Improved Deep Belief Network Model and Its Application in Named Entity Recognition of Chinese Electronic Medical Records

Wusuo Li, Shenghui Shi*, Ziqiao Gao, Wei Wei, Qunxiong Zhu, Xiaoyong Lin, Daguang Jiang, Shang Gao
College of Information Science and Technology
Beijing University of Chemical Technology
Beijing, 100029, China
e-mail: zmsnlws@163.com, shish@mail.buct.edu.cn, fbigaoziqiao@gmail.com, weiw4180@163.com, zhuqx@mail.buct.edu.cn, linxy@mail.buct.edu.cn, jiangdg@mail.buct.edu.cn, gaoshang1996@hotmail.com

***Abstract*—The common named entity recognition(NER) methods of Chinese electronic medical records(EMR), are often based on conditional probability of the likelihood of estimation. It is difficult to obtain deeper characteristics between words. In order to obtain deeper level of semantics from Chinese words, this paper purposes an improved deep belief networks (DBN) model by adding the part-of-speech (POS) node, so as to identify the named entities better. This study compares three experiments: Improved DBN model method, original DBN model method and conditional random field (CRF) based method. The result shows that the F1-score of the improved DBN method surpasses both the original DBN and the CRF method's, reaching 91.749%. The experiment shows that the POS node is beneficial to the NER, and the improved DBN model is validated in NER.**

Keywords-named entity recognition(NER); deep belief network(DBN); Chinese electronic medical records

I. INTRODUCTION

The wide range applications and development of medical information technology have produced massive amounts of electronic medical records (EMR) data. Electronic medical records are detailed records of patient's medical data generated by medical practitioners, such as doctors and nurses, engaged in medical activities. By analyzing the electronic medical records, the researcher can excavate a large number of medical knowledge that is closely related to patients, which has always been a consensus [1]. However, a large number of patient's medical and health information is stored in the narrative text recorded by the medical staff, rather than the forms that the computer system program could be used directly. For these large amounts of data, Named Entity Recognition (NER) in the field of Natural Language Processing (NLP), can extract medical information from important documents and medical texts, and has been introduced in the medical field, showing great practical value in many practices [2]. With the continuous development of NLP, NLP will be better integrated with biomedical Big Data Analysis (BDA), to jointly serve the fields of medical diagnosis and treatment. The study of different named entity recognition methods is an important research direction for each relative researcher.

II. BACKGROUND

Named entity (NE) was originally defined at MUC-6 [3]. It refers to an entity with a specific meaning and has now become one of the most important tasks in NLP. In the medical field, the named entity in the electronic medical records text, mainly refer to the entity that expresses the specific meaning in the records of the patient's medical treatment, such as disease names, symptoms, drug names, check names, medical means. [4]

The earlier proposed methods for named entity recognition of electronic medical records are the ones based on dictionary and rules, like NTU [5], Oki [6]. Then some systems, tools or organizations were created, such as MedKAT (Medical Knowledge Analysis Tool) [7] developed by IBM and the cTAKES (clinical Text Analysis and Knowledge Extraction System) [8] developed by the Mayo Clinic.

In 2009 of the Center of Informatics for Integrating Biology and the Bedside (I2B2), the systems of the named entity recognition were mainly based on rules or on the combinations of rules and machine learning(ML) [9]. However, in 2010 of I2B2 [10], many efficient systems were mainly used ML-based methods [11-13].

So far, there have been many ML-based advanced medical NER system algorithms, such as Conditional Random Fields (CRFs) [13], Support Vector Machines (SVMs) [14], Maximum Entropy Markov Models (MEMMs) [15] and so on.

Compared to the English name of the medical entity recognition, the Chinese name entity recognition has also attracted increasingly attention from researchers. The traditional ML-based named entity recognition method has also been widely used. Such as Lin et al [16] applied the support vector machine to achieve excellent accuracy in identifying Chinese named entities. Zhou et al [17] build large-scale named entity recognition corpora from Chinese Wikipedia, and the NER models achieved the best performance when combining their silver-standard corpora with gold-standard corpora. Wang et al [18] presented a detailed analysis for recognizing symptom names from free-text clinical records of traditional Chinese medicine through analyzing labeling results of CRFs.

