A Survey on Video Classification Methods Based on Deep Learning

Qiuyu Ren, Liang Bai, Haoran Wang, Zhihong Deng, Xiaoming Zhu, Han Li and Can Luo

ABSTRACT

With the development of multimedia and communication technology, the amount of multimedia data, for example, video data has been rapidly emerging. How to process these data accurately and efficiently has attracted more and more researchers' attention. Video classification is an important part of video data processing. After deep learning method successfully applied in image and audio fields, the study of video classification has gradually shifted to deep learning direction. This paper summarized the video classification methods based on deep learning and analyzed the differences in performance of typical algorithms, meanwhile, summarized the commonly used video classification data sets.

KEYWORDS

Video classification, Deep learning, Data sets, Compact models.

INTRODUCTION

With the development of video recording and playback devices, more and more people choose to shoot video and upload it to the network to share their life. People prefer to get information by watching videos over reading books, listening to audios and other forms. In recent years, the evolvement of 4G/5G has spawned a large number of mobile video apps. The short video shooting and sharing APPs, including TicTok, have turned the social network from sharing text and pictures to videos. Under these circumstances, the amount of video data has increased significantly. Video data have the characteristics of large data volume and complex data modality. A large amount of video data carries a huge amount of information. Faced with such a huge and growing amount of data, figuring out how to deal with it quickly and effectively has become an important issue for researchers.

Video classification is an important part of the video processing field. The purpose of which is to classify videos into preset categories based on video content. Video classification is mainly divided into two parts in: feature extraction and classification. Traditional video classification methods rely on artificially designed features and typical machine learning methods to learn behavior recognition and event detection. However, the video itself has a large
amount of data and complex data modalities. It is difficult to obtain good results by using artificially designed features and machine learning methods. Since deep learning has achieved good results in computer vision tasks such as image classification in 2012, deep learning methods have gained more and more researchers’ attention[1][2]. Deep learning relies on deep neural network back propagation to autonomously extract the depth features of input video. The features obtained this way have better generalization and representation, and contain rich semantic information. Therefore, deep learning has become the main research direction in computer vision issues.

With the rise of deep neural networks, especially the successful application of CNN, LSTM[3], GRU[4] in video classification, the traditional methods have gradually faded out of people's horizons.

DATA SETS

Deep learning is a data-based method. Many scientific research institutions have spent a lot of manpower and resources to collect and label video data sets for research in video related fields. Commonly used data sets are YouTube-8M[5], HCF-50[6], HCF-101[7], HMDB51[8] and so on. Among them, the small-scale data sets are Weizmann, KTH, and Hollywood, and the total amount and video types are small but are well-labeled. Medium data set includes more than 50 video type, such as UCF101, Thumos'14 and HMDB51. The large dataset such as the YouTube-8M dataset (collected by Google), Sports-1M, ActivityNet, Kinetics, etc.

More detailed information is summarized in table I.

<table>
<thead>
<tr>
<th>Name</th>
<th>Video amount</th>
<th>Categories amount</th>
<th>Published year</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weizmann</td>
<td>81</td>
<td>9</td>
<td>2005</td>
<td>static</td>
</tr>
<tr>
<td>KTH</td>
<td>2361</td>
<td>6</td>
<td>2004</td>
<td>static</td>
</tr>
<tr>
<td>Hollywood</td>
<td>430</td>
<td>8</td>
<td>2008</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Hollywood2</td>
<td>3669</td>
<td>12</td>
<td>2009</td>
<td>Dynamic</td>
</tr>
<tr>
<td>UCF50</td>
<td>6676</td>
<td>50</td>
<td>2012</td>
<td>Dynamic</td>
</tr>
<tr>
<td>HMDB51</td>
<td>6474</td>
<td>51</td>
<td>2013</td>
<td>Dynamic</td>
</tr>
<tr>
<td>UCF101</td>
<td>13320</td>
<td>101</td>
<td>2012</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Thumos’14</td>
<td>18394</td>
<td>101</td>
<td>2014</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Youtube-8M</td>
<td>8264650</td>
<td>4800</td>
<td>2016</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Sports-1M</td>
<td>1133158</td>
<td>487</td>
<td>2014</td>
<td>Dynamic</td>
</tr>
<tr>
<td>ActivityNet</td>
<td>27901</td>
<td>203</td>
<td>2015</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Kinetics</td>
<td>306245</td>
<td>400</td>
<td>2017</td>
<td>Dynamic</td>
</tr>
</tbody>
</table>

