Dense Convolutional Networks for Efficient Video Analysis

Tian Jin
Master Student
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+86-15257220611
Tian.jin@strath.ac.uk

Zhihao He
Master Student
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+44-7712734792
zhihao.he@strath.ac.uk

Amlan Basu
Ph.D. Candidate
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+44-7459802138
amlan.basu@strath.ac.uk

John Soraghan
Professor
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+44 (0)141 548 2514
j.soraghan@strath.ac.uk

Gaetano Di Caterina
Research Fellow
CeSIP, Royal College Building
University of Strathclyde, Glasgow
01415484458
gaetano.di-caterina@strath.ac.uk

Lykourgos Petropoulakis
Senior Knowledge Exchange Fellow
CeSIP, Royal College Building
University of Strathclyde, Glasgow
+44-7459802138
l.petropoulakis@strath.ac.uk

Abstract - Over the past few years various Convolutional Neural Networks (CNNs) based models exhibited certain human-like performance in a range of image processing problems. Video understanding, action classification, gesture recognition has become a new stage for CNNs. The typical approach for video analysis is based on 2DCNN to extract feature map from a single frame and through 3DCNN or LSTM to merging spatiotemporal information, some approaches will add optical flow on the other branch and then post-hoc fusion. Normally the performance is proportional to the model complexity, as the accuracy keeps improving, the problem is also evolved from accuracy to model size, computing speed, model availability. In this paper, we present a lightweight network architecture framework to learn spatiotemporal feature from video. Our architecture tries to merge long-term content in any network feature map. Keeping the model as small and as fast as possible while maintaining accuracy. The accuracy achieved is 91.4% along with an appreciable speed of 69.3 fps.

Keywords- CNN, LSTM, 3D-Net, DenseNet, 2D Layer, Video Analysis.

I. INTRODUCTION

Thanks to larger datasets and multiple deep learning algorithms, video understanding and, specifically, action classification which have reported outstanding performance in recent years. Meanwhile, efficient and low cost video analysis algorithm has become a meaningful subject, especially for real time human-robot computer interaction. Short time action recognition is depending on the present action within a short time window, however video understanding is concerned with longer-term frames information. Several architectures have been proposed to capture spatiotemporal information between frames, however the speed and efficiency always remained at the bottleneck. Such as 3D CNN [5,6], theoretically time window should cover whole video, due to the cost of calculation people have to use a small window, each convolution computation contains several frames, the information is merged by convolution and pooling computation layer by layer. Obviously, this algorithm is suboptimal for exploiting relationship between frames. In fact, ordinary LSTM [3, 4] structure still cannot get satisfactory results, even if it has outstanding performance in several sequence problems, some algorithms choose to sacrifice speed to improve accuracy, these ideas are through add an optical flow branch to achieve 'Two-Stream' structure [7, 8] and the get excellent results, however the speed of 'Two-Stream' is too slow that these algorithms lack of practicality.

Our approach is inspired by DenseNet [1] and ECO [2], we proposed a simple end-to-end trainable architecture to address previously mentioned dilemma, DenseNet provides a good idea to reduce calculation, normally feature map will increase as the number of layers increase, but in this way only a few number of feature maps are calculated per layer and through concat operation to merge them, on the other hand unlike other combine operations, concat can pass the results of each computation to the output without change. Effectively processing the information in each frame is the contribution of ECO. Temporal neighbourhood of a frame almost redundant information. So normal 3D CNN method is very inefficient to slide from the first frame to the last. Most calculation will produce similar results and don't carry any useful information. It will waste a lot of computing resources. Select a single frame of a temporal neighbourhood and use bidirectional ConvLSTM to yield a representation for each sampled frame, each representation will have a strong ability to extract spatiotemporal information. Not like ECO, ECO yield the feature map with a 2D convolutional architecture and then to capture the contextual relationships between distant frames, re-stacked feature maps and feed into 3D CNN. The overview of our approach architecture is illustrated in Figure 1. Consequent network will become faster and still have satisfactory accuracy. This makes more sense for real time video retrieval.

