V-CNN: Data Visualizing based Convolutional Neural Network

Guanxiong Feng, Bo Li, Mao Yang*, Zhongjiang Yan
Northwestern Polytechnical University
Xi’an China
*Corresponding author: yangmao@nwpu.edu.cn

***Abstract*—Recently, artificial intelligence technology has aroused wide attention and application worldwide, and is considered to be the next technology to create a new paradigm in the industry. The convolutional neural network (CNN), which is beneficial in fields such as imaging and voice analysis, is a type of representative algorithm of artificial intelligence. Increasing fields of study are introducing CNN into their research. However, CNN primarily handle image data, which is entirely different from the data form generated in other fields of study. Blindly processing the data by directly using CNN leads to incorrect training results or instances where training efficiency is too low. In this study, we use the idea of “making data fit model,” putting forward CNN based on data visualization, named V-CNN. V-CNN integrates the data visualization front before CNN model so that the data in the system is suitable for the CNN, which is, in turn, suitable for image recognition. This article further uses intelligent network intrusion detection as an example to verify the V-CNN performance. The results show that all the four categories of invasion of the AWID data set in each type of the recall rate is more than 99.8%, which is significantly better than that in the existing literature. To the best of our knowledge, this article is the first to propose V-CNN based on data visualization. V-CNN is general to handle data from almost all fields. Therefore, we call it “All can be image”.**

***Index Terms*—CNN, data visualization, deep learning, artificial intelligence, intelligent system**

I. Introduction

With the development of artificial intelligence and machine learning technology, artificial intelligence has become one of the most important technologies that will be affecting people’s lives in the future [1]. In the current era of big data, deep learning is extremely important in the field of artificial intelligence that processes big data. Scholars hope to introduce artificial intelligence to solve the problems they encounter in their sprcific field, which not only reduces the cost of human resources but also provides superior service. Moreover, with the continuous expansion of the system and function of every field, it has become increasingly complex and difficult to obtain and optimize system parameters through simple analysis and modeling method. Therefore, more and more areas urgently need to find means to optimize the system more intelligently. Among these methods, the convolution neural network (CNN) [2] is widely used in image, natural language processing, and other fields because of its sparse connection. For example, the R-FCN-3000 network, based on the fully connected convolutional network, has realized 3000 categories of detection [3], covers almost all categories in daily life, and its speed is up to 30 fps. CNN, in the field of natural language processing and in machine translation, not only significantly improves the running speed, but makes the network structure can more easily find information in the sentence structure [4].

CNN has been widely used in many industries. For example, medical diagnosis has introduced 3D-CNN to detect brain images to diagnose adhd [5]. The classification algorithm based on CNN is applied to mechanical fault diagnosis [6]. Although the introduction of CNN can be very beneficial in these areas, it always uses data in the form of images. In more general areas, however, the data do not have the format of image data (e.g., in the field of wireless networks). Thus, they mainly use non-deep learning algorithms. For example, the operation and maintenance of a cellular network in a wireless network and enterprise network WiFi, the network security problem can be solved using a machine learning approach [7], [8]; in the field of cognitive radio, machine learning can be used to solve the problem of a non-Markov environment and decentralized network [9]; in the field of WiFi indoor positioning, the machine learning method is used to realize accurate indoor positioning [10]; and in terms of network throughput prediction, machine learning is used to estimate the throughput and latency of its customers on a 2.4-Ghz WiFi channel [11].

Although the CNN in several fields has achieved great success, in view of the intelligent problem of image data, fault prediction diagnosis, system optimization control, and so on, for example, although the ability to use the general machine learning algorithm “intelligent,” but given that these algorithms have inherent defects, we hope to use the deep learning methods, such as CNN to deal with these problems. However, owing to the issue regarding data format, in the process of using CNN directly, there is blindness in the model optimization and model training, which leads to a poor training effect or low training efficiency.

It is difficult to use the CNN directly for general field intelligence. In this study, we use the idea of “making data fit model”, and propose a CNN based on image data, named V-CNN. The data collected by the system, after converting to image form, can not only fit with convolution neural network well but also convert complex data into a simple image .

