Video Knowledge Discovery Based on Convolutional Neural Network

JinJiao Lin[1,4], ChunFang Liu[1], LiZhen Cui[2,5](✉), WeiYuan Huang[3], Rui Song[4], and YanZe Zhao[1]

[1] School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 10456CN, CO, China

[2] School of Software, Shandong University, Jinan, China
clz@sdu.edu.cn

[3] School of Marxism, Shandong University of Finance and Economics, Jinan 10456CN, CO, China

[4] School of Control Science and Engineering, Shandong University, Jinan 10422CN, CO, China

[5] Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, China

Abstract. Under the background of Internet+education, video course resources are becoming more and more abundant, at the same time, the Internet has a large number of not named or named non-standard courses video. It is increasingly important to identify courses name in these abundant video course teaching resources to improve learner efficiency. This study utilizes a deep neural network framework that incorporates a simple to implement transformation-invariant pooling operator (TI-pooling), after the audio and image information in course video is processed by the convolution layer and pooling layer of the model, the TI-pooling operator will further extract the features, so as to extract the most important information of course video, and we will identify the course name from the extracted course video information. The experimental results show that the accuracy of course name recognition obtained by taking image and audio as the input of CNN model is higher than that obtained by only image, only audio and only image and audio without ti-pooling operation.

Keywords: Knowledge discovery · TI-pooling · Convolutional nerve

1 Introduction

Online education platforms, forums, personal homepages, Weibo, various training groups, live broadcast platforms, etc. are all scattered with a large number of course video resources. Some of the course resources are normative, with course names, course knowledge points, and course evaluations. However, there are many video resources that are not standardized and are uploaded spontaneously by individuals on the Internet. Therefore, when searching for learning resources, the search may be incomplete due to the irregular description of the video resources, the irregularity or lack of naming, so it is

X. Zhang et al. (Eds.): CloudComp 2019/SmartGift 2019, LNICST 322, pp. 341–348, 2020.
https://doi.org/10.1007/978-3-030-48513-9_28

increasingly important to identify the courses name from the video for us to effectively use the Internet learning resources.

Identifying course names from video is a category of knowledge discovery, and knowledge discovery is the process of identifying effective, novel, potentially useful, and ultimately understandable knowledge from the data [1]. At present, most researches on knowledge discovery focus on text documents. For example, Wang et al. [2] proposed a convolutional neural network event mining model using distributed features, which uses word embedding, triggering word types, part of speech characteristics and multiple features of topic model to conduct event mining in text. Li et al. [3] used gated recurrent neural network (GRU) with attention mechanism to identify events in texts. However, few people study video, audio and other multimedia files. Video and audio generally contain rich knowledge, especially courses video, which is not only rich in content but also related to knowledge. At present, there are a large number of courses video on the Internet, and these course resources have the phenomenon that the course name does not correspond to the content or lacks the course name. Research on how to identify the course name in video will help learners make better use of learning resources. In recent years, deep neural networks have made remarkable achievements in many machine learning problems, such as image recognition [4], image classification [5] and video classification. However, identifying course names from courses video is still a challenge.

Based on the above analysis, this study uses a deep neural network model to collect video fragments of different courses from MOOC of China University and input the pictures and audio of course video into the model for training. After the completion of convolution and pooling, a TI-pooling operation is added. The TI-pooling operation can automatically find the best "standard" instance for training input, reduce the redundancy of learning features, and reduce the parameters and training time of the model. Ti-pooling operation will be introduced in detail in Sect. 3.2. In terms of the selection of activation function, we choose FReLU activation function. Compared with traditional ReLU function, FReLU function has the advantages of rapid convergence, higher performance, low calculation cost and strong adaptability. To verify the effectiveness of the method we used, we compared it with only images, only audio, and with images and audio but no TI-pooling model. Experimental results show that the performance of our method is better than the other three methods. Generally, this study offers at least three contributions as follows.

1. The CNN is applied to the course name recognition of course video.
2. The images and audio of course video are used as the input of the model to identify the name of course video.
3. The course name is automatically recognized from the course video.

2 Related Work

Massive data and poor knowledge lead to the emergence of data mining and knowledge discovery research. Knowledge discovery originates from artificial intelligence and machine learning. It is a new interdisciplinary subject with strong adaptability formed by the integration of machine learning, artificial intelligence, database and knowledge base. There are two main branches of knowledge discovery research at present, namely knowledge discovery based on database (KDD) and knowledge discovery based on literature (KDT).

