비디오 분류에 기반 해석가능한 딥러닝 알고리즘

김택위, 조인휘
한양대학교 컴퓨터소프트웨어학과
jinzewei1996724@hanyang.ac.kr, iwjoe@hanyang.ac.kr

An Explainable Deep Learning Algorithm based on Video Classification

Jin Zewei, Inwhee Joe
Department of Computer Science, Hanyang University

요 약

The rapid development of the Internet has led to a significant increase in multimedia content in social networks. How to better analyze and improve video classification models has become an important task. Deep learning models have typical "black box" characteristics. The model requires explainable analysis. This article uses two classification models: ConvLSTM and VGG16+LSTM models. And combined with the explainable method of LRP, generate visualized explainable results. Finally, based on the experimental results, the accuracy of the classification model is: ConvLSTM: 75.94%, VGG16+LSTM: 92.50%. We conducted explainable analysis on the VGG16+LSTM model combined with the LRP method. We found VGG16+LSTM classification model tends to use the frames biased towards the latter half of the video and the last frame as the basis for classification.

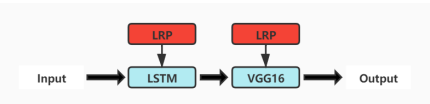
1. Introduction

In this article, two video classification models based on deep networks are first proposed. First, Convolutional neural network is used as the encoder to extract the feature space of the frame in the video information, and then it is combined with LSTM model to process the data of time series and generate the corresponding description as the decoder. In the second model, a CNN network (VGG16) is combined with LSTM and trained using time series images. In addition, in this article, we also emphasized the importance of Explainable analysis and summarized and classified the Explainable research of deep learning models. In the Explainable analysis section of this article, we used the Explainable model for networks Layer-Wise Relevance Propagation (LRP). This model is used to visually interpret video classification results and obtain relevance score results about video frames. Based on the relevance score results obtained. It is possible to explore the important features that the classification model focuses on and use them as the main reference information for the final classification.

2. Model Introduction

2.1 Model Introduction Based on VGG16+LSTM Network

In our model structure, we define the hybrid model structure for video classification as two parts. Firstly, VGG16 is used to extract spatial dimension features, and then the features are input into LSTM to extract temporal dimension features. Therefore, in the process of explaining the network, we also divided the video classification model into two parts. Firstly, in the first part, LRP networks are used to explain LSTM. Then, the interpretation results obtained from the LSTM network are used as the initial weights for the second VGG16 network for interpretation. As shown in the figure below, this is the overall structure of our LRP network model.



2.2 The LRP-based explainable network based on the VGG16

In the LRP-based explainable network based on the VGG16, it can be seen from the original model diagram that the output of the VGG16 network is the input of the LSTM network, so we need to use the explanation result obtained from the LSTM network in the previous step as the initial weight of our VGG16 network.

The propagation rules at different layers are as follows:

Pooling layer : Similar to the previous method, the back-

propagation signal is redirected to the position recorded during the forward propagation.

Activation layer : The back-propagation signal is simply passed to the next layer without rectification. This propagation rule satisfies the conservation law.

Convolutional layer : Bach [7] proposed two correlation propagation rules for this layer, which also satisfy the conservation law. Let $z_{ij} = a_i^{(l)} w_{ij}^{(l,l+1)}$ be the weighted activation of the i neuron to the j neuron in the next layer. The first propagation equation shown at next:

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j} + \epsilon sign(\sum_{i'} z_{i'j})} R_j^{l+1}$$

If a neuron in the previous layer makes a major contribution to a neuron in the latter layer, then the neuron i should account for a larger share of the relevance Rj of the j neuron. And the second propagation equation is to separate positive and negative activations in the relevance propagation equation. That is, the following equation:

$$R_i^{(l)} = \sum_j (\alpha \cdot \frac{z_{ij}^+}{\sum_i {}'z_{i'j}^+} + \beta \cdot \frac{z_{ij}^-}{\sum_i {}'z_{i'j}^-}) R^{(l+1)_j}$$

Among them, zij+ and zij− respectively represent the positive part and the negative part of zij . In addition, the parameter setting α + β = 1 makes the relevance between the layers conserved.