978-1-5386-7946-3/18/$31.00 ©2018 IEEE

Our contribution:

1. To the best of our knowledge, this is the first paper to propose the idea of “making data fit model”. We propose a CNN based on visualization, called V-CNN.

2. We take network intrusion detection as an example, this study verified the intelligence system data visualization method. By using V-CNN, the method is proven to have a good classification effect, each type of classification accuracy is achieved as more than 99.8%, which is significantly better than the existing literature.

3. We emphasize that the V-CNN approach presented in this paper is versatile and can be easily applied in many other fields.

II. Introduction of V-CNN

As shown in Fig.1, the core idea of V-CNN is to process the data of the input CNN. The data visualization module is mainly responsible for representing the data in the form of images; the CNN module reference LeNet - 5 network [5] is mainly responsible for using graphical data to solve the original problem. The following section focuses on this aspect of the study.

A digital image is usually made up of pixels and is similar to the two-dimensional array of data structures. The colors of each pixel consist of R, G, B, three colors of gray image combination, said something, digital images of very strong correlation between two adjacent pixels.

The process of data visualization is divided into four steps: data cleaning, correlation analysis, pretreatment, and composition, and its main functions are summarized as follows:

- A. Data cleaning
- B. Correlation analysis
- C. Preprocessing
- D. Visualization

The basic principles of each step are described below.

A. Data cleaning

Normally, we can extract a large number of attributes from running the system, but there are a number of properties for the characteristics of the machine learning algorithm. Therefore, learning is meaningless, and we need to eliminate some of these properties in the first place. At the same time, some of the attributes in the system may be missing data, or some data cannot be represented in the requisite format. At this point, the data need to be cleaned, and the missing data are completed using the mean completion and specified value completion methods. The abnormal values are removed and complete data are obtained.

B. Correlation analysis

The correlation coefficient is calculated by

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}},$$

Where the term $Cov(X,Y)$ is the covariance of X and Y, $Var[X]$ is the variance of X, and $Var[Y]$ is the variance of Y.

Thus, the correlation coefficient between each pair of attributes is obtained for the data set in section A .

C. Preprocessing

Data sets processed in section A contain categorical and numeric data; therefore, the data need to be preprocessed.

- The categorical data are one-hot encoded, and a new data set is obtained. It is important to note that in the process of encoding, the number of categories generated by the code is not too large.
- The new data set is processed as a whole; all the data are evaluated on the basis of the attribute values, and the range is reduced to [0,255].

D. Visualization

Based on the number of attributes obtained in section C, an image size relative to the number of attributes is constructed. The image is a square of size

$$size = 3 \times \sqrt{\frac{x + \sum_{i=1}^{y} w_i}{x+y} \times y + x},$$

where

x: number of numeric variables

y: number of class variables

w_i: number of values of the ith class variable

- Numeric data are placed directly in any one of the new pixel RGB channels.
- The categorical data are encoded and assigned multiple dimensions; these are placed in different color channels of the same pixel or in different color channels adjacent to the image .
- The number of pixels required to represent each type of data is

$$\begin{cases} 1, w_i \leq 3 \\ \lceil \frac{w_i}{3} \rceil, w_i > 3, \end{cases}$$

For example, consider the following two situations:

1. For category number less than or equal to three (e.g., Fig.2, which shows three categories), the data from one dimension are encoded into three dimensions. The data are padded with zeros and expressed in standard form as [255, 0, 0]; the three dimensions of the image in the middle, shown with red marker pixels, represent the three channels corresponding to R, G, and B. The number in the upper right corner of the image is the number of dimensions after encoding.

2. For category number greater than three (e.g., Fig.3, which shows five categories), the required number of zeros are encoded, and the standardized data are expressed as [255, 0, 0, 0, 0]. When a pixel cannot be completely expressed in this manner, it is placed in additional color channels represented by the extra dimensions. As shown in the figure, the five dimensions include the two extra channels to represent the pixels.

