A Multilevel Deep Learning Method for Data Fusion and Anomaly Detection of Power Big Data

Dong-Lan LIU[1]*, Xin LIU[1], Hao YU[1], Wen-Ting Wang[1], Xiao-Hong ZHAO[1] , Jian-Fei CHEN[2]

[1]State Grid Shandong Electric Power Research Institute. Jinan 250003, PR China

[2]State Grid Shandong Electric Power Company. Jinan 250021, PR China

*Email: liudonglan2006@126.com

Keywords: power big data, restricted Boltzmann machine, recurrent neural networks, anomaly detection, deep learning, data fusion.

Abstract. With the expansion of the power information network scale, various network threats are also increasing. In order to excavate security threats in power grid by making full use of heterogeneous data sources in power big data, this paper maps heterogeneous data in different formats to a unified embedded vector space with deep restricted Boltzmann machine, and achieves the fusion of heterogeneous data sources. Then, it draws a profile for embedded vector dataset using recurrent neural networks, and achieves the anomaly detection of big data. Experimental results show that the proposed anomaly detection approach has the biggest value in our proposed mutual information metric, and it is obviously better than other anomaly detection algorithms in accuracy, false positive rate and false negative rate. The method of this paper can effectively detect the security threat in the power grid, and it is conducive to the safe and stable operation of power grids.

1 Introduction

So far, there is no agreement on the definition of big data. But the industry generally thinks big data has five characteristics: Volume, Variety, Value, Velocity, Veracity, or "5V" [1]. The veracity of data is the core of big data, and it is mean that the data records what happens in the real world. For example, the data in the e-commerce site records the user's purchase behavior, while various network devices record the access behavior of the network. By analyzing the big data of the e-commerce site, we can predict the user's buying patterns and make recommendations for the product. By analyzing the log of network equipment, we can analyze cyber attacks and take countermeasures for the network.

Electricity is an important infrastructure of the country, and it is necessary to maintain the normal life of people. The security of power grid is related to national economy and people's livelihood. The most important event in the safety of the power grid was the massive power outages in the United States and Ukraine, both of which caused huge economic losses and social impact [2]. With the comprehensive construction of intelligent, digital and information power grid, the amount of power grid data has exploded. The power industry enters the era of big data. At the same time, the development of power grid has become more and more integrated with the information network, so that the power grid faces many security threats both in and out of the world [3]. The rapid development of open source big data platform, such as Hadoop and Spark, has provided important technical support for the research of power big data.

In the research of power big data, the main domestic jobs are as follows. The literature [4] illustrated the characteristics of big data of smart grid. And it analyzed the general framework used in the analysis of big data of smart grid. The literature [5] analyzed the relationship between and among big data, cloud computing and smart grid. And this research discussed the key technologies that meet the development demand of power enterprises from the power of big data integration management, data analysis, data processing technology, the data presentation technology four aspects. The literature [6] investigated mining technology for the big data of power grid. In this research, the Bootstrap sampling, partitional clustering and hierarchical clustering are used to analyze the power load curve in the power distribution network. The literature [7] proposed a discretization scheme based on information accuracy by using Hadoop as a platform for big data analysis and processing, and studied the property entity recognition of power big data.

The main foreign research work on the big data of the power grid can be referred to in the review literature [8], [9] and [10]. A detailed overview of the research process of smart grid before 2011 (including) has been reviewed in the literature [8]. The literature [9] discussed the cyber security threats in the smart grid, and analyzed the challenges facing these threats. With the advent of the era of big data, the literature [10] made further elaboration on the security threat in smart grid from the perspective of data driven.

In this paper, we rely on the Hadoop [11] and Spark [12] big data processing platform, as well as the operation on the Spark platform Deeplearning4j [13] deep learning framework, and study the anomaly detection of security threats in power grid based on heterogeneous data sources in power big data. In order to integrate data from heterogeneous data sources, a deep restricted Boltzmann machine structure is adopted. In order to establish a profile for embedded vector dataset using recurrent neural networks, and achieve the anomaly detection of big data.