EVALUATION NORM

A commonly used evaluation norm for video classification algorithms is mAP (mean Average Precision). There are two commonly used indicators: precision(P) and recall(R). The calculation formulas of the two are as follows:
\[ P = \frac{TP}{TP + FP} \]  
\[ R = \frac{TP}{TP + FN} \]

In which:

<table>
<thead>
<tr>
<th>Real</th>
<th>Predicted</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td></td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>False</td>
<td></td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Generally the precision ratio refers to the proportion of the classified tasks that are correctly classified, and the recall rate is the number of predicted true samples out of all the real true samples. Normally precision and recall won’t be both good, the mAP comprehensively characterizes the precision and recall. The larger the value, the higher the accuracy of the classification. mAP is calculated as by:

\[ mAP = \frac{\sum_{m=1}^{M} AveP(m)}{M} \]

M is the total number of categories. AveP(m) is the Average Precision(AP). According to PASCAL VOC 2010[9], AveP(m) is calculated as follow:

- Calculate the precision-recall curve, i.e. the P-R curve, with the precision as y axis, the recall as x axis and precision is monotonically decreasing. As for each recall r, the corresponding precision is set to the maximum value of the precision in all the \( r' \geq r \).
- Calculate the area under the P-R curve as the AP for the corresponding category.

**VIDEO CLASSIFICATION BASED ON DEEP LEARNING**

Compared with image classification task, the video classification task has a larger amount of data to be processed, and the data modal is complicated. In addition to spatial information, video has temporal information as well, which is also important for the understanding of video content. Therefore, how to effectively use the temporal information of the video and how to combine the temporal information with the spatial information is the main problem faced by the video classification task. Considerable researches have been done from two main aspects: supervised learning and unsupervised learning according to whether the data is labeled.

**VIDEO CLASSIFICATION BASED ON SUPERVISED LEARNING**

With the successful application of deep learning methods in image and audio field, depth features for video classification have also received more and more attention. Models such as two-stream convolutional neural networks and 3D
Convolutional neural networks have been proposed and are becoming mainstream methods. The two-stream convolutional neural network was first proposed in 2014 by Simonyan et al. [10]. It models the dynamic and static information of video data separately. The static features are extracted from the video frames, and the dynamic features are extracted from the optical flow data. The network structure is as shown in Figure 1.

The accuracy of the two-stream convolutional neural network defeated the manual features for the first time on the video classification standard data sets UCF101 and HMDB51. Due to the input of dynamic information extraction stream of two-stream neural network is optical flow extracted from several consecutive frames, long-term dynamic information is nearly impossible to be obtained. Wang et al. [11] proposed a temporal segment networks (TSN). Based on the two-stream convolutional neural network, TSN divides the video with longer duration into several segments, and each segment is sparsely sampled and classified, and the final classification results are obtained by merging the segment classification scores. Meanwhile, TSN improves the computational efficiency by sparse sampling of video.

Different information in video has different influence on classification results, and the dynamic and static information in the video are related to each other rather than separated. Based on this thought, Peng introduced the attention mechanism into the two-stream convolutional neural network [12]. In the process of extracting static and dynamic features, the key regions and key frames are located separately to obtain more representative features. Then the dynamic and static features are optimized and adapted jointly to improve the classification accuracy.

The input optical flow of dynamic feature extraction in the two-stream network structure was first proposed by Gibson in 1950. It is a method to find the corresponding relationship between the previous frame and the current frame according to the change of the pixels in the time domain and the correlation between adjacent frames in the image sequence, so as to calculate the motion information of objects in video between adjacent frames. The calculation of optical flow consumes lots of computational and storage resources. For this reason, some researchers consider using other information instead of optical flow. Sun [15] directly calculate the pixel-level spatial-temporal gradient on the feature map, so that the convolutional neural network can obtain motion information directly from the image frame. In the case of comparable performance, the optical flow guided feature model proposed by Sun is 15 times faster than the two-stream convolutional neural network.

