JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

"Video Classification Using Deep Ensemble Machine and Convolutional Neural Network"

Miss. Priti K. Mudkanna, Prof. Smita S. Sangewar

Student, Professor of Computer Science Department.
Department of Computer Science, D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji, India.
Department of Computer Science, D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji, India.

Abstract: Video classification has been extensively researched in computer vision due to it's wide spread application. It always remains an outstanding task because of great challenges in effective spatial-temporal feature extraction efficient classification with high dimensional video representation. in this paper, we propose an end-to-end learning framework called deep ensemble machine (DEM) for video classification. It is important to extract spatial as well as temporal features from video for that two deep convolutional neural networks are proposed i.e. Vision and Graphics Group and C3-D.

Index Terms **- Convolutional Neural networks, Ensemble Machine, Vision and Graphics Group, C3-D, Video Classification.**

1. INTRODUCTION

Video classification has been extensively researched due to its wide spread use in many Important applications such as human action recognition and dynamic scene classification. It is highly desired to have an end-to-end learning framework that can establish effective video representation while simultaneously conducting efficient classification. Here introduce the ensemble technology into deep learning and design a single end-to-end learning architecture without explicitly extracting optical flow, which would be computationally more efficient. The convolutional 3-D and VGG (Vision and Graphics Group) are first deployed to extract temporal and spatial features from the input videos cooperatively, which establishes comprehensive and informative representations of videos.

VGG and C3-D are chosen due their strong capability of extracting complementary spatial and temporal features for comprehensive video representations. The introduced RLE layer is further deployed to encode the initial outputs of classifiers which is followed by a weighting layer jointly learned in the end-to-end framework to combine classification results.

The end-to-end learning framework for video classification is illustrated in fig.1. The convolution 3-D(C-3D) and VGG are first deployed to extract spatial and temporal features from the input videos cooperatively, that is input to the deep ensemble machine. The resultant high-dimensional representations are further reduced by random projections into a set of lower dimensional subspaces, on which an ensemble of efficient classifier is trained with these lower dimensional features. RLE layer is further deploy to encode the initial outputs of classifiers, which is followed by weighting layer jointly learned in the end-to-end framework to combine classification result. This approach combines the strength of deep CNNs for effective feature extraction and ensemble learning for efficient classification.

2. LITERATURE REVIEW

Krizhevsky, I. Sutsakever [10] proposed Imagenet classification with deep convolutional neural networks. Here trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers employed a recently-developed regularization method called "dropout" that proved to be very effective. Also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry. Proposed work's results show that a large, deep convolutional neural network is capable of achieving record breaking results on a highly challenging dataset using purely supervised learning. It is notable that this network's performance degrades if a single convolutional layer is removed. For example, removing any of the middle layers results in a loss of about 2% for the top-1 performance of the network. So the depth really is important for achieving proposed results.

Joe Yue-Hei Ng [9] proposed and evaluate several deep neural network architectures to combine image information across a video over longer time periods than previously attempted. Proposed two methods capable of handling full length videos. The first method explores various convolutional temporal feature pooling architectures, examining the various design choices which need to be made when adapting a CNN for this task. The second proposed method explicitly models the video as an ordered sequence of frames. For this purpose, employ a recurrent neural network that uses Long Short-Term Memory (LSTM) cells which are connected

to the output of the underlying CNN. LSTM is used to process the features extracted from frames by CNNs. Those feature-level fusion methods concatenate the features simply for the following processing, which do not investigate the ensemble essence from the classification point of view. When it comes to video classification tasks, one has to consider not only the appearance information in the spatial domain but also the complex temporal evolution of the scenes. However, due to the fusion of spatial and temporal features, it would unavoidably induce high-dimensional video representations, which raise great challenge for efficient and accurate video classification.

The task of classifying videos of natural dynamic scenes into appropriate classes has gained a lot of attention in recent years. The problem especially becomes challenging when the camera used to capture the video is dynamic, A.Gangopadhyay[7] analyses the performance of statistical aggregation (SA) techniques on various pre-trained convolutional neural network(CNN) models to address this problem. The proposed approach works by extracting CNN activation features for a number of frames in a video and then uses an aggregation scheme in order to obtain a robust feature descriptor for the video. The final descriptor obtained is powerful enough to distinguish among dynamic scenes and is even capable of addressing the scenario where the camera motion is dominant and the scene dynamics are complex. Further, this paper shows an extensive study on the performance of various aggregation methods and their combinations. As compared to the previous spatio-temporal approaches, in this paper focus on capturing temporal variations of very powerful spatial descriptors provided by CNN. This method is computationally efficient than the traditional local feature extraction, encoding and pooling approaches. Here observed that CNN spatial descriptors are excellent representatives of spatial information, as demonstrated by the accuracies obtained using only a single frame per video.