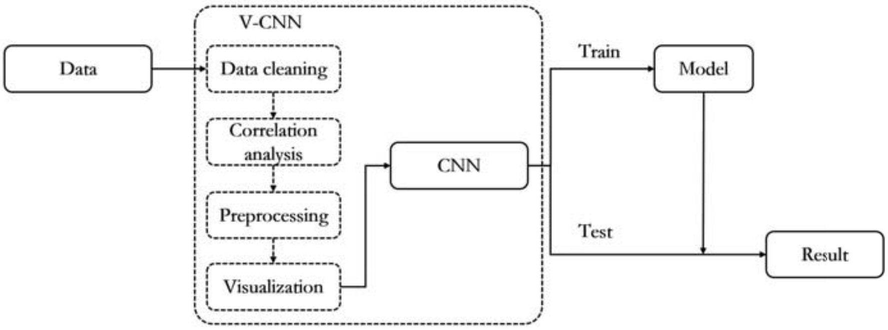

Fig. 1. V-CNN structure

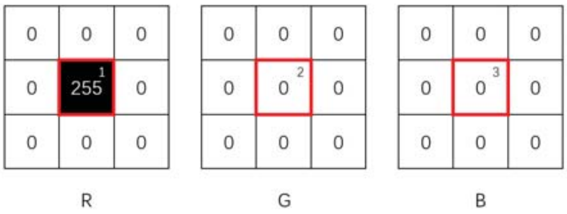

Fig. 2. Number of categories = 3

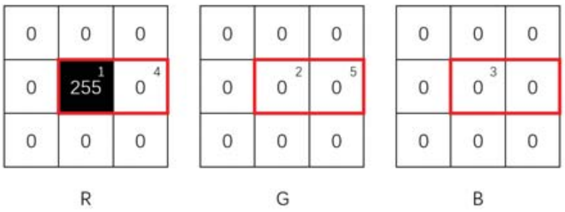

Fig. 3. Number of categories = 5

The basic image is initially obtained by placing the highly relevant data in the additional dimensions. The interpolation of this basic image not only increases the size but also produces an image that can be easily resolved by the human eye. As shown in Fig.4, the red mark (5 dimensions) and the blue mark (3 dimensions) belong to two different original attributes; their correlation meets the processing requirements and is placed in the adjacent dimensions.

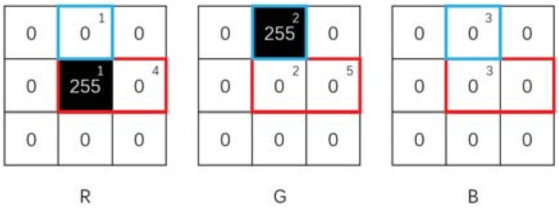

Fig. 4. Combination of two categories

III. Example verification based on network intrusion

In order to verify the validity of the data visualization method, we use the network intrusion detection data set to verify the performance of this method.

A. Data set introduction

The AWID [12] data set contains complete and simplified sets of data, which are discrete data collected in different environments using different devices. The data set provides a tag that corresponds to different attacks (ATK) and organizes the attack tags into four main classes (CLS) (including the normal class), as shown in TABLE I.

TABLE I

Name	Classes	Type	Number	Time(h)
AWID-ATK-F-Trn	10	train	162,375,247	96
AWID-ATK-F-Tst	17	test	48,524,866	12
AWID-CLS-F-Trn	4	train	162,375,247	96
AWID-CLS-F-Tst	4	test	48,524,866	12
AWID-ATK-R-Trn	10	train	1,795,575	1
AWID-ATK-R-Tst	15	test	575,643	1/3
AWID-CLS-R-Trn	4	train	1,795,575	1
AWID-CLS-R-Tst	4	test	575,643	1/3

Each data contain 155 attributes, and if an attribute is not applied to a particular record, the marker “?” is used in the record.

The data we use are the simplified version (-R) of the main class data (-CLS), which is divided into four categories, and the number of each category is shown in TABLE II:

TABLE II

AWID-CLS-R-Trn		AWID-CLS-R-Tst	
Class	Number	Class	Number
Flooding	48,484	Flooding	8,097
Impersonation	48,522	Impersonation	20,079
Injection	65,379	Injection	16,682
Normal	1,633,190	Normal	530,785

B. System model

The model used in this example is shown in Fig.6.

