Short-Form video classification based on Gate shift module and Semantic embedding

Jun Tao¹,², Lixin Han²,³*, Jun Zhu³,⁴

¹,²,³ School of Computer and Information, Hohai University, Nanjing, Jiangsu Province, 211100, China
⁵e-mail: taoguang_0916@163.com, ⁴e-mail: zj_zijin@163.com
⁶e-mail: lhan@hhu.edu.cn

Abstract. Most of the existing video classification methods are based on large-scale data training, which can better realize the recognition and classification of known categories. However, data labelling is cumbersome and most things are unknown. Therefore, the existing video classification methods fall into a data bottleneck. This paper proposes a short video classification method based on GSM and semantic embedding. It uses super large-scale text information to assist the recognition process of the video classification model. This is an important development in the classification effect of knowledge categories. Specifically, this article expands the video classification method, adds category semantic embedding in the video feature extraction process, and trains to continuously fit the word vector of the corresponding category, and then uses semantic similarity to realize the classification of unknown categories. Multi-angle comparative experiments verify the effectiveness of this model, which can achieve good classification of unknown video categories.

1. Introduction

Intelligent understanding and recognition of video content has become the cornerstone of research in many fields [1]. The video contains not only the static features of the object, but also the motion features of the object, that is, the feature information in space and the feature information in time. Therefore, in order to achieve good classification and recognition of videos, it is necessary to extract these two kinds of feature information.

At present, in the field of video classification, a trained neural network model can be used to classify videos of known categories well. Commonly used methods include convolution of continuous static video frames to extract image features or manual extraction of features such as IDT [2], and even some research laboratories use spatio-temporal convolution architecture LTC [3], SMN [4] and C3D [5] to aggregate higher-level feature representations. On public datasets such as YouTube-8M, UCF-101, and CCV, the accuracy of video detection classification algorithms based on deep learning can reach more than 80%, and even the recognition rate of certain categories can reach more than 95%. However, deep learning often requires large-scale sample data to train a good enough model, but the cost of labelling video data is extremely expensive, and it is difficult for models trained on annotated video datasets to be good in actual application scenarios.

Video recognition research is gradually developing in the direction of unsupervised learning, and the difficulty is also increasing, and its research results are less compared with other directions. The research of unsupervised video classification attempts to use zero-sample learning and semantic clustering technology to realize the recognition of similar categories, such as the bag of words model,
E2E [6]. Recent studies have shown that category semantic integration can achieve aggregate representation of unknown categories within a certain accuracy, so as to perform video classification better, but this method relies on massive GPU resources, and training optimization is also an important goal. In order to make unsupervised video classification be widely used, we must tap the advantages of traditional algorithms, combine the characteristics of deep neural networks closely, and put forward more and more ideas to achieve greater breakthroughs.

Our contribution involves the following aspects:

- We have introduced the GSM [7] architecture, which aggregates the spatial information and timing information to better extract video features.
- We introduced a powerful corpus, and used the Word2Vec tool trained from the Google News Corpus containing 3 million words to map the categories to word vectors, and use the aggregation relationship of the categories to achieve the description of the unknown category, and the realization of the unknown category Recognition.
- In-depth analysis of the model and pre-training of this article is carried out. In a series of comparative verification experiments, the good performance of this model is explored.

2. Related work

Video classification is mainly completed by three parts: feature extraction, feature fusion and feature classification.

In terms of feature extraction, it is necessary to extract features from the two-dimensional space of the video frame, but also from the time dimension of the video itself. Feature extraction methods are divided into overall representation and partial representation. Feature fusion is also called feature coding. In order to make features have higher distinguishing ability, remove redundant information, and improve the computational efficiency of target recognition, the extracted features need to be fused. Mainstream feature coding methods include BOF [8] and FV [9]. In terms of feature classification, direct classification methods need to perform dimensionality reduction processing (PCA) on the extracted or encoded behavior features to reduce computational complexity and remove noise, and then use traditional classifiers such as KNN and SVM for classification. Time domain state space models mainly use Dynamic Time Warping (DTW) or Dynamic Space-Time Warping (DSTW) to align time dimensions of different scales, or use generative models (HMM) to distinguish models (CRF, MEMM) make classification judgments.