Knowledge discovery based on database (KDD) can be defined as using data mining methods to identify valid, potentially useful, and ultimately understandable patterns from the database [7]. Knowledge discovery technology based on database is very mature and has been applied in many industries. For example, Wu Dan [8] used database knowledge discovery technology to predict employee turnover based on the basic information database of employees, and identified important factors that affect employee turnover, including the company's equity ownership, monthly salary, work environment satisfaction, work participation and so on. Xu et al. [9] developed the PhenoPredict system, which can infer the therapeutic effects of therapeutic drugs for diseases with similar phenotypes on schizophrenia from the knowledge base. Li Xiaoqing [10] studied bank data mining and knowledge discovery, and pointed out that data mining and knowledge discovery provide a basis for bank decision-making and customer relationship management. Knowledge discovery based on database has its limitation that it can only deal with structured data.

However, in the real world, knowledge does not all appear in the form of structured data in traditional databases, and quite a lot of knowledge is stored and presented in various forms, such as books, journals, newspapers, research papers, radio and television news, WEB pages, E-mail and so on. There is also a large amount of valuable information in these unstructured data sources. Therefore, data mining from these unstructured data sources to extract useful knowledge for users has become a new research hotspot in data mining, which is knowledge discovery based on text. For example, Kerzendorf [11] has developed a tool that can find similar articles based entirely on the text content of the input paper. By mining Web server logs, Novanto Yudistira et al. [12] found the correlation knowledge in the indicators of e-learning Web logs. Strong typed genetic programming (STGP) is used as a cutting edge technique to find precise rules and summarize them to achieve goals. The knowledge displayed may be useful to teachers or scholars, and strategies can be improved according to course activities to improve the use quality of e-learning. Enrique Alfonseca et al. [13] describes a combination of adaptive hypermedia and natural language processing techniques to create online information systems based on linear text in electronic formats, such as textbooks. Online information systems can recommend information that users may want based on their interests and background. Text-based knowledge discovery can process a variety of unstructured data. However, the current social data volume is growing exponentially. Traditional knowledge discovery technology based on database and opportunity text has been difficult to process massive data.

In recent years, deep learning technology has achieved good results in image recognition, image classification and audio processing, and promoted the application of knowledge discovery in video and audio. We use a two-channel convolutional neural network model to process the pictures and audio in the course video, and realize the automatic recognition of the course names without naming or non-standard naming of video from a large number of course video.

3 Methodology

3.1 The Network Architecture

For video knowledge discovery, CNN-related technology usually adopts multi-channel network structure, and has the following three main characteristics: first, weight sharing, second, local reception field (LRF), and third, pooling operation. CNN generally uses local information rather than global information.

The CNN model we use consists of two channels, picture and audio, which share parameters, The model consists of five convolution layers, each of which is followed by a maximum pooling layer after convolution. After the five convolution layers, a TI-pooling operation is conducted, and then the full connection layer is connected. The CNN model is shown in Fig. 1:

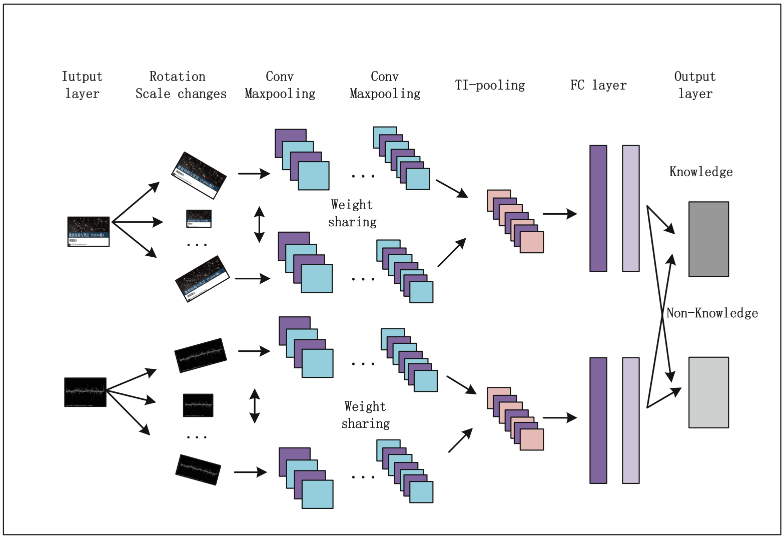

Fig. 1. CNN architecture

3.2 TI-Pooling Operation

In this study we represent features in a convolutional neural network as invariant transformations, which means that the machine learning algorithm only processes inputs that have not changed for some transformations. The most famous examples of general-purpose transformation-invariant features are SIFT (scale-invariant feature transform) [14] and its rotation-invariant modification RIFT (rotation-invariant feature transform) [15].

Because we did some processing on the sample before entering the data into the model, such as rotation, scaling and other changes to enhance the richness of the sample. The goal of TI-pooling is to carry out exhaustive search on the transformed samples to obtain the instance corresponding to the current response of the feature, and then only improve the performance of the feature with this instance.