2 A Multilevel Deep Learning Method

2.1 Problem Description

There is a lot of heterogeneous data in the grid. According to the source of data, there are firewall logs, intrusion detection system logs, and logs generated by the business system, etc. According to the types of logs, there are traffic logs, configuration management logs, and security attack class logs, and so on. These logs are designed for specific needs, and they record different kinds of network information, user behavior, and system operations. Thus, these data are heterogeneous.

When the power grid is attacked by Advanced Persistent Threats (APT), it is difficult to analyze individual logs because of the latency and persistence of the attacks. However, all the attack marks of the attacker are recorded in various logs. If heterogeneous log information is integrated, it is possible to detect the security threats in the grid as soon as possible and take countermeasures for the network.

In this paper, we integrate heterogeneous data in different formats in the grid, and map to a unified embedded vector space by using deep restricted Boltzmann machine. Then, it draws a profile for embedded vector dataset using recurrent neural networks, and achieves the anomaly detection of big data. When the new data deviates from the portrait to a certain threshold, it is considered an anomaly data.

2.2 Heterogeneous data fusion based on deep restricted Boltzmann machine

Different logs of the electrical power system may contain the same fields, such as time, hosts, etc. It may also include the specific fields of the logging system, such as the possible attack types in the intrusion detection system, the unique user operation behavior of the business system, and so on. These logs contain different fields, and the single log sizes in different logging systems are not equal, so it is not appropriate to use the same fields to describe all the logs. In order to describe all types of logs with a uniform format, we use a multilevel deep restricted Boltzmann machine inserts different types of logging into a single vector space. maps different types of logs to a unified embedded vector space.

First, we convert any log into a binary vector. For a log containing n fields $\boldsymbol{x}(x_1, x_2, ..., x_n)$, the i th （$1 \le i \le n$） field is x_i, we use the binary representation for x_i, and we concatenate all of the binary fields of $\boldsymbol{x}$ into a binary value. For example, when $n=2$, $\boldsymbol{x}=(7,15)$, and both of the fields of $\boldsymbol{x}$ are represented by 8-bit binary representation. And the transformation of the binary is $\boldsymbol{x}'=$ 0000011100001111, the length of $\boldsymbol{x}'$ is $\boldsymbol{len}(\boldsymbol{x})=16$.

Next, we extend the logs with different lengths to the same length. When the length of the log is extended, the extended log length is the maximum log length for all logs, and we need to add 0 in the front of the binary vector if the length of the log is not sufficient. For example, there are two format logs, $x_1'=$0000011100001111, and the length of x_1' is 16. $x_2'=$00001243000001110000l111, and the length of x_2' is 24. Then, we extend the x_1' = 000000000000011100001111, and the length of expanded x_1' also is 24.

Then, we use the multilevel deep restricted Boltzmann machine [14] to map the extended log vectors $\boldsymbol{x}'$ with uniform length to the embedded vector space. And the embedded vector space of length is m ($m < len(\boldsymbol{x}')$). The structure of embedded mapping of vector with a multilevel deep restricted Boltzmann machine is shown in figure 1.