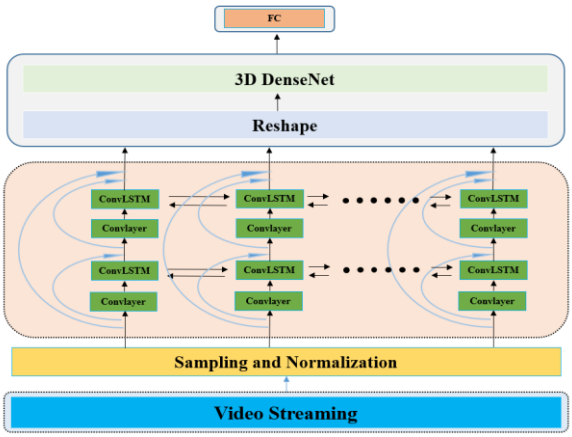

Figure 1. Overview of Dense Convolutional Networks, Dense bidirectional ConvLSTM are utilized to learn long-term spatiotemporal features and sent original information to the high-level structure, all feature will be re-stacked and fed to 3D DenseNet which classifies the action.

II. Type Style and Fonts

Although the development of deep learning is very rapid, but there is still not a unified video analysis framework, the current architecture is mainly based on convolutional layers and layers operators with 2D or 3D convolution kernel. Also, either the network is only for analysing RGB videos or includes pre-computed light flow. 2D convolution merges the space and time information which is evident from LSTM network structure in 2D convolution, spatiotemporal information (can be learned by LSTMs), or aggregation over time.

For video classification problems, there are some generally accepted algorithm based on deep learning method [8, 12-15]. In order to extract more spatiotemporal features from a sequence of frames 3D convolution network is introduced. 3D kernel can learn the spatiotemporal feature. Tran et al. [3] proposed a 3D architecture using 3D convolution kernels which slide whole video. They studied the 3D structure of Resnet system and improved the structure of C3D [15]. In addition, recurrent networks are also a good approach to extract temporal relation between frames [16, 17, 12]. Donahue et al. [12] proposed an approach using LSTM to integrate features from CNN. In fact, LSTM-based network to extract the features of spatiotemporal is inefficient and unsatisfactory. In action recognition field, the performance of LSTM-based method is always behind the CNN-based methods. Recently introduced 3D CNN has some new approach and new methods emerge one after another [10, 14, 18]. The characteristic of these methods is the use of sliding window to obtain short term temporal context. However, such methods consume all computing resources because the average score of these Windows needs to be calculated before final fusion.

In fact, these methods do not utilize video's spatiotemporal information very efficiently and require very large computing resources. Especially through the post-hoc fusion, not only inefficient, but also reduces the operation speed which cannot be used in the scene of fact work.