As shown in Fig. 1, in the CNN model, the original sample and the transformed sample are input together. Instead of considering all the inputs as independent samples, but all the responses of the original sample and the transformed sample are accumulated and the maximum response is taken. Compared with data expansion, TI-pooling operation can learn fewer parameters without the disadvantage of losing relevant information after sample conversion, because it uses the most representative strength for learning.

Assume that, given a set of possible transformations Φ, we want to construct new features g_k (x) in such a way that their output is independent from the known in advance nuisance variations of the image x. We propose to formulate these features in the following manner:

$$g_k(x) = \max_{\phi \in \Phi} f_k(\phi(x)) \tag{1}$$

Where ϕ(x) is the input sample x according to a set of transform Φ transform after get the sample, f_kϕ(x) is the input sample characteristics of the model, and TI-pooling ensures that we use the best instance ϕ(x) for learning.

3.3 Activation Function

ReLU is an activation function widely used in CNN, but due to the zero-hard rectification, it cannot obtain the benefits of negative values. ReLU simply restrains the negative value to hard-zero, which provides sparsity but results negative missing. The variants of ReLU, including leaky ReLU (LReLU) [16], parametric ReLU (PReLU) [17], and randomized ReLU (RReLU) [18], enable non-zero slope to the negative part. It is proven that the negative parts are helpful for network learning. In this paper we use a new activation function called flexible rectified linear unit (FReLU), FRELU extends the output state of the activation function, adjusts the output of the ReLU function by adjusting the rectifying point, captures negative information and provides 0 features. It has the advantages of fast convergence, high performance, low calculation cost and strong adaptability [19].

As shown in Fig. 2(a), the input is x and the ReLU function is:

$$relu(x) = \begin{cases} x & if\, x > 0 \\ 0 & if\, x < 0 \end{cases} \tag{2}$$

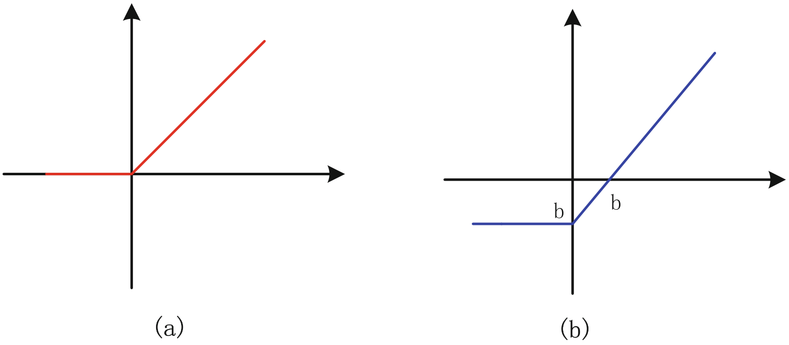

Fig. 2. ReLU and FReLU function images

The FReLU activation function we use is shown in Fig. 2(b). The function is:

$$frelu(x) = \begin{cases} x + b_l & if\ x > 0 \\ b_l & if\ x < 0 \end{cases} \tag{3}$$

where b_l is the l_{th} layer-wise learnable parameter, which controls the output range of FReLU. Note that FReLU naturaly generates ReLU when $b_l = 0$.

3.4 Loss Function

We choose the cross entropy function as the loss function of the training network. The specific function is:

$$\mathrm{C} = \frac{1}{n}\sum_{x}[ylna + (1 - y)\ln(1 - a)] \tag{4}$$

Where x is the input to the training, y is the output of the training, a is the actual output of each neuron, and n is the entire number of samples trained.

3.5 Back Propagation

Let $\nabla f_k(x)$ be the gradient of the feature $f_k(x)$ defined in Eq. 1 with respect to the outputs $\mathrm{O}(\cdot, \theta_j^{l-1})$ of the previous layer. This gradient is standard for convolutional neural networks and we do not discuss in details how to compute it [20]. From this gradient we can easily formulate the gradient $\frac{dg_k(x)}{df_k(x)}$ of the transformation-invariant feature $g_k(x)$ in the following manner:

$$\frac{dg_k(x)}{df_k(x)} = \nabla f_k(\phi(x)) \tag{5}$$

$$\phi = arg \max_{\phi \in \Phi} f_k(\phi(x)) \tag{6}$$

4 Experiments

The method we used is to input the images and audio of course video into the model. In order to verify the accuracy of the model, we conducted a comparative experiment with the model that only images, only audio, only images and audio but without TI-pooling. The detailed process of the experiment is shown below.

4.1 Data Set

We collected 15 video clips from MOOC of China University, processed the video into 324 pictures and 62 pieces of audio, and marked the picture and audio according to the course name. In order to increase the richness of the sample, we will make the picture and audio. After the rotation and scaling changes, 1296 pictures and 248 pieces of audio were obtained, and then 70% of the samples were selected into the training set, and 30% of the samples entered the test set.