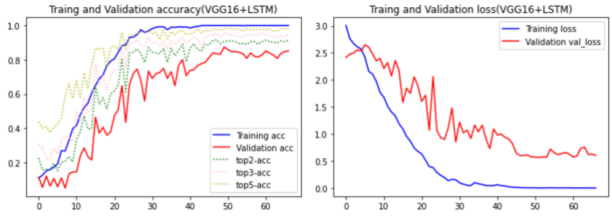
3. Experimental Results

In the experiment of this study, in order to reduce the cost of model training time, we chose a smaller dataset similar to UCF101, UCF11, as a substitute. The experiment in this article is based on the sub dataset of UCF101: UCF11.

In fact, video classification requires not only spatial dimensional information of the image, but also temporal motion information between consecutive frames. Therefore, it is necessary to capture the contextual and spatiotemporal information of video frame sequences, so that classification models can better complete individual action recognition in videos. When designing a video motion recognition model, an important idea is that the network used must be able to capture spatiotemporal information. Therefore, the video classification algorithms studied in this article are all based on spatiotemporal feature extraction, and finally, Softmax classifier is used for recognition.

3.1 Experimental results for VGG16+LSTM

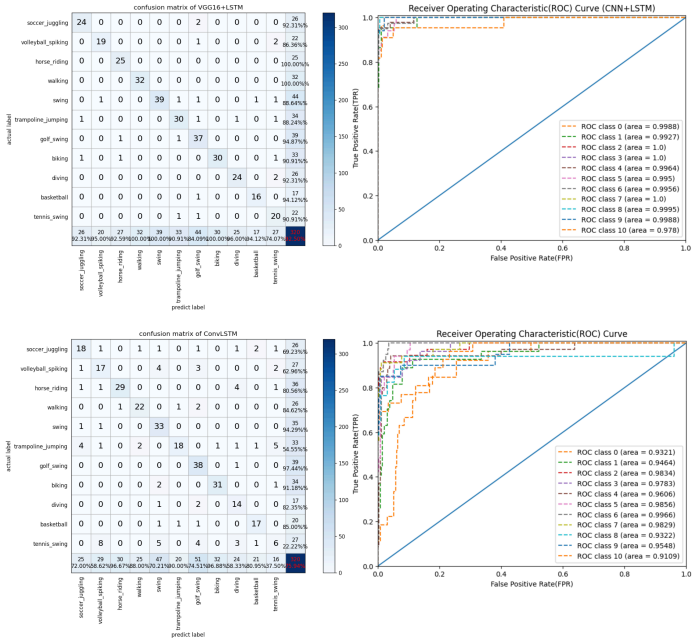
The training validation accuracy and training validation loss of the VGG16+LSTM model are shown in Figure. It achieved approximately 100% training accuracy after 35 epochs and reached a maximum validation rate of over 80% after 60 epochs. It can also be seen that the model does not show overfitting. The model was trained over 100 epochs, with a batch size of 32 and an early stop of 15.



As shown in Figure, the video classification task performance of VGG16+LSTM model on UCF11 dataset is described by Confusion matrix.

The results obtained from the Confusion matrix can be easily seen. The average prediction accuracy of the model is 92.50%. The accuracy, recall, and F1-score reports for each class are shown in Figure. The average accuracy, recall rate, and F1-score are about 93%

we also plotted receiver operating characteristics (ROC) curves that reflect the performance of the VGG16+LSTM model and evaluated them, as shown in Figure. From this figure, we can see that the curves of most classes on the ROC chart converge to the top-left corner. According to the principle of ROC, this indicates that the model has a high prediction accuracy for most classes.



3.2 Comparative analysis of the performance of the two models

In general, according to the training accuracy curve, Confusion matrix, and ROC curve from the two models, it can be concluded that the final accuracy of the VGG16+LSTM video classification model is higher than that of the ConvLSTM model, which shows better performance on the UCF11 dataset. In the next interpretability analysis

stage, we use LRP to better explain the VGG16+LSTM model.

From Table, we can see the performance comparation between two models.

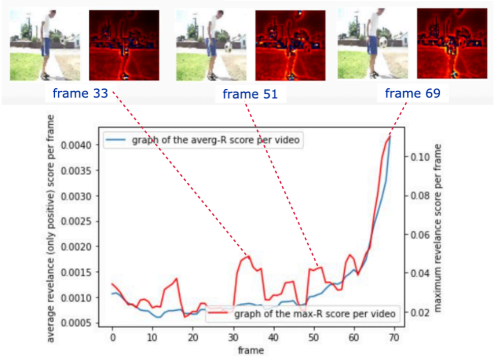
Model Name	Total Parameters	Final Accuracy	F1-Score
VGG16+LSTM	15,802,955	92.50%	93%
ConvLSTM	1,280,239	75.94%	76%

4. Explainable Analysis Based on VGG16+LSTM Network Model

We used the LRP algorithm to explain the prediction of football category videos from the UCF11 test set. In order to obtain a first impression of the explanatory power of the model prediction, a specific video with obvious Object behavior in the football category was first selected and explained separately.