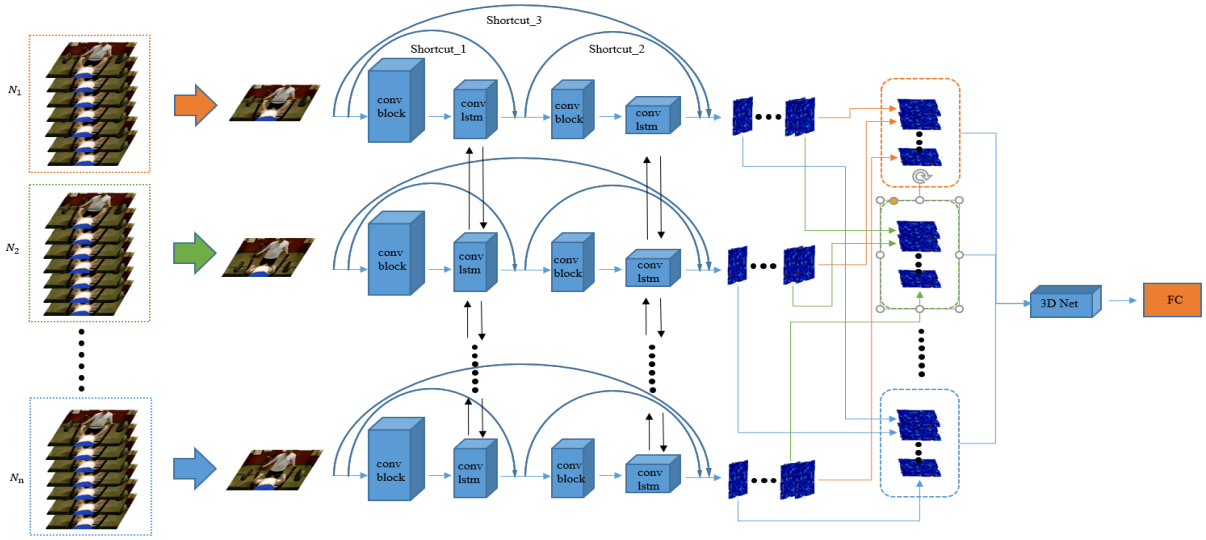

Figure 2 A video will be divided into N subsection, and sampled one frame form each subsection randomly, 2 DenseConvLSTM part will be used to learn sample spatiotemporal feature, the feature from different frame have same place. will be re-stacked and 3D DensNet used to post-hoc fusion.

III. Architecture

The network architecture is shown in Fig. 2. a whole video will be split into N subsections N_n n = 1, 2,……, N each subsection containing the same number of video frames and then randomly samples a frame from each subset to represent the entire subset. Since there are a lot of duplicate frames in whole video, it is very wasteful to send the whole video into the network for calculation, and it will get a lot of repeated useless information. By this method, redundant information can be avoided and useful

information can be extracted, and spatiotemporal information of video can be extracted effectively through repeated random sampling iteration in training.

After sampling, N frames will be obtained and will be sent to network. The network has two main components, the first is the DenseConvLSTM, and this part is used for learning simple spatiotemporal information. And second is the 3D DenseNet, this part is for high dimensional spatiotemporal information fusion. Concat is also used in the DenseConvLSTM block, which means that the original information can be passed on to the next structure irrespective of whether the results of the feature are good or not.

Since convolution layer is very useful in learning the features of image, it is necessary to extract the features of video using the convolution block before bi-direction ConvLSTM. Each convolution block has three different convolution layers, and connect to ConvLSTM block to learn spatiotemporal features. ConvLSTM consists of two bi-direction LSTM. All features will be re-stacked using DenseNet architecture with 3D convolutions. The DenseNet contains 4 dense blocks with growth rate of 32 and 3D DenseNet network yields the final action class label. It is a straightforward architecture and it can be trained end-to-end on action classification and large data base.

A. 2D Layer

Each Conv block contains three convolution layers and two bi-direction LSTM layers. The parameters of convolution layers and shortcuts are shown in table 1.

TABLE I. THE PROPOSED ARCHITECTURE OF CNN

Layer Name	Output Size	Parameter
Conv1_1	109×109	7×7 Conv, Stride 2, Padding 0,16
Conv1_2	54×54	3×3 Conv, Stride 2, Padding 0,16
Conv1_2	54×54	3×3 Conv, Stride 1, Padding 1,16
Conv2_1	26×26	3×3 Conv, Stride 1, Padding 1,16
Conv2_2	26×26	3×3 Conv, Stride 1, Padding 1,16
Conv2_2	26×26	3×3 Conv, Stride 1, Padding 1,16
Shortcut_1	54×54	11×11 Conv, Stride 4, Padding 0,32
Shortcut_2	26×26	3×3 Conv, Stride 2, Padding 0,32
Shortcut_3	26×26	11×11 Conv, Stride 4, Padding 0,32 3×3 Conv, Stride 2, Padding 0,32

B. ConvLSTM

In the traditional fully connected LSTM, input features will be vectorized, which is very unfavorable for learning space-time features. In the transmission process of each layer, vectorized features will lose spatial relevance, however, object posture and position is significance for recognition. So ConvLSTM [19] is used to learn long-term temporal and spatial properties, and convolution and recursion can take full advantage of spatiotemporal information. This is also shown using the mathematical equations which are,

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ct} \circ C_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma\big(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\big) \quad (2)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \quad (4)$$

$$H_t = o_t \circ \tanh(C_t) \quad (5)$$

The feature map of hidden layer we adopt 32 as each signal LSTM output

C. 3D-Net

For the 3D network, we adopted the 3D structure based on DenseNet. DenseNet itself has many advantages, reducing model size and increasing classification effect. This is a very effective structure and the network parameters are shown in table 2.