In the field of deep learning, in order to extract the spatiotemporal features of videos, researchers proposed three different ideas, namely the CNN-LSTM [10] method, the Two-Stream [11] method and the C3D [12] method (three-dimensional convolution kernel method). However, these three methods have a common flaw, that is, they cannot classify invisible classes. A recent effective method is to use the idea of zero-sample learning to solve this problem. Biagio [6] et al. The first end-to-end e2e trainable model for zero-sample action recognition is proposed. This model can be used to predict unknown categories by training known categories. However, in the experiment, the training set and the test set do not intersect, and it is not considered the real environment often has both known classes and unknown classes. Inspired by this idea, learn from its related work experience, improve and optimize its existing shortcomings and unrealized functions, based on the high-precision recognition of known classes, and realize the detection and classification of unknown classes.

3. Approach

In this part, this article will first formalize the problem and explain each step as follows. It also elaborates on visual feature extraction, word vector construction, classification function and loss function. **Training:** The training set $T_r = \{(x_1, l_1), (x_2, l_2), \ldots, (x_N, l_N)\}$ is composed of pairs of videos $x$ and the category labels of the videos $l$, where the number of training set samples is $N$. $L_r$ is collection of training set category labels. For any $l_i, i \in \{1, 2, \ldots, N\}$, $l_i \in L_r$. $L_t$ is a collection of test set category labels. **Inference:** For a given test data set $T_e$, it contains both known and unknown categories. Given a test video $x$, firstly converting it into the corresponding semantic embedding.
according to the visual encoder \( g_v \) and semantic encoder \( g_s \) of the model, and then measure the similarity between \( s \) and the word vector of the test set category label through a visual semantic compatibility function to realize reasoning.

3.1. Visual feature extraction
In order to solve the shortcomings of C3D networks that are requiring a large amount of parameters and high computational cost, many improved C3D nuclear decomposition methods have been proposed in recent years [13-16], but all these methods are to learn the structured kernel through hard-coded connection. The network structure is very solid and cannot be adjusted based on the training data. The features output from the previous block to the next block are all fixed networks. Therefore, an improvement was made to this point, and the Gate-Shift Module (GSM) was proposed.

GSM uses the time shift of TSM [16] for reference, and at the same time replaces the hard-wired channel split of GST [15] with a learnable gating block, and controls different network structures through the gating signal output by the gating block. The GSM structure is shown in figure 1.

![Figure 1. The GSM structure.](image1)

GSM combines TSM and GST, as shown in Figure 2. The first part is 2D convolution, as shown in the top box, the second part is the gating block, as shown in the green box, and the third part is shift, as shown in the white box on the left. The specific formulas are as follows.

\[
Y_1 = \tanh(W_1 \cdot X_1) \odot X_1; \quad Y_2 = \tanh(W_2 \cdot X_2) \odot X_2 \\
R_1 = X_1 - Y_1; \quad R_2 = X_2 - Y_2 \\
Z_1 = \text{shift}_\text{fw}(Y_1) + R_1; \quad Z_2 = \text{shift}_\text{bw}(Y_2) + R_2
\]

Where \( \text{shift}_\text{fw} \) and \( \text{shift}_\text{bw} \) represent forward and backward temporal shift respectively.

3.2. Word vector construction
This article uses Word2Vec to map category labels to word vectors. The basic idea is to map each word into a K-dimensional real number vector through training, and judge the semantic similarity between them by the distance between them. In this research, Word2Vec uses the Google News Corpus to train word vectors, which contains about 3 million most commonly used words.

