a more accurate evaluation for this task I computed the F1

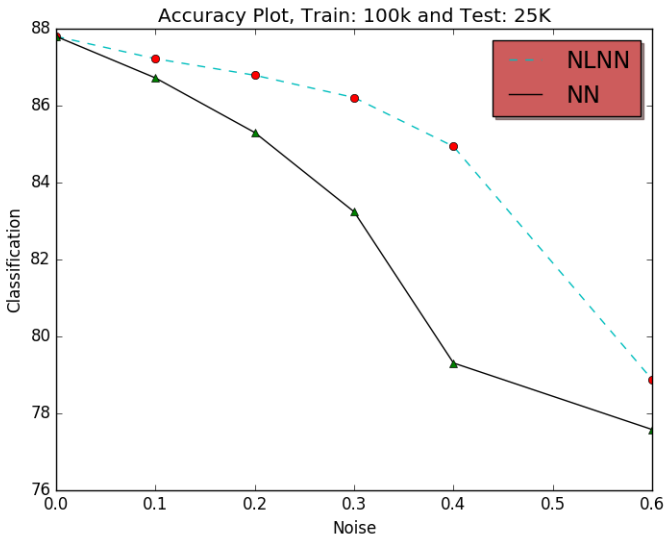

Figure 4.3: Accuracy of NLNN and NN on Penn Tree bank data set

-score, which is shown in 4.4.

4.4 Discussions

The performance of NLNN depends on many factors. Good initialization of parameter for EM model is one of them. A good initialization depends on how well we pre-train NN (BiLSTMs) as its output determines the quality of initial estimation of hidden true labels. Since on the mixed training data (Penn tree bank and American News Corpus) NLNN make a relative improvement compare to NN, it indicates that NLNN has a positive effect on automatic labelling.

On the Penn tree bank, when I artificially (brill tagger) labelled a section of Penn tree bank data and also replaced the labels with the clean labels in every iterations of section it appears the initialization and iterative improved estimation of true labels. The reason behind the improvement is Penn tree bank data has well separated sentences which are used as input features of the NN. The other factor is keeping the clean section of training data intact in every iteration as it decrease the possibility of NN to learn from noisy features.

There has been lot of work related to make NN robust to noise. In best of my knowledge there has not been much work related to make Multitasking Learning NN

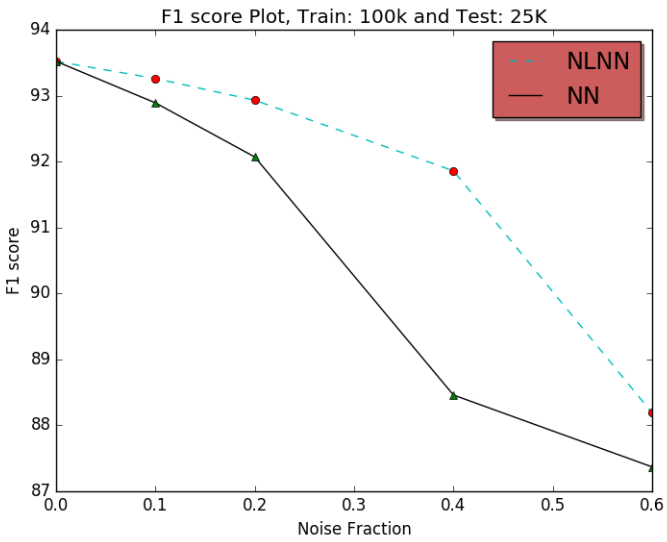

Figure 4.4: F1 score of NLNN and NN on Penn Tree bank data set

robust to noise. In the next chapter, I have discussed the idea behind handling noise for MTL.

Chapter 5

Multi-tasking Learning and Future Plan

5.1 Multi-tasking Learning

Multi-task learning (MTL) is a machine learning technique that aims at improving the generalization performance of a task using other related tasks [MULTI-TASK SEQUENCE TO SEQUENCE LEARNING]. In the field of Natural Language Processing, there are many related tasks, which makes MTL more effective. Generally, features used for one task can be useful for another task, MTL leverages this idea. Multi-tasking Learning systems are commonly designed to train a single neural network for multiple tasks, using a share representation of features. In a recent study, by Hctor Martnez Alonso & Barbara Plank in [1] shown that not all combination of tasks can produce significant results. They proposed that in MTL there are two kind of tasks one is auxiliary tasks (POS, Chunk, etc.) and another is main tasks (NER, Frames, etc.).

5.1.1 NLNN for MTL

The advancement of MTL in the field of NLP helped to improve and learn several tasks in parallel. The performance of MTL depends on the size of the training data. Generally, artificial annotators are used to label large data. However, that introduces unreliable labels in the training data. Therefore, the main idea behind exploring NLNN for MTL is to improve the performance on unreliable labels. NLNN can be used to find the noise distributions of the labels and also estimating the neural network parameters.

5.2 Future Plan

While up-till now this thesis dealt with NLNN for Single Task Learning (POS tagging task). The future road map for the thesis is as follows:

- Study and Implementation of Multi-tasking Learning for Natural Language Processing tasks. The study will explore which two or more NLP tasks trains well together. It will take 4 weeks to implement MTL.
- Implementation of NLNN for MTL. This experiment will explore the strength of NLNN, as it will deal with more than one set of unreliable labels. Duration of this experiment will take another 6 weeks.
- Study and analyzing the results will take another 3 weeks.

Bibliography

[1] Alonso, Hctor Martnez, and Barbara Plank. "When is multitask learning effective? Semantic sequence prediction under varying data conditions."

[2] Bekker, Alan Joseph, and Jacob Goldberger. "Training deep neural-networks based on unreliable labels." Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[3] D. Nettleton, A. Orriols-Puig, Fornells, A. (2010) A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial intelligence review.

[4] Larsen, J., Nonboe, L., Hintz-Madsen, M., Hansen, L. K. (1998). Design of robust neural network classifiers. In IEEE International Con- ference on Acoustics, Speech and Signal Processing, 2, pp. 1205-1208.

[5] Lee, D. H. (2013). Pseudo-label: The simple and efficient semi- supervised learning method for deep neural networks. In Workshop on Challenges in Representation Learning, ICML, 3, p. 2.

[6] Mnih, Volodymyr, and Geoffrey E. Hinton. "Learning to label aerial images from noisy data." Proceedings of the 29th International Conference on Machine Learning (ICML-12). 2012.

[7] Plank, Barbara, Anders Sgaard, and Yoav Goldberg. "Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss." arXiv preprint arXiv:1604.05529 (2016).

[8] Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A. (2014). Training deep neural networks on noisy labels with bootstrap- ping. arXiv preprint arXiv:1412.6596.

[9] Sukhbaatar, S., Fergus, R. (2014). Learning from noisy labels with deep neural networks. arXiv preprint arXiv:1406.2080, 2(3), 4.

[10] Thanda, Abhinav, and Shankar M. Venkatesan. "Multi-task Learning Of Deep Neural Networks For Audio Visual Automatic Speech Recognition." arXiv preprint arXiv:1701.02477 (2017).