978-1-5386-4794-3/18/$31.00 ©2018 IEEE

In recent years, various NER researchers have become increasingly interested in the NLP system that involve deep learning [19, 20], which has excellent performance in image processing [21-23]. Deep learning is one part of machine learning, and it can be used to design advanced neural networks to learn advanced features of words or graphics. So, it is recognized as a great prospect in the field of NLP.

The fundamental approach of the traditional machine learning NER models, such as the Hidden Markov model (HMM), the maximum entropy Markov model (MEMM) and the Conditional random field (CRF) and so on, is to maximize the likelihood of the conditional probability, not to extract the meaning of words and sentences. To solve this problem, based on the Deep Belief Networks (DBN) proposed by Geoffrey Hinton et al [24], this paper proposes a NER method for Chinese electronic medical records with an improved DBN model, then they are compared with the traditional CRF-based NER system to verify the feasibility of the improved DBN model approach.

III. Research Foundation

Various researchers have proposed various methods for designing deep neural networks for NLP systems. This paper base the one of the earliest depth neural network proposed by Geoffrey Hinton [24] – DBN, which is composed of several layers of RBM and a layer of BP. This structure has been widely used in a variety of tasks, and achieved excellent performance. When calculating whether a word is a named entity, the purpose is to minimize the number of unknown data errors.

Set f: $R^D \rightarrow \{0, 1, \ldots, L\}$ is a predictor function, then this loss function can be written as: $l_{0,1} = \sum_{i=0}^{|D|} I_{f(x^{(i)}) \neq y^{(i)}}$, D is the training set. The indicator function I is defined as: $I_x = \begin{cases} 1 & if\ x\ is\ true \\ 0 & otherwise \end{cases}$. f defined as: $f(x) = argmax_k\ P\ (Y = k|x, \theta)$, Since the 0-1 loss is not differential, the optimization is too time-consuming for large models (thousands or millions of parameters). So, the logarithmic likelihood $\mathcal{L}(\theta, D) = \sum_{i=0}^{|D|} logP(Y = y^{(i)}|x^{(i)}, \theta)$ should be maximized.

Among them D: the dimension of the input, $D_h^{(i)}$: the number of nodes in the i-th layer hidden layer; $f_\theta(x)$, $f(x)$: the relevant classification function. $P(Y|x, \theta)$: The number of annotations; the logarithmic likelihood of the model defined by the parameter $argmax_k P(Y = k \vee x, \theta)$; $L(\theta, D)$: the logarithmic degree D of the model defined by the parameter θ; $l(\theta, D)$: the loss value of the predictive function f of the parameterized data set D, θ: all parameter sets for a given model.

IV. Improved Deep Belief Networks Model

Considering the POS should be a very important factor for the NER, we improved the DBN model by adding the

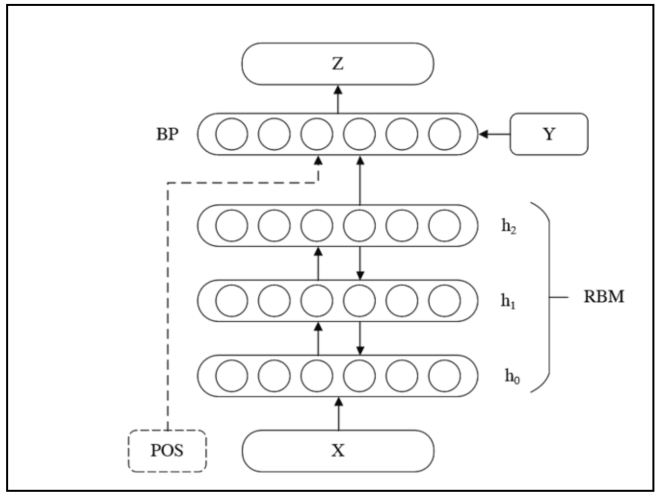

Figure 1. the structure of improved DBN.