3. PROPOSED METHODOLOGY

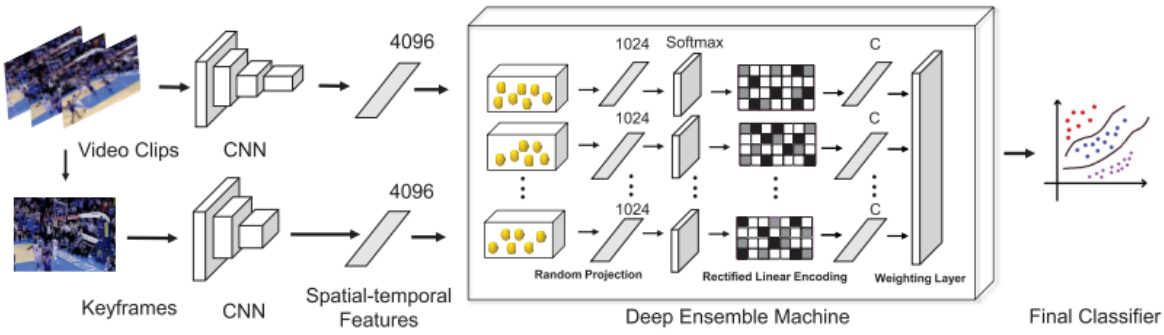

Fig1: Proposed System architecture.

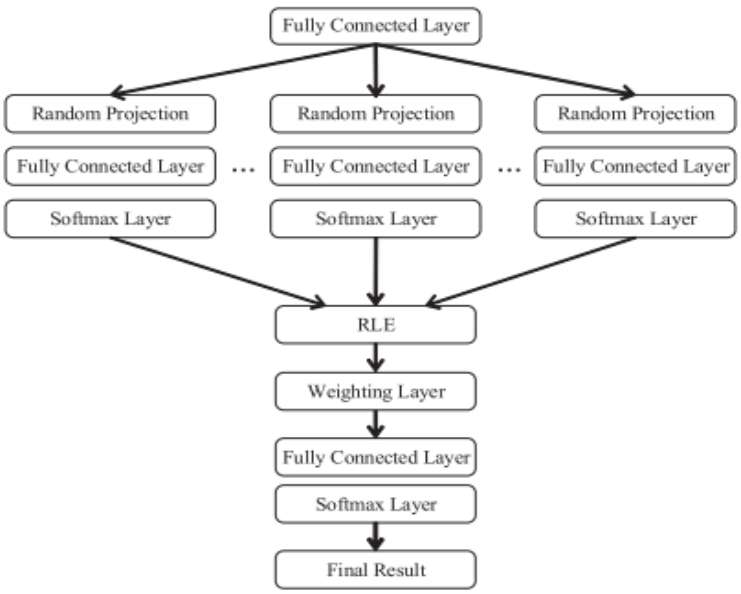

Figure2. Learning Architecture of Deep Ensemble Machine Module.

The convolution 3-D(C-3D) and VGG are first deployed to extract spatial and temporal features from the input videos cooperatively, that is input to the deep ensemble machine. The resultant high-dimensional representations are further reduced by random projections into a set of lower dimensional subspaces, on which an ensemble of efficient classifier is trained with these lower dimensional features. RLE layer is further deploy to encode the initial outputs of classifiers, which is followed by weighting

layer jointly learned in the end-to-end framework to combine classification result. This approach combines the strength of deep CNNs for effective feature extraction and ensemble learning for efficient classification.

4. MODULES

4.1 CNN (Convolutional Neural Network):

The CNN (Convolution Neural Network) algorithm is used in order to detect the particular part of the frame. Then the maximum weight values are taken from the feature extraction frames by using the Convolution neural network. A variant of CNN C3-D neural network used 3-D convolutional and 3_D pooling to extract spatiotemporal features from videos in an end-to-end framework. Here CNN takes input from video frames and after processing on that frame by using C3-D and VGG it provides the output to the Ensemble machine learning framework.

4.2 Random Projection:

Random projection can efficiently and effectively transform a high –dimensional vector to a low dimensional vector while retaining essential information. High dimensional representations of videos are obtained simultaneously. It is computationally inefficient by directly conducting classification on the high-dimensional features. Here to address this issue random projection layer is used to reduce the representation into set of low-dimensional subspaces, in which an ensemble of classifiers is built to improve the effectiveness and robustness of Deep Ensemble Machine. By doing so, the redundancy in features is reduced, and at the same time more representative features are selected.

More specifically, random projection is conducted by a sparse random matrix, which could be defined as follows:

$$r_{ij} = \sqrt{s} * \begin{cases} 1, & p = \frac{1}{2s} \\ 0, & p = 1 - \frac{1}{s} \\ -1, & p = \frac{1}{2s} \end{cases}$$

Where rij is the element of the random sparse matrix, p is the probability of rij, and s is set to a suitable value to adjust the sparseness, which not only satisfies the Johnson–Lindenstrauss lemma rule but also makes the matrix relatively sparse.