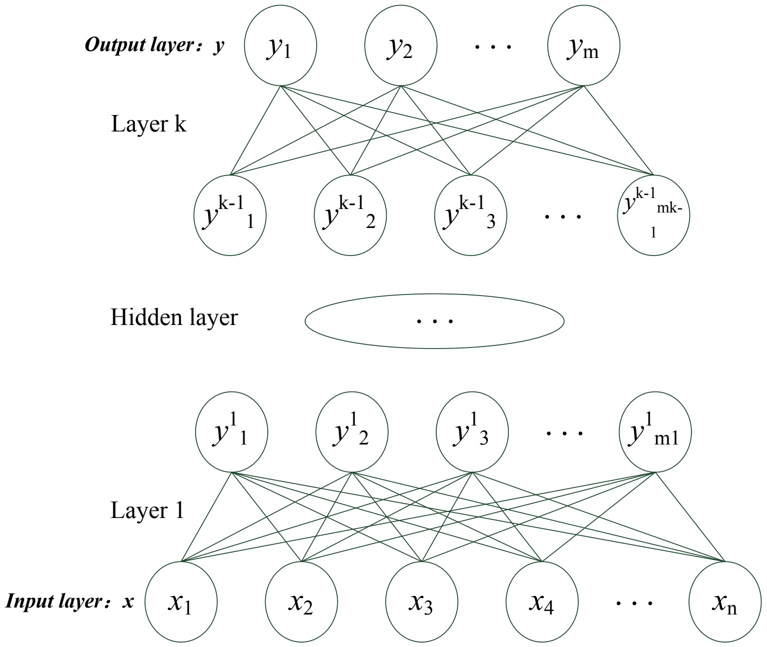

Fig.1. Embedded vector map of the multilevel deep restricted Boltzmann machine

The embedded vector mapping process of the multilevel deep restricted Boltzmann machine is depicted as follows in Fig. 1. The superscript of the letter in figure 1 is just for the sake of distinction, not a power function of a variable.

Here, input layer $\boldsymbol{x}(x_1, x_2, ..., x_n)$ is an extended n dimensional log binary vector. Each of these dimensions x_i ($1 \le i \le n$) is equal to 0 or 1. Output layer $\boldsymbol{y}(y_1, y_2, ..., y_m)$ is an embedded binary vector after mapping. Each of these dimensions y_i ($1 \le i \le m$) is also equal to 0 or 1.

The structure of embedded mapping of vector with a multilevel deep restricted Boltzmann machine is a k layer network. Each of these layers is a restricted Boltzmann machine. Take the first layer for example, the input layer of the restricted Boltzmann machine is $\boldsymbol{x}(x_1, x_2, ..., x_n)$, the hidden layer is $\boldsymbol{y}^1(y_1^1, y_2^1, ..., y_{m_1}^1)$, then they satisfy the following joint distribution function:

$$P(\boldsymbol{x}, \boldsymbol{y}^1; \theta) = \frac{1}{Z(\theta)} \exp(-E(\boldsymbol{x}, \boldsymbol{y}^1; \theta)) \tag{1}$$

Among them, the formula for the partition function $Z(\theta)$ is

$$Z(\theta) = \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}^1} \exp(-E(\boldsymbol{x}, \boldsymbol{y}^1; \theta)) \tag{2}$$

The energy function of the restricted Boltzmann machine $E(\boldsymbol{x}, \boldsymbol{y}^1; \theta)$ is

$$E(\boldsymbol{x}, \boldsymbol{y}^1; \theta) = \boldsymbol{\text{-}bx \text{-} c}^{\mathrm{T}} \boldsymbol{y}^1 \boldsymbol{\text{ - } x}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{y}^1 \tag{3}$$

The prior parameter is $\theta = (\boldsymbol{b,c,W})$.

In the joint distribution function $P(\boldsymbol{x}, \boldsymbol{y}^1; \theta)$, we can derive the conditional probability distribution of $\boldsymbol{y}^1$ as follows if $\boldsymbol{x}$ is known.

$$P(\boldsymbol{y}^1 \mid \boldsymbol{x};\theta) = \prod_i P(y_i^1 \mid \boldsymbol{x};\theta) = \prod_i \frac{1}{1+\exp(-\sum_j W_{ji}x_j - b_i)} \tag{4}$$

As mentioned above, the vector value $\boldsymbol{y}^1$ of the first hidden layer is calculated by the input vector $\boldsymbol{x}$. The vector value $\boldsymbol{y}^2$ of the second hidden layer is calculated by the input vector $\boldsymbol{y}^1$. Repeating the procedure until the vector value $\boldsymbol{y}^{\mathrm{k}}$ of the k th hidden layer is calculated and making the output value vector satisfy for $\boldsymbol{y} = \boldsymbol{y}^{\mathrm{k}}$.

2.3 Data portrait and anomaly detection

In the process of embedded vector mapping of different types of log data, the integration of heterogeneous data is realized by applying the multilevel restricted Boltzmann machine. We obtain the data portrait of log data by learning the embedded vector space. Thus, we can detect the anomaly data for future security events.

The vector in the embedded vector space is sorted by the timestamp of the original data, and the time sequence of an embedded vector is obtained. For these time series, the recurrent neural network [15] was used to model. Figure 2 is a structure diagram of the application of recurrent neural network structure for data portrait and abnormal detection.