POS node as input to the layer named BP. Since POS is already done by the training of neural network results, no longer need to be reduced, or clustered by RBM layer, the POS is not included in the process of RBM training. So, the results of RBM layer and the POS are inputs to the final BP layer for training. As shown in Fig. 1, the POS node represents the input of POS corresponding the word vector X.

The training process shown in Fig. 1: first input the X, and then train the RBM layers of the DBM. After the training of the RMB layers, input X to RBM layers, POS to the POS node and Y to the BP layer. Then train the BP and DBM fine-tuning.

In the Fig. 1, input X: word vector trained by word2vec, POS: X corresponding part of speech, RBM: RBM layer of DBN, hi: i-th layer of RBM, BP: BP layer of DBN, Y: corresponding mark of X, Z: the output obtained by training BP.

V. Experiment

A. General Overview

This paper studies three kinds of named entity recognition methods: the improved DBN model method with POS, the original DBN model method and the CRF-based method. Both improved DBN and original DBN use Tomas Mikolov's word2vec training word vectors. And all three named entity recognition methods are compared with each other.

The word vectors used by DBN were trained from the unlabeled corpus. Three models were trained using the same data set, and their performances on the test set were recorded. All three NER systems were evaluated by using standard: precision, recall, and F1-score in this study.

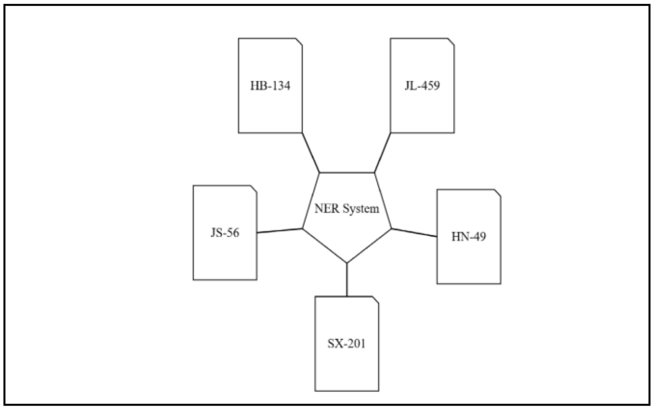

Figure 2. The distribution of and the number of documents in each region.

B. Data Set

In this study, the clinical data of Chinese single disease were used, as shown in Fig. 2, which contains 19287 records of 1066 clinical files from five Hospital Information System (HIS) databases in five different regions of China. These large amounts of data are divided for two uses: the first one is to use the clinical records to train word vectors; the second one is to extract 300 medical records from the data, and to build a labelled corpus by labelling these records. For each medical record, the named entities related to clinical (diseases, symptoms, drugs, examination, medical means) were marked out.

In this study, 300 documents were further divided into two parts, two thirds for training and one third for testing. Only a very small amount of preprocessing (for example, deleting redundant blank lines, newlines, replacing sensitive information) is used to modify these records.

C. Deep Belief Networks-based Named Entity Recognition

In this paper, two NER methods based on DBN model are studied. One is to improve the DBN model with POS. The second is based on the original DBN model. As shown in Fig. 3, the two DBN-based NER is trained as follows: Firstly, the Chinese Words are segment and the parts of speech are taken out. Secondly, the word vectors are trained by words after word segmentation. Then, all the second use of the training data are manually marked. Finally, the word vectors corresponding to the words are used as DBN inputs, and the marks of words are trained and tested as output standards. The detailed process is as Fig. 3.

1) Word segmentation with part of speech

Unlike English, Chinese characters are connected together, and there is no space as a separator. Therefore, the first step in the named Chinese entities recognition is the word segmentation. In this study, we first use the jieba in the python language to segment all the data of the two different datasets. Jieba using HMM and Viterbi algorithm is one of the most advanced word segmentation system, and it can add our own dictionary. Moreover, it can segment the word at the same time marking the word's Part-of-Speech (POS). In this paper, the medical dictionary is 74732, the number of words is 4581011.