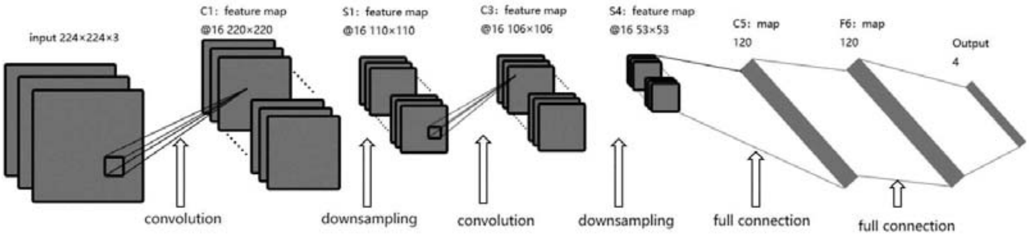

Fig. 5. CNN model

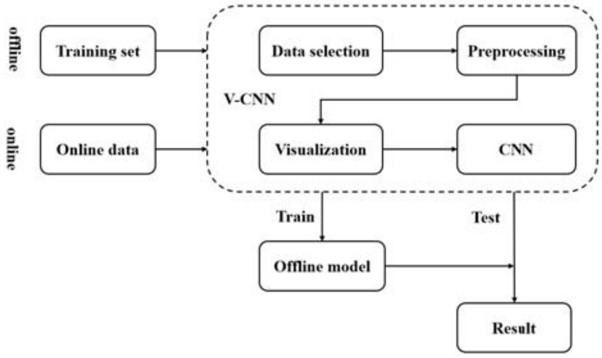

Fig. 6. System model

C. Data preparation

- Attributes for the original data set are selected, and 71 dimensional properties are analyzed for the correlation between pairs of attributes. The result is shown in Fig.7 (the correlation coefficient for the opening part of the figure is computed in cases where the denominator is zero).

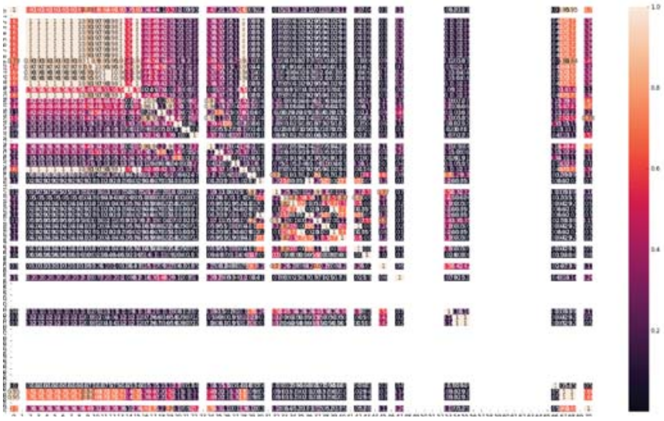

Fig. 7. Correlation coefficient

- Substitute the “?” attributes of a particular record with 0 to obtain a data set without missing values. The categorical values in the data set are then examined for their scope within the investigation, and the categories of the numeric and categorical data are obtained. The categorical data are then re-encoded with one-hot encoding, and 116 properties are obtained.
- The data obtained from the previous step are processed by image processing, and the data are obtained in an image format. The original data size is 1.10GB, and the data size obtained after image processing is 11.10GB.
- Figs.8-15 are randomly selected from the processed data; these can be observed by the human eye to detect differences among the different categories of data.

D. Model training

During the training phase, the training data of data set are input into the V-CNN; the CNN model is shown in Fig.5.

E. Training results and evaluation

Because of the large number of samples between categories, we use Recall to evaluate the effectiveness of the classification model.

$$R = \frac{TP}{N_+}$$

R:Recall rate, also known as TPR.

TP:Correctly predict the result as the number of positive classes.

N_+:The real category is the number of positive classes.

The TPR results for each category are shown in TABLE III.