As the name implies, the 3D convolutional neural network extends the convolutional neural network into a 3-dimensional space, which stacks several frames into a cube, and uses a 3D convolution kernel in the cube [13]. Extend the...
convolutional neural network to 3D and increase the time dimension so that it can model spatial-temporal information jointly. In this structure, each feature map in the convolutional layer is connected to several adjacent frames in the upper layer, in which way the gotten feature contains motion information, as shown in Figure 2 C3D is a typical model in 3D convolutional neural networks with a convolution kernel of $3\times3\times3$. However, C3D relies on well-labeled video data sets for pre-training. In order to alleviate the dependence of the 3D convolutional neural network on data sets, Qiu et al.[14] proposed a pseudo 3D residual network, which decomposed the $3\times3\times3$ convolution kernel into a $1\times3\times3$ spatial convolution. With a time domain convolution of $3\times1\times1$, these 2D convolution kernels can be pre-trained on the image dataset.

Inspired by the successful application of recurrent neural networks in natural language processing field for the processing of sequence information, researchers have applied recurrent neural networks to process video data. Ng et al.[16] put the output of the convolutional neural network to the LSTM in time series for time series modeling, which confirms the feasibility of using LSTM to achieve video feature fusion.

In addition to the above methods, researchers also combined some models to obtain better classification results. Carreira et al.[17] combined the two-stream method with the 3D convolutional neural network and proposes an I3D model. Although the 3D convolution can extract time information by itself, adding optical flow information can improve performance.

![Figure 2. How 3D convolution works. The time dimension of the 3D convolution in the figure is 3, which means 3 frames are combined together into one cube. We can see that the value of a pixel of a convolution map is obtained by convolving the local receptive field of the same pixel of three consecutive frames of the previous layer.](image)

All the methods we covered only pay attention to the characteristics of the image modality in the video, and ignoring the audio and text information attached to the video. But in reality, human brain can detect events or conceptual inspiration in video by visualizing spatial-temporal data and listening to audio and reading its description. Reference[18] proposes a multimodal deep learning framework using different sources of information including visual, audio and textual data. It solves the problem that many modal data in multimedia resources cannot be fully utilized. Especially in the case of a modal data is damaged or missing, using multiple data types can significantly improve the final detection and retrieval performance.

**VIDEO CLASSIFICATION BASED ON UNSUPERVISED LEARNING**

The use of unsupervised learning methods to integrate spatial and temporal context information is a promising approach to discovering and describing video structures. Unsupervised classification, also known as cluster analysis, is to measure the similarity of features, and the attributes of the categories are determined by visual interpretation or field investigation after classification[19]. In reference[20], the feature composition maps obtained from multiple video-
related self-supervised learning tasks are used to measure similarity, and the result guides the training procedure of a lighter classification network. Compared with the features obtained by the supervised learning method, the features obtained by unsupervised method is weaker, but can make full use of various video data.

The supervised learning classification model trained on labelled data is difficult to expand to new categories. In order to solve this problem, zero-shot video classification has gradually become a hotspot[21][22]. The zero-shot classification method is mainly divided into object attribute based prediction and word vector based prediction. It aims to establish the connection between given labels and unknown label. By mapping the video features of the unknown category and the training set features to the same space and measuring their similarity measure, we can get the classification result of new category.

In addition to classification accuracy, researchers also think highly of the compactness of the model. Whether the model can be as compact as possible without cutting down performance is a prerequisite for the model to be practical. The compact model can achieve by repeatedly applying small weight matrix operations to all frames in the video[23]. In this way models will take up very little in memory, the FLOPs (floating point operations) are still large. Shweta et al.[24] used the teacher-student model compact a model by training a see-very-little student model with a see-it-all teacher model. The student model requires only a small amount of memory and FLOPs. It also maintains high classification accuracy. In Shweta's work, the student network can reduce inference time by 30%, FLOP by about 90%, and performance degradation is negligible.

CONCLUSION

Nowadays video classification method based on deep learning is still a hotspot in video understanding field. Common methods are two-stream convolutional neural network and 3D convolutional neural network method. These methods make use of both spatial and temporal information in video. To solve the classification of new categories, the researchers employ the zero-shot method to study the association between video content and labels. After obtaining better classification accuracy, the researchers considered a compact model with smaller memory footprint and smaller FLOPs. However, video classification task remains a challenge. One of the difficulties is that a better result relies on large-scale and well-marked data sets for model training, but the collection and labeling of data sets requires a lot of human and material resources. Another difficulty can be eliminating the influence of blur and camera motion on classification results. How to get more representative video features is also a concern of researchers.

REFERENCES

1. Dahl GE, Yu D, Deng L, Acero A. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition [J].