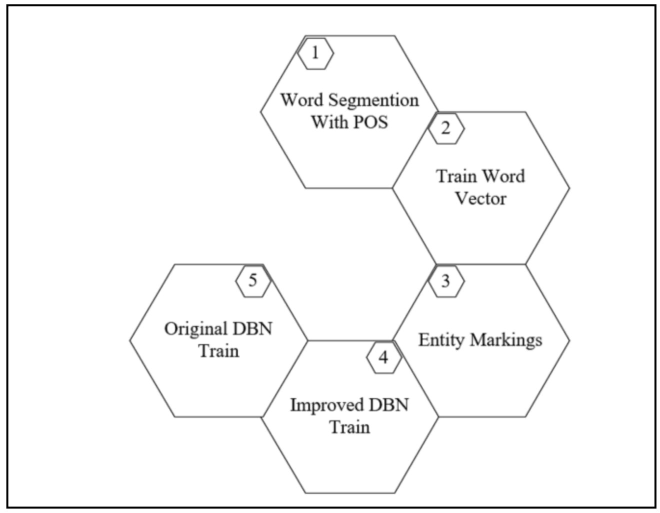

Figure 3. DBN training process.

2) Word vector

Since the text can't be directly used by deep neural network, the traditional method, like word bag model (bag of words), is to represent the text numerically. While, in 2013 Tomas Mikolov [25] put forward the word2vec model, which can better express grammar information. In this study, word2vec is implemented with java version to train word vectors.

3) Entity marking and part of speech

After the word segmentation, for the data to be trained, it is necessary to mark whether the word is an entity or not. A number of staffs used 300 electronic medical records data to mark the named entities that can be used for training and testing.

Before DBN training, we have to deal with this problem: The POS are also text that can't be directly used as BP input. In this study, a simple method is used: the POS is simply replaced by a number. Sorting all n POS in the natural order, and then dividing 0 to 1 into n equally spaced points, each of which corresponds to one of POS, for example, the numerals of the three parts of speech (POS) (a, n, v) is (0, 0.5, 1).

4) Improved and original deep belief networks train

Based on the DBN model, this paper studies two named entity recognition methods, one is improved DBN model and the other one is original DBN model. As shown in Fig. 1, the dashed part of the structure belongs to the improved DBN model rather than original DBN model. Both DBN model methods uses the word vector of the word2vec training's result as X, and takes whether the word is entity as Y. Both X and Y are trained as inputs to the models, first training the RBM layers, and then fine-tuning. The main difference between the improved DBN model (shown in Fig. 1) and the original DBN model is whether the inputs of BP include POS.

In this study, the learning rate of two DBNs is set to 0.01 and the vector dimension is 200. The number of hidden nodes is set to 300,300,300. All DBN parameters are updated by reverse spread using the small batch random gradient drop.

TABLE I. RESULTS OF METHODS.

Methods	Results (%)		
	F1-score	*Precision*	*Recall*
Original DBN	90.280	90.242	90.318
Improved DBN	91.749	91.685	91.813
CRF	91.386	92.667	90.141

D. Conditional Random Field -based Named Entity Recognition

The CRF model decodes the sequence by the invariant Markov chain and the Viterbi algorithm. The training criterion maximizes the likelihood of the conditional probability of the output variable y for given observation x (For example, B, I, O; B - the beginning of the entity; I - inside the entity; O - outside the entity). The x represents the words and contextual words in the sentence. CRFs are basically designed for sequence tagging because they can sequentially establish the relationships between adjacent blocks. Therefore, it has been widely used in a variety of NER tasks. Currently it is recognized as one of the state-of-the-art NER method. Therefore, this study uses CRF as a benchmark method to compare with the DBN-based NER method. This study uses CRF ++, which is one of the most popular versions of each implementation of CRF.

VI. RESULT

The training data of the above 5 regions are processed by word segmentation, named entity tagging and part of speech reference. Then, the improved DBN model method, the original DBN model method and the benchmark CRF method are trained and tested with their results recorded.

As shown in Table I and Fig. 4, the results of the benchmark method CRF, the original DBN method, and the improved DBN method with POS in the test and verification data. The F1-score of the reference method (CRF) is

91.386%. The original DBN is 90.280%, not better than CRF performance. The F1-score of the improved DBN method with POS was increased to 91.749%, which is about 0.4% higher than the CRF.