TABLE III

class	total number	[17]	V-CNN
Flooding	16,682	100.00%	99.99%
Impersonation	8,097	73.78%	100.00%
Injection	20,079	99.99%	99.83%
Normal	530,785	22.00%	99.95%

IV. Discussion

In view of the current fields, there is an urgent need to use machine learning techniques to deal with the problems encountered by demand. The proposed V-CNN approach can solve the model selection problem and reduce the training speed and incorrect results. V-CNN integrates the data visualization front end with the efficient CNN so that the data in the system are suitable for the CNN, thus enabling the CNN adapt to more diverse scenarios. T he innovation technique

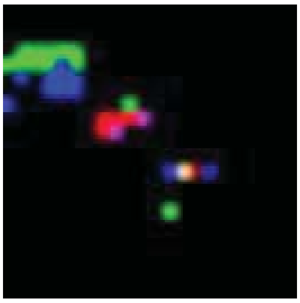

Fig. 8. Normal-train data

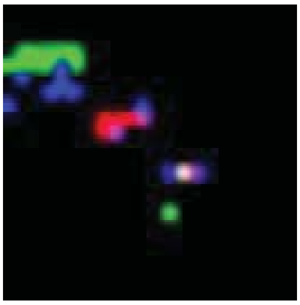

Fig. 9. Normal-test data

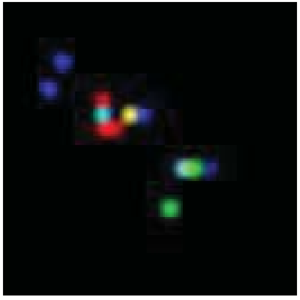

Fig. 10. Injection-train data

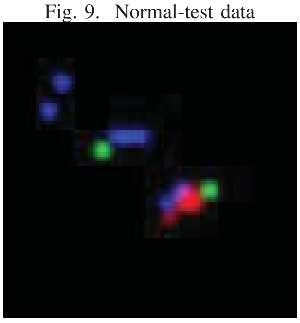

Fig. 11. Injection-test data

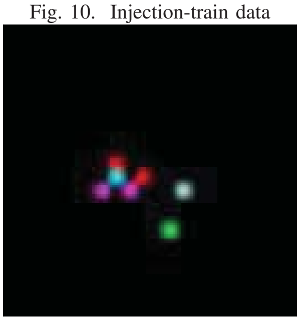

Fig. 12. Impersonation-train data

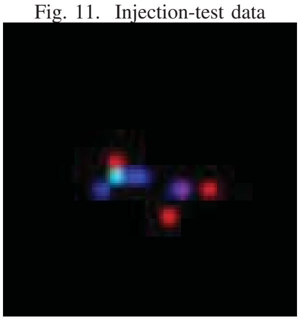

Fig. 13. Impersonation-test data

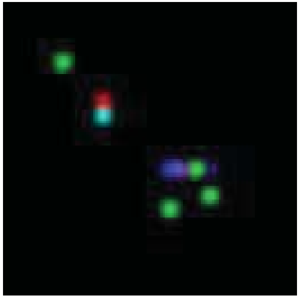

Fig. 14. Flooding-train data

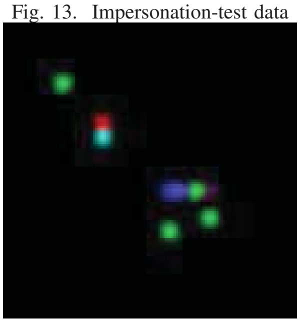

Fig. 15. Flooding-test data

of VCNN is "making data fit model," from the data rather than a network model. This study obtains the data of visual expression, relative to the direct use of data of CNN's method has better effect. We classified AWID network intrusion data sets to observe the performance of V-CNN. Compared with the existing experiments [13]–[16], the accuracy of classification was improved, and competitive results were obtained using the proposed approach. At the same time, except for the wireless network system in other fields, the V-CNN method converts data generated by the system to image processing and using CNN. Owing to the excellent image classification ability of CNN, they can be used to deal with the problems in other areas.