3.3. Classification function
This module aims to infer the category of the test video by calculating the difference between the semantically encoded video feature of the test set and the category label word vector. This model is calculated by using KL divergence, the formula is as follows:

\[
L(s, h) = \sum_{i=1}^{M} s(x_i) \log \left( \frac{s(x_i)}{h(x_i)} \right)
\]

Where \( s \) is the semantic embedding of the test video \( x \), \( h \) is the word vector mapped by the category label using the Word2Vec tool, and \( M \) is the dimension of the word vector.

The test video is inferred into the category of the word vector with the smallest difference. At this point, for the test video, the model prediction output is:
\[ O(\cdot) = \arg\min_{l \in L_T} L(s, h_l) \]  

(5)

3.4. Loss function

In this paper, the model needs to make the semantic embedding corresponding to the training video continuously fit the word vector of the video label, so the model uses the mean square error loss function. In addition, according to the characteristics of the mean square error, the loss function will penalize larger errors. Therefore, using the mean square error loss function model will converge faster. The loss function formula used in this article is as follows:

\[ L = \sum_{l \in L_T} \| h_l - (g_y, g_s)(x) \|^2 \]  

(6)

4. Experiments

4.1. Experimental environment

The experimental configuration is: CPU: Intel (R) Xeon (R) CPU e5-2660 V4; Main frequency: 2.00GHz; The memory is 32g, and the processing is accelerated with the help of two Tesla V100 PCIe 32GB graphics cards (V100 Support 16-bit floating point model training).

4.2. Datasets and parameter settings

The datasets used in this experiment were public video datasets UCF101 and HMDB51. UCF101 has 13320 videos, which are divided into 101 categories. HMDB51 contains 6766 samples. The training dataset was 50 categories selected from UCF101 dataset by clustering method, and the test datasets were 50 categories of video data randomly selected from UCF101 and HMDB51. Figure 3 shows the results of visualizing the category embedding of experimental training dataset and test dataset.

Figure 3. Training and test classes, UCF101 and HMDB51 visualization of Word2Vec embeddings. The red dots represent the training data classes we selected after clustering UCF101, the blue dots represent the video category that does not participate in training, the green crosses represent test classes, and some random UCF101 classes are not shown in the figure.

Implementation details: (1) Data pre-processing operation used the sparse sampling method in the TSN network; (2) The non-C3D network used ImageNet pre-trained Inception-V2 as the back bone; (3) Adam algorithm was used for training, the learning rate was initially set to 0.001, and each time it was reduced to 1/10 of the original learning rate; (4) Used SGD momentum=0.9; (5) Used gradient clipping and batch normalization before each nonlinear layer.
4.3. Experimental results and analysis

We compared the current mainstream Action2Vec, TARN, URL and E2E method in Table 1. The algorithms were trained on the same training dataset and verified on the same test dataset. We repeated this process 3 times and average the results.

It could be seen from Table 1 that our model and the URL model have similar experimental results and are better than Action2Vec model, TARN model and E2E model. But the URL model used a very deep ResNet-200, and the parameters far exceed other algorithms. As the number of layers increases, the features extracted by the network become more abstract. The semantics are high-dimensional features, so it can be speculated that it will be easier to learn the mapping between visual features and semantic features in high dimensions. Follow-up research will follow this line of thinking.

| DataSet | VisualFeat | UCF101 Top-1 | UCF101 Top-5 | HMDB51 Top-1 | HMDB51 Top-5 |
|---------|------------|--------------|--------------|--------------|--------------|
| Action2Vec | C3D        | 41.01        | --           | 22.23        | --           |
| TARN    | C3D        | 42.23        | --           | 26.11        | --           |
| URL     | ResNet-200 | **48.48**    | **56.25**    | 33.60        | 39.33        |
| E2E     | C3D        | 43.17        | 54.06        | **38.73**    | 43.21        |
| Ours    | GST        | 47.78        | 55.65        | 36.89        | **43.64**    |