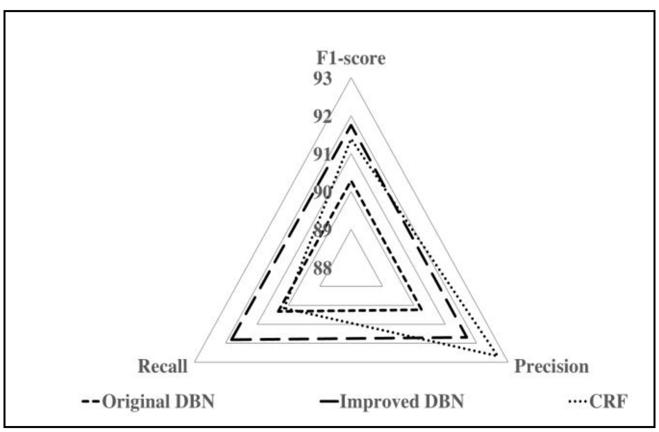

Figure 4. Results of CRF, original DBN and improved DBN.

TABLE II. SEMANTIC CORRELATION OF "COUGH" CALCULATED BY WORD VECTOR DISTANCE

Entity (English Meaning)	Correlation (%)
咳嗽(cough)	100.000
咳痰(sputum)	95.615
乏力(fatigue)	93.451
症状(symptom)	92.439
关节疼痛(joint pain)	88.064
无尿(no urine)	88.000
头晕(dizziness)	87.612
心悸(palpitation)	86.572
便血(stool blood)	86.185

Compared with the original DBN model method and the improved DBN model method, it is easy to find that the development of the improved DBN model method comes from the records of POS. The further analysis of DBN-based NER system makes it easy to find that word vector can automatically capture semantic information. Table II and Fig. 5 show an example of a "cough" semantic related word obtained by word vector training. The word distance is obtained by using the word vector to calculate the cosine similarity.

Most of the neighbors in Fig. 5 are related to the target word, which is an important difference between the word vector representation and the traditional word bag model representation. It shows that the DBN model is feasible in the Named Entity Recognition of Chinese electronic medical records.

VII. SUMMARY AND PROSPECT

This paper purposes an improved DBN model method of NER in Chinese electronic medical records. Without the POS, the original DBN method is not superior to one of the state of the art methods, CRF. After adding the POS, the improved DBN method is not only superior to the original DBN method, but also exceeds the CRF method on the F1-value, reaching 91.749%. Further analysis shows that word vector training can automatically capture the semantic information, and the performance of the improved DBN with POS is better.

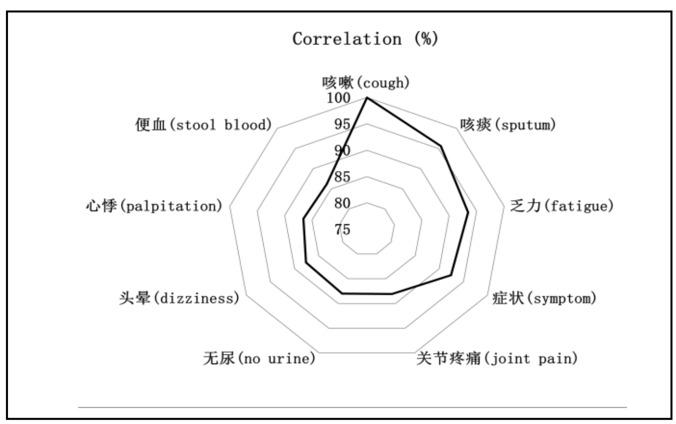

Figure 5. Correlation of "cough" calculated by word vector distance.

In the future, we plan to study different methods to improve the processing power of Chinese electronic medical records based on DBN NER system. An easy way is to increase the size of labeled and unlabeled medical corpora. Another good research direction is to fix the recognition of named entity in some particular small areas, such as single disease. In this way, the knowledge content is richer, the direction is clearer, so it is easier to produce accurate results.

ACKNOWLEDGMENT

This work was supported by Single Disease Multi-Source Heterogeneous Data Fusion System and its Application in Clinical and Pharmacy Research (GWGJ201608), from Beijing Municipal Education Commission.