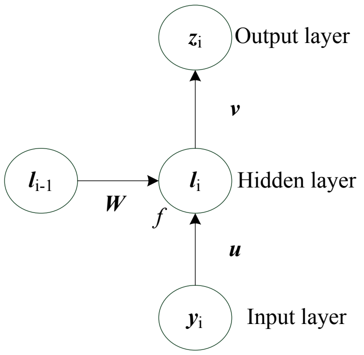

Fig.2. The structure diagram of recurrent neural network

Next, we introduce the process of data portrait and anomaly detection in Figure 2. Given embedded vector time series $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \ldots]$, the subscript represents the time sequence of the embedded vector. The input vector $\boldsymbol{y}_{\mathrm{i}}$ is combined with the last hidden layer vector $\boldsymbol{l}_{\mathrm{i-1}}$ to get the current hidden layer vector $\boldsymbol{l}_{\mathrm{i}}$. The calculation formula is as follows.

$$\boldsymbol{l}_i = f(\boldsymbol{u}\boldsymbol{y}_i + \boldsymbol{W}\boldsymbol{l}_{i-1}) \tag{5}$$

The output layer of the network adopts the linear structure, its formula is

$$z_i = \boldsymbol{v}\boldsymbol{l}_i \tag{6}$$

where $f(\cdot)$ is activation function.

This paper adopts the logistic regression function, that is

$$f(\boldsymbol{x}) = (1+\exp(-\boldsymbol{x}))^{-1} \tag{7}$$

Among of the vector $\boldsymbol{u}$, $\boldsymbol{v}$, and the matrix $\boldsymbol{W}$ is the parameters to be learned in the network.

When the input vector is $\boldsymbol{y}_1$, the hidden layer vector $\boldsymbol{l}_0$ is a randomly generated vector. The resulting hidden layer vector after calculation is the portrait of the data. For example, if the data sequence contains n embedded vectors, ultimate $\boldsymbol{l}_n$ is the portrait of the data. According to the obtained data portrait, the linear model of the output layer can be used to discriminate the abnormal situation of the data.

3 Experimental Results and Analysis

In order to verify the big data fusion and anomaly detection method proposed in this paper, we take the log data of a power company as an example to test and contrast the performance of the proposed algorithm. The experiment includes firewall logs, intrusion detection system logs, business system operation logs and database access logs.

During the log preprocessing, all log records are extended. We uniformly extend the log length to the maximum log length, and we need to add 0 in the front of the binary vector if the length of the log is not sufficient. When the logging length is extended, it is stored in the Hadoop file system. For the extension logs in the Hadoop file system, the Spark platform is applied to analyze the data to generate the embedded vectors. We apply deep learning framework Deeplearning4j to implement depth Boltzmann network on Spark platform.

In order to evaluate the proposed data portrait and anomaly detection algorithm, we compare our approach with other algorithms. It includes support vector machine (SVM), logistic regression (LR), naive Bayesian (NB), decision tree (DR), k-means, gaussian mixture (GM) and principal component analysis algorithm (PCA). And our approach is recorded as MDL(Multilevel Deep Learning). In the implementation of the algorithm, the embedded vector data is analyzed by using Deeplearning4j to obtain the portrait of the data. Other comparison algorithms use Spark's own algorithm. The experiment compared the abnormal detection effect of different algorithms, and the evaluation criteria had mutual information metric, accuracy, false positive rate and false negative rate.

In order to evaluate the effect of different algorithms, the following mutual information evaluation metric are defined when we using multiple methods for anomaly detection. Suppose we have k anomaly detection algorithms, every algorithm can get a prediction result. And the result set is $S=[R_1,...,R_k]$. If $R_1=(1,2)$, $R_2=(2,3)$, then $S=[1,2,2,3]$. With the idea of integration learning, the integration of multiple algorithms can better reflect the predicted results. We use the $S-R_i$ represent for deleting i result. This paper evaluates the evaluation result of R_i by the mutual information metric of R_i and $S-R_i$. Given random variables R and S, and the joint distribution of R and S is $p(r,s)$, the marginal distribution is $p(r)$ and $p(s)$ respectively. Then, the formula for the mutual information metric is as follows.

$$I(R,S)=\sum_{r\in R}\sum_{s\in S}p(r,s)\log\frac{p(r,s)}{p(r)p(s)} \tag{8}$$

Fig. 3 is the mutual information metric comparison result of multiple anomaly detection algorithms. It can be seen from the experimental results that the algorithm proposed in this paper has the largest mutual information metric. Next is the principal component analysis, again is the support vector machine. Principal component analysis (PCA) is a method of data compression, which creates a portrait by removing noise from the data. In this paper, we proposed the multilevel deep restricted Boltzmann machine is the essence of data compression through multi-layer network, and it can express more complex structures, so the anomaly detection result is better than principal component analysis method. The key and difficult point of support vector machine (SVM) is the construction of kernel function, and the kernel function of SVM in Spark is linear, so the effect of anomaly detection is not very good.