In Figure we show the frames intercepted by the time series labels corresponding to the example video, as well as the LRP interpretation corresponding to the prediction label “soccer juggling”. Football is recognized as relevant, and we can clearly see that the Object individual and football are marked in red, especially in frame 69, which is very obvious. Other parts of the frame image, such as buildings and sky in the background, are not highlighted, so it can be concluded that these two parts are not related to the video.

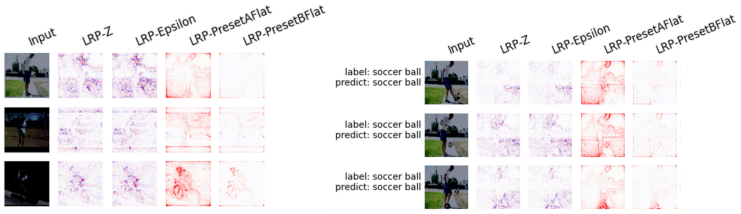
In addition, in order to further understand the distribution of relevance scores on timestamps. We plotted a cross time relevance score curve for soccer videos. From this Figure, it can be seen that the blue curve shows the average relevance score (only consider the positive) over the captured 70 frame time range, while the red curve shows the maximum relevance score over the entire time.



From Figure, we can obtain some important conclusions about video explainable analysis. There are three marked higher points in the figure, and the frames corresponding to these positions are an important part of the conclusions drawn by the classification model. Identify the timestamps corresponding to these three higher points and generate corresponding heatmaps. (The timestamps for this example are 33, 51, 69). It can be observed that the VGG16+LSTM video classification model tends to use the last few frames of the video as the main basis for final classification decisions.

However, there are still some important frame information located in the early and middle stages that also play an important role in video classification task.

Based on the traditional interpretability analysis of LRP, this article aims to compare the LRP performance under various parameter configurations and variations. LRP-Z, LRP-Epsilon, LRP-PresetAFlat, and LRP-PresetBFlat were used for comparison with four different methods. By comparing the four LRP variants using the UCF11 dataset with "soccer juggling" and "biking" data, it was found that for the current analyzed VGG16+LSTM model, the LRP -PresetAFlat method is better and the decision boundaries are clearer.



5. Conclusion

In this paper, we use deep learning network models based on LSTM technology video classification. The two Hybrid model are ConvLSTM model and VGG16+LSTM model. We conducted experiments using the same hyper-parameter settings and trained these two models on the UCF11 video dataset. The final experimental results showed that the ConvLSTM model had a final accuracy of 75.94%, while the VGG16+LSTM model achieved a final accuracy of 92.50%, demonstrating better classification performance.

Afterwards, in the explainable analysis section of the model, the LRP algorithm was combined with the VGG16+LSTM model to obtain a visual classification decision. Obtained heat maps of different frame positions in the selected video. And the relevance score of the corresponding time curve.

Finally, through a large number of experiments comparing the score curves and heat map analysis of different videos, we found that the explainable conclusion of the VGG16+LSTM classification model can be summarized as that the classification model tends to use the frames biased towards the latter half of the video and the last frame as the basis for classification.

References

[1] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems, pages 802–810, 2015.

[2] Gr´egoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Mu¨ller. Layer-Wise Relevance Propagation: An Overview, pages 193–209. Springer International Publishing, Cham, 2019.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In
2005 IEEE Computer Society Conference on Computer Vision and Pattern Recog- nition (CVPR'05), volume 1, pages 886–893 vol. 1, 2005.

[4] Heng Wang, Alexander Kl¨aser, Cordelia Schmid, and Cheng-Lin Liu. Dense tra- jectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 103(1):60–79, May 2013.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[6] Dan C. Cire¸san, Alessandro Giusti, Luca M. Gambardella, and Ju¨rgen Schmidhu- ber. Mitosis detection in breast cancer histology images with deep neural networks. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, pages 411–418, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[7] Sebastian Bach, Alexander Binder, Gr´egoire Montavon, Frederick Klauschen, Klaus-Robert Mu¨ller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE, 10(7):1– 46, 07 2015.