REFERENCES

[1] R. C. Wasserman, "Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research," Academic Pediatrics, vol. 11, no. 4, p. 280, 2011.

[2] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research," Yearb Med Inform, pp. 128-44, 2008.

[3] B. Sundheim and B. Sundheim, "Message Understanding Conference-6: a brief history," in Conference on Computational Linguistics, 1996, pp. 466-471.

[4] J. F. Yang, Q. B. Yu, Y. Guan, and Z. P. Jiang, "An overview of research on electronic medical record oriented named entity recognition and entity relation extraction," Zidonghua Xuebao/acta Automatica Sinica, vol. 40, no. 8, pp. 1537-1562, 2014.

[5] H. H. Chen, Y. W. Ding, S. C. Tsai, and G. W. Bian, "Description of the NTU System Used for MET2," 1998.

[6] J. Fukumoto, F. Masui, M. Shimohata, and M. Sasaki, "Oki Electric Industry: Description of the Oki System as Used for MUC-7," 1998.

[7] Jagannathan and A. S. Elmaghraby, "MEDKAT: multiple expert DELPHI-based Knowledge Acquisition Tool," in ACM Ne Regional Conference, 1985.

[8] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," Journal of the American Medical Informatics Association Jamia, vol. 17, no. 5, p. 507, 2010.

[9] O. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," Journal of the American Medical Informatics Association Jamia, vol. 17, no. 5, p. 514, 2010.

[10] B. De Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu, "Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010," Journal of the American Medical Informatics Association Jamia, vol. 18, no. 5, pp. 557-562, 2011.

[11] YukunChen, MeiLiu, S. Trentrosenbloom, SubramaniMani, and J. Cdenny, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," Journal of the American Medical Informatics Association Jamia, vol. 18, no. 5, p. 601, 2011.

[12] M. Torii, K. Wagholikar, and H. Liu, "Using machine learning for concept extraction on clinical documents from multiple data sources," Journal of the American Medical Informatics Association Jamia, vol. 18, no. 5, p. 580, 2011.

[13] J. D. Lafferty, A. Mccallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in Eighteenth International Conference on Machine Learning, 2001, pp. 282-289.

[14] G. Bakir, T. Hofmann, B. Schölkopf, and A. J. Smola, "Support Vector Machine Learning for Interdependent and Structured Output Spaces," in International Conference on Machine Learning, 2004, p. 104.

[15] A. Mccallum, D. Freitag, and F. C. N. Pereira, "Maximum entropy markov models for information extraction and segmentation," Proc of Icml, pp. 591--598, 1999.

[16] X. D. Lin, H. Peng, and B. Liu, "Chinese Named Entity Recognition using Support Vector Machines," in International Conference on Machine Learning and Cybernetics, 2009, pp. 4216-4220.

[17] J. Zhou, B. C. Li, and G. Chen, "Automatically building large-scale named entity recognition corpora from Chinese Wikipedia," Frontiers of Information Technology & Electronic Engineering, vol. 16, no. 11, pp. 940-956, 2015.

[18] Y. Wang, Y. Liu, Z. Yu, L. Chen, and Y. Jiang, "A preliminary work on symptom name recognition from free-text clinical records of traditional chinese medicine using conditional random fields and reasonable features," in The Workshop on Biomedical Natural Language Processing, 2012, pp. 223-230.

[19] N. Jones, "Computer science: The learning machines," Nature, vol. 505, no. 7482, pp. 146-8, 2014.

[20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, p. 504, 2006.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 2, p. 2012, 2012.

[22] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning Hierarchical Features for Scene Labeling," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 35, no. 8, pp. 1915-29, 2012.

[23] K. Kavukcuoglu, Y. L. Boureau, Y. L. Boureau, K. Gregor, and Y. Lecun, "Learning convolutional feature hierarchies for visual recognition," in International Conference on Neural Information Processing Systems, 2010, pp. 1090-1098.

[24] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, 2012.

[25] T. Mikolov, G. Corrado, K. Chen, J. Dean, T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," in International Conference on Learning Representations, 2013, pp. 1-12.