Next, we compare the accuracy, false positive rate and false negative rate of several anomaly detection algorithms through experiments, and their formulas are as follows.

$$accuracy=\frac{The\ correct\ number\ of\ exceptio\ ns\ in\ the\ test\ results}{The\ number\ of\ except\ ions\ in\ the\ test\ res\ ults} \tag{9}$$

$$false\ positive\ rate=\frac{The\ wrong\ number\ of\ exception\ s\ in\ the\ test\ results}{Number\ of\ test\ results} \tag{10}$$

$$false\ nega\ tive\ rate=\frac{The\ number\ of\ abnormal\ mis\sin g\ in\ the\ test\ results}{The\ number\ of\ except\ ions\ in\ the\ test\ res\ ults} \tag{11}$$

The comparison results of accuracy, false positive rate and false negative rate are shown in Figures 4, 5 and 6. As can be seen from these figures, the proposed algorithm not only has high accuracy, but also has very low false positive rate and false negative rate. Besides the proposed algorithm, the Gauss mixture method and the principal component analysis method also have higher accuracy in the accurate comparison. Because of the two methods are unsupervised learning, and unsupervised learning does not require prior definitions of exceptions. False positive rate and false negative rate are commonly used anomaly detection indicators in the field of information security. The method proposed in this paper has low false positive rate and false negative rate, so it can be applied to anomaly detection of power big data.