4.3 Rectified Linear Encoding:

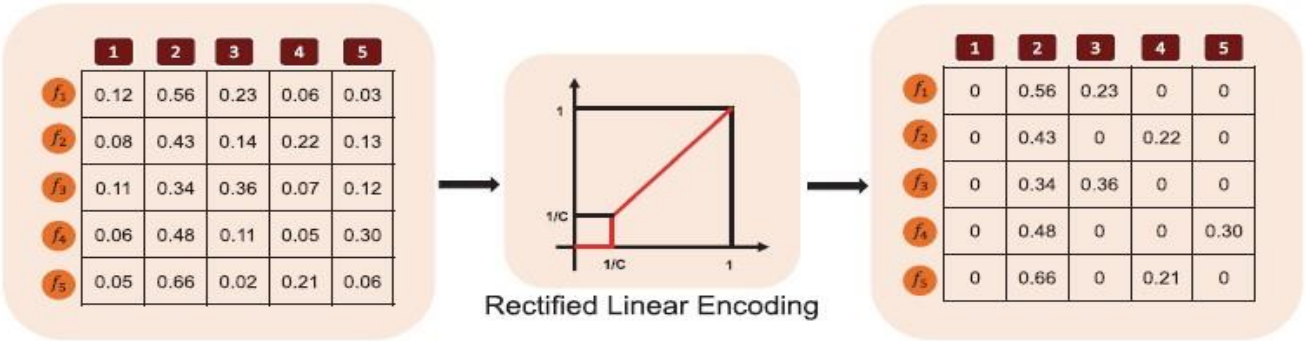

Fig. 3. Principle of RLE Encoding. Left disc: activations of softmax layers, in which softmax output of a base classifier is indicated by each row, and category is represented by each column. Middle disc: our proposed RLE, by which the encoded features can be obtained in the right disc.

The classification result from the ensemble of classifiers could be simply concatenated into long feature vector and fed into fully connected layer to produce the final result. These base classifiers with C-dimensional probability vectors are the initial outputs that produce the coarse classification, where C is the total number of categories. The C-dimensional vectors are further processed by the RLE layer, which aims to reduce the computational cost. In the RLE layer, the probabilities to different categories are treated as features for the following layer. If the probability, i.e., the value of the feature, is larger than 1/C, the sample has a high probability belonging to the corresponding categories. This RLE layer brings two desirable merits. On the one hand, the RLE layer encodes the outputs of base classifiers into a single feature vector, which provides an effective way to fuse the initial classification results from base classifiers. On the other hand, although it plays a similar role to the ReLU in CNNS activation and suppresses most inactive neurons. This character largely reduces the number of parameters in the following fully connected layer, which enables efficient training for classification.

4.4 Ensemble of classifier:

Instead of directly concatenating the encoded results from classifiers, here propose learning the weights associated with base classifiers for ensemble. This idea is well-founded by the ensemble learning theory: when none of the models in an ensemble results into a correct classification, there usually exists a combination that can obtain better performance than any individual model or classifier on its own. Since low-dimensional features in different subspaces are obtained by distinct random projections, they are different from each other and play distinctive roles in the final prediction. A separate weighting layer is useful to assign weights for base classifiers in a backpropagation process during the training stage, by which base classifiers are combined optimally for better performance.

5. CONCLUSION

In this paper, we have presented an end-to-end deep learning framework, the DEM, for video classification. Based on the spatial-temporal features extracted by heterogeneous deep CNNs, VGG, and C3-D, we propose an ensemble learning module based on random projections to work on the top of CNNs. Random projections reduce high-dimensional video representations into a set of low-dimensional subspaces based on which an ensemble of classifier is learned for prediction. The initial results from the classifiers are further encoded by a newly proposed RLE layer, which is followed by a fully connected layer to combine all the classifiers to produce the final classification results. Our approach leverages the strengths of deep CNNs for feature extraction and ensemble learning for efficient classification. We have conducted extensive experiments on four challenging data sets for video classification including human action recognition and dynamic scene classification. The proposed approach achieves high performance on all the data sets and surpasses state-of-the art methods, showing great effectiveness for diverse video classification tasks.

6. REFERENCES

[1] Z. Lin, Z. Jiang, and L. S. Davis, “Recognizing actions by shapemotion prototype trees,” in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep./Oct. 2009, pp. 444–451.

[2] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in Proc. Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 1933–1941

[3] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 568–576.

[4] C. Thériault, N. Thome, and M. Cord, “Dynamic scene classification: Learning motion descriptors with slow features analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 2603–2610.

[5] A. Gangopadhyay, S. M. Tripathi, I. Jindal, and S. Raman, “Dynamic scene classification using convolutional neural networks,” in Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP), Dec. 2016, pp. 1255–1259.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105

[7] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 4694–4702.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 4489–4497

[9] Jiewan Zheng, X. Cao, B. Zhang, “Deep Ensemble Machine for Video Classification”, in Proc. IEEEE transaction on neural networks and learning system, 2018.