Table 2 Model ablation experiment. Shows our original reference method and our improved and optimized results

| DataSet | VisualFeat | UCF101 Top-1 | HMDB51 Top-1 |
|---------|------------|--------------|--------------|
| E2E     | C3D        | 41.19        | 32.25        |
| Ours-GST| C3D        | 43.23        | 36.13        |
| Ours-KL | GST        | 45.93        | **37.06**    |
| Ours    | GST        | **47.78**    | 36.89        |

Table 2 showed that using KL divergence as the classification function of the model had a partial improvement in the classification accuracy of the model compared to the original model that directly used the cosine value. Although the accuracy of using GST was not significantly improved compared to the original model, it could greatly alleviate the large amount of parameters and high computational cost of using the C3D network while effectively extracting the timing characteristics of the video. Compared with the original model, the model in this paper improved the accuracy of partial classification while reducing the parameters of the model, making the model much more practical.

5. Conclusion

Video content understanding and analysis is a hot issue in the field of computer vision. Based on unsupervised video classification and large-scale data labelling, through the observation of E2E model and C3D network, we proposed a short video classification method based on GSM and semantic embedding. Based on the design and implementation of large-scale category semantic database, this method effectively explored the recognition of a wide range of unlabelled unknown video classes. The final fusion model could further improve the recognition performance. Difference between visual feature manifold and semantic feature manifold is an important factor affecting the accuracy of video classification model based on zero sample learning. Therefore, reducing the difference between visual feature manifold and semantic feature manifold is the focus of follow-up research.
Acknowledgments
This work was supported by the Science Research Project of Nanjing University of Science and Technology Zijin College (No. 2020ZRKX0401001) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China, No. 20KJB520035.

References
[1] Hainxin H., Ruipeng W., Xiaoyang Liu. (2020) Overview of human behavior recognition technology based on 3D convolution. computer Science, 47(S2):139-144.
[2] Wang, H., Schmid, C. (2013) Action recognition with improved trajectories. In: IEEE International Conference on Computer Vision. Sydney. pp. 3551-3558.
[3] Feichtenhofer, C., Pinz, A., Wildes, R.P. (2017) Spatiotemporal multiplier networks for video action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. Honolulu. pp. 4768-4777.
[4] Varol, G., Laptev, I., Schmid, C. (2017) Long-term temporal convolutions for action recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6): 1510-1517.
[5] Tran, D., Bourdev, L., Fergus, R., et al. (2015) Learning spatiotemporal features with 3d convolutional networks. In: IEEE International Conference on Computer Vision. Santiago. pp. 4489-4497.
[6] Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., & Chalupka, K. (2020). Rethinking zero-shot video classification: End-to-end training for realistic applications. In: IEEE Conference on Computer Vision and Pattern Recognition. Seattle. pp. 4613-4623.
[7] Sudhakaran, S., Escalera, S., & Lanz, O. (2020). Gate-shift networks for video action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. Seattle. pp. 1102-1111.
[8] Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York. pp. 2169-2178.
[9] Perronnin, F., & Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In: IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis. pp. 1-8.
[10] Donahue, J., Anne Hendricks, L., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston. pp. 2625-2634.
[11] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199.
[12] Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In: IEEE International Conference on Computer Vision. Honolulu. pp. 5533-5541.
[13] Tran, D., Wang, H., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 6450-6459.
[14] Tran, D., Wang, H., Torresani, L., & Feisli, M. (2019). Video classification with channel-separated convolutional networks. In: IEEE International Conference on Computer Vision. Seoul. pp. 5552-5561.
[15] Luo, C., & Yuille, A. L. (2019). Grouped spatial-temporal aggregation for efficient action recognition. In: IEEE International Conference on Computer Vision. Seoul. pp. 5512-5521.
[16] Lin, J., Gan, C., & Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In: IEEE International Conference on Computer Vision. Seoul. pp. 7083-7093.