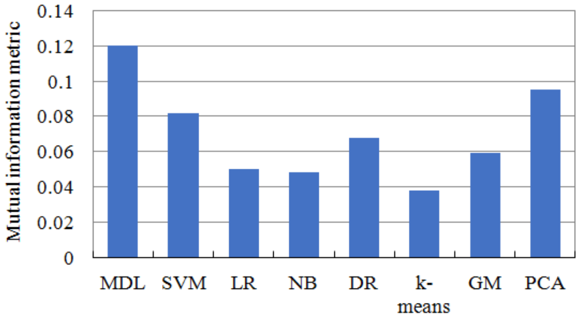

Fig.3. The comparison of mutual information metric

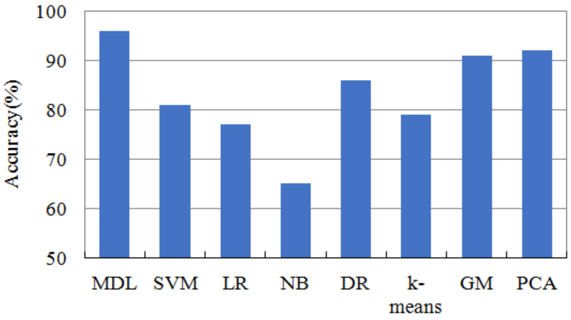

Fig.4. The comparison of the accuracy

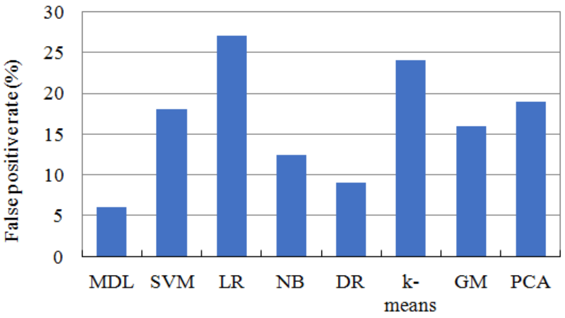

Fig.5. The comparison of the false positive rate

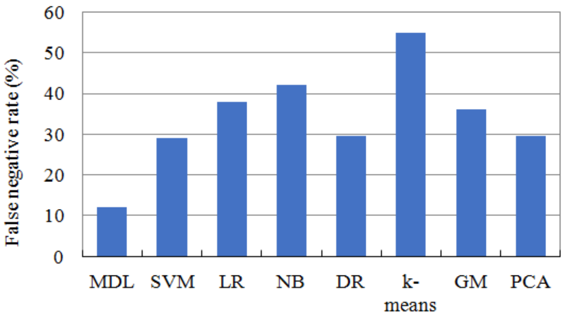

Fig.6. The comparison of the false negative rate

4 Conclusion

With the gradual integration of power grid and information network, information network brings more convenience and more security threats to the power grid. In this paper, the anomaly detection of security threat events is studied by using heterogeneous data source in power grid. We can map heterogeneous data in different formats to a unified embedded vector space with deep restricted Boltzmann machine, and achieve the fusion of heterogeneous data sources. Then, we can draw a profile for embedded vector dataset using recurrent neural networks, and achieve the anomaly detection of big data. Finally, the effectiveness of the proposed method is verified by experiments. Experimental results show that the proposed anomaly detection approach is obviously better than other anomaly detection algorithms in accuracy, false positive rate and false negative rate.

Acknowledgments

Most of this work was done during my working period at State Grid Shandong Electric Power Research Institute. The authors wish to acknowledge the anonymous reviewers and all co-workers for their helpful comments.

References

[1] Meng Xiaofeng, Ci Xiang. Big data management: concepts, techniques and challenges[J]. Journal of Computer Research and Development, 2013, 50(1): 146-169.

[2] GUO Qinglai, XIn Shujun, WANG Jianhui, SUN Hongbin. Comprehensive Security Assessment for a Cyber Physical Energy System: a Lesson from Ukraine's Blackout[J]. Automation of Electric Power Systems, 2016, 40(5): 145-147.

[3] ZHU Xiaoyan, FANG Quan. Study on Mechanism and Strategy of Cybersecurity in U.S. Electric Power Industry[J]. Electric Power, 2015, 48(5): 81-88.

[4] SUN Hongfei, GONG Lidong, ZHANG Haitao, WU Huijuan. Research on big data Analysis Platform for Smart Grid and Its Application Evolution[J]. Modern Electric Power, 2016, 33(6): 64-73.

[5] PENG Xiaosheng, DENG Diyuan, CHENG Shijie, WEN Jinyu, LI Zhaohui, NIU Lin. Key Technologies of Electric Power Big Data and Its Application Prospects in Smart Grid[J]. Proceedings of the CSEE, 2015, 35(3): 503-511.

[6] ZHANG Bin, ZHUANG Chijie, HU Jun, CHEN Shuiming, ZHANG Mingming, WANG Ke, ZENG Rong. Ensemble Clustering Algorithm Combined With Dimension Reduction Techniques for Power Load Profiles[J]. Proceedings of the CSEE, 2015, 35(15): 3741-3749.

[7] QI Jun, QU Zhaoyang, LOU Jianlou, WANG Chong. A kind of attribute entity recognition algorithm based on Hadoop for power big data[J]. Power System Protection and Control, 2016, 44(24): 52-57.

[8] Fang X, Misra S, Xue G, et al. Smart grid—The new and improved power grid: A survey[J]. IEEE communications surveys & tutorials, 2012, 14(4): 944-980.

[9] Wang W, Lu Z. Cyber security in the Smart Grid: Survey and challenges[J]. Computer Networks, 2013, 57(5): 1344-1371.

[10] Tan S, De D, Song W Z, et al. Survey of Security Advances in Smart Grid: A Data Driven Approach[J]. IEEE Communications Surveys & Tutorials, 2016.

[11] Shvachko K, Kuang H, Radia S, et al. The hadoop distributed file system[C]//Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. IEEE, 2010: 1-10.

[12] Zaharia M, Chowdhury M, Das T, et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing[C]//Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012: 2-2.

[13] Team D J D. Deeplearning4j: Open-source distributed deep learning for the JVM[J]. Apache Software Foundation License, 2.

[14] Fiore U, Palmieri F, Castiglione A, et al. Network anomaly detection with the restricted Boltzmann machine[J]. Neurocomputing, 2013, 122: 13-23.

[15] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.