V. Summary and prospect

The current rapid development in all areas of intelligent systems will produce huge amounts of data. Owing to the intelligent problems, such as model selection and training, faced by the machine learning algorithms, this paper proposes a CNN based on data visualization. The V-CNN, from different data transformation to become suitable for use in the field of image form of CNN processing. In the AWID data set each type of the recall rate is more than 99.8% by using V-CNN, which is significantly better than that in the existing literature. In the future, we will apply V-CNN into wireless network operation maintenance.

VI. Acknowledgement

This work was supported in part by the National Natural Science Foundations of CHINA (Grant No. 61771390, No. 61501373, No. 61771392, and No. 61271279), the National Science and Technology Major Project (Grant No. 2015ZX03002006-004, and No. 2016ZX03001018-004), and the Fundamental Research Funds for the Central Universities (Grant No. 3102017ZY018).

References

[1] Chen X W, Lin X. Big data deep learning: challenges and perspectives[J]. IEEE access, 2014, 2: 514-525.

[2] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[3] Singh B, Li H, Sharma A, et al. R-FCN-3000 at 30fps: Decoupling Detection and Classification[J]. arXiv preprint arXiv:1712.01802, 2017.

[4] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[J]. arXiv preprint arXiv:1705.03122, 2017.

[5] Zou L, Zheng J, McKeown M J. Deep learning based automatic diagnoses of attention deficit hyperactive disorder[C]//Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference on. IEEE, 2017: 962-966.

[6] Liu Z, Wang J, Duan L, et al. Infrared Image Combined with CNN Based Fault Diagnosis for Rotating Machinery[C]//Sensing, Diagnostics, Prognostics, and Control (SDPC), 2017 International Conference on. IEEE, 2017: 137-142.

[7] Siadati H, Saket B, Memon N. Detecting malicious logins in enterprise networks using visualization[C]//Visualization for Cyber Security (VizSec), 2016 IEEE Symposium on. IEEE, 2016: 1-8.

[8] Yakasai S T, Zheng F C, Guy C G. Towards policy unification for enterprise network security[C]//Network Softwarization (NetSoft), 2017 IEEE Conference on. IEEE, 2017: 1-5. Security (VizSec), 2016 IEEE Symposium on. IEEE, 2016: 1-8.

[9] Bkassiny M, Li Y, Jayaweera S K. A survey on machine-learning techniques in cognitive radios[J]. IEEE Communications Surveys & Tutorials, 2013, 15(3): 1136-1159.

[10] Wang X, Gao L, Mao S, et al. DeepFi: Deep learning for indoor fingerprinting using channel state information[C]//Wireless Communications and Networking Conference (WCNC), 2015 IEEE. IEEE, 2015: 1666-1671.

[11] Kajita S, Yamaguchi H, Higashino T, et al. Throughput and delay estimator for 2.4 ghz wifi aps: A machine learning-based approach[C]//IFIP Wireless and Mobile Networking Conference (WMNC), 2015 8th. IEEE, 2015: 223-226.

[12] AWID, "Awid-wireless security datasets project data set," 2014.[Online]. Available: http://icsdweb.aegean.gr/awid/features.html

[13] Alotaibi B, Elleithy K. A majority voting technique for wireless intrusion detection systems[C]//Systems, Applications and Technology Conference (LISAT), 2016 IEEE Long Island. IEEE, 2016: 1-6.

[14] Aminanto M E, Kim K. Detecting impersonation attack in WiFi networks using deep learning approach[C]//International Workshop on Information Security Applications. Springer, Cham, 2016: 136-147.

[15] Thing V L L. IEEE 802.11 Network Anomaly Detection and Attack Classification: A Deep Learning Approach[C]//Wireless Communications and Networking Conference (WCNC), 2017 IEEE. IEEE, 2017: 1-6.

[16] Thanthrige U S K P M, Samarabandu J, Wang X. Machine learning techniques for intrusion detection on public dataset[C]//Electrical and Computer Engineering (CCECE), 2016 IEEE Canadian Conference on. IEEE, 2016: 1-4.

[17] Kolias C, Kambourakis G, Stavrou A, et al. Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset[J]. IEEE Communications Surveys & Tutorials, 2016, 18(1): 184-208.