4.2 Parameter Settings

The optimizer of the whole model uses the stochastic gradient descent method. The initial learning rate of the stochastic gradient descent method is set to 0.005, and the learning rate is attenuated by 1×10^{-6} after each update. The batch size of the data set read by the neural network during training is 16. The training data is transmitted to the neural network we use in the form of “sample-tag” for training the network model. The number of iterations is 10^3.

4.3 Experimental Result

To evaluate the effectiveness of the method we used, we compared the model using only images, using only audio, and using images and audio without increasing the TI-pooling operation. The experimental results show that the model we used is identified. Course names are more accurate than other methods. As shown in Table 1:

Table 1. Experimental result

Methods	Accuracy
Image only	61.3%
Audio only	57.7%
Image and Audio(without TI-pooling)	71.6%
Image and Audio(with TI-pooling)	77.4%

5 Conclusion

Identifying course names from a large number of non-naming or naming non-standard course videos can help learners improve the efficiency of resource retrieval and thus improve learning efficiency. In this paper we use a two-channel convolutional neural network model to process the image and audio signals of the course video. The framework adds a TI-pooling operation after all convolutional pooling layers. TI-pooling can Extract the most important features from the course video. The experimental results show that the CNN framework we use can better identify the course name from the course video, thus helping learners to better utilize the video learning resources on the Internet.

Acknowledgement. This work was financially supported by the Teaching Reform Research Project of Undergraduate Colleges and Universities of Shandong Province (Z2016Z036), the Teaching Reform Research Project of Shandong University of Finance and Economics (jy2018062891470, jy201830, jy201810), Shandong Provincial Social Science Planning Research Project (18CHLJ08), Scientific Research Projects of Universities in Shandong Province (J18RA136).

References

1. Fayyad, U, Shapiro, G.P., Smyth, P.: From data mining to knowledge discovery in databases [EB/OL]. http://www.kdnuggets.com/gpspubs/imag-kdd-overview-1996-Fayyad.Pdf. Accessed 22 Jun 2003
2. Wang, A., Wang, J., Lin, H., et al.: A multiple distributed representation method based on neural network for biomedical event extraction. BMC Med. Inform. Decis. Mak. **17**(S3), 171 (2017)
3. Lishuang, L., Yang, L., Meiyue, Q.: Extracting biomedical events with parallel multi-pooling convolutional neural networks. IEEE/ACM Trans. Comput. Biol. Bioinf. 1–1 (2018)
4. LeCunand, Y. Bengio, Y.: Convolutional networks for images, speech, and time series. In: The Handbook of Brain Theory and Neural Networks, vol. 3361, no. 10, p. 1 (1995)
5. Schmidhuber J. Multi-column deep neural networks for image classification. In: Computer Vision & Pattern Recognition (2012)
6. Karpathy, A., Toderici, G., Shetty, S., et al.: Large-scale video classification with convolutional neural networks. In: Computer Vision & Pattern Recognition (2014)
7. Peng, S.: Application of knowledge discovery in subject service. Northeast Normal University
8. Wu, D.: Prediction of employee turnover based on database knowledge discovery. Sci. Technol. Innov. 14 (2019)
9. Xu, R., Wang, Q.Q.: PhenoPredict: a disease phenome-wide drug repositioning approach towards schizophrenia drug discovery. J. Biomed. Inform. **56**(C), 348–355 (2015)
10. Li, X.: Decision analysis of banks based on data mining and knowledge discovery. Fintech Times 1, 56–59 (2014)
11. Kerzendorf, W.E. Knowledge discovery through text-based similarity searches for astronomy literature (2017)
12. Yudistira, N., Akbar, S.R., Arwan, A.: Using strongly typed genetic programming for knowledge discovery of course quality from e-Learning's web log. In: 2013 5th International Conference on Knowledge and Smart Technology (KST) (2013)
13. Alfonseca, E., Rodríguez, P., Pérez, D.: An approach for automatic generation of adaptive hypermedia in education with multilingual knowledge discovery techniques. Comput. Educ. **49**(2), 0–513 (2007)
14. Lowe, D.G.: Object recognition from local scale-invariant features. In: The proceedings of the seventh IEEE international conference on Computer Vision 1999, vol. 2, pp. 1150–1157. IEEE (1999)
15. Lazebnik, S., Schmid, C., Ponce, J., et al.: Semi-local affine parts for object recognition. In: British Machine Vision Conference (BMVC 2004), vol. 2, pp. 779–788 (2004)
16. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of ICML, vol. 30, no. 1 (2013)
17. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
18. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
19. Qiu, S, Xu, X, Cai, B.: FReLU: flexible rectified linear units for improving convolutional neural networks (2017)
20. Laptev, D., Savinov, N., Buhmann, J.M., et al.: TI-POOLING: transformation-invariant pooling for feature learning in convolutional neural networks (2016)