TABLE II. NETWORK PARAMETERS FOR 3D-NET

Layer Name	3D-DenseNet (growth rate=32)
3D-Dense Block (1)	$\begin{bmatrix} 1\times1\times1\ conv \\ 3\times3\times3\ conv \end{bmatrix}$ × 4, stride 1, padding 1
Transition Layer (1)	1 × 1 × 1 stride 1, padding 0
	3 × 3 × 2 average pool, stride 2 × 2 × 2, padding 1
3D- Dense Block (2)	$\begin{bmatrix} 1\times1\times1\ conv \\ 3\times3\times3\ conv \end{bmatrix}$ × 8, stride 1, padding 1
Transition Layer (2)	1 × 1 × 1 stride 1, padding 0
	3 × 3 × 2 average pool, stride 2 × 2 × 2, padding 1
3D- Dense Block (3)	$\begin{bmatrix} 1\times1\times1\ conv \\ 3\times3\times3\ conv \end{bmatrix}$ × 8, stride 1, padding 1
Transition Layer (3)	1 × 1 × 1 stride 1, padding 0
	3 × 3 × 2 average pool, stride 2 × 2 × 2, padding 1
3D- Dense Block (4)	$\begin{bmatrix} 1\times1\times1\ conv \\ 3\times3\times3\ conv \end{bmatrix}$ × 8, stride 1, padding 1
Classification Layer	7 × 7 global average pool
	FC, softmax

IV. TRAINING DETAILS

We used mini-batch of SGD to train our network and used dropout in each fully connected layer. Each video is divided into 10 segments and randomly selected one frame from each segment. The purpose of this sampling is to avoid redundant information and improve network robustness. Furthermore, we applied the data augmentation technique introduced in resizing the frame of dataset to 240 × 320 and it adopts the scale jittering with horizontal flipping. Then we use per-pixel mean subtraction and randomly crop a 224×224 figure from frames. The initial learning rate is 0.001. When loss is stable, it is decreased by a factor of 10. The momentum of our training network is 0.9, the weight

attenuation is 0.0005, and the minus-sized is 32. All the training was done on a tesla P100 GPU.

Test time inference: In order to improve the accuracy, many start-of-the-art algorithms prefer to run some post-processing. For instance, TSN [9] and ARTNet [10], each segment video collects 25 individual frames. Each frame/volume sample crops 10 areas of corner and center and their horizontal flipping. Then based on the average scores of the 250 samples results are obtained. It is clear that methods for calculation are very expensive and that is not the best way for fast video analysis algorithm. However, our algorithm without too much additional operations, is based on many widely recognized ideas to solve the problem of speed. During the test period, our algorithm in video is divided into N subsection and randomly extracts a frame from each section, these frames only perform center cropping and send the cropped part to the network. This is an end-to-end prediction.

V. EXPERIMENT

In order to test the generalization ability of our method, we evaluated it on different data sets to make it more convenient and intuitive to compare with other excellent methods. We tested UCF101 [11] on the original RGB video as input.

We also tested speed of our approach at the same time. A lot of methods always select fps (frames per second) as a standard comparison for speed. Actually such comparisons have problems because of the different algorithms applied on same video to calculate the number of frames which is always different. So here we are more concerned with the processing speed of each video. Our approach can reach 790 FPS on a Tesla P100 GPU. All the results are shown in table 3.

TABLE III. THE SIMULATION RESULTS

Method	Pre-Train	Accuracy (%)	Speed (fps)
Our Approach	**Kinetics**	**91.4**	**69.3**
HOG [20]	-	72.4	-
ConvNet+LSTM (AlexNet) [21]	ImageNet	68.2	-
C3D (VGGNet-11) [3]	Sports-1M	82.3	-
Res3D [15]	Sports-1M	85.8	<2
TSN Spatial Network [9]	ImageNet	86.4	21
TSN Spatial Network [9]	ImageNet + Kinetics	91.1	21
RGB-13D [22]	ImageNet+ Kinetics	95.6	0.9
ECO_{12F} [2]	Kinetics	92.4	52.6
ARTNet w/o TSN [10]	Kinetics	93.5	2.9

CONCLUSION

We have put forward a structure which is highly efficient for video analysis. And that is evident form the result obtained. The proposed structure can be used not only in video classification but also can be used in various video understanding tasks. The advantage is that network only needs an RGB image and no other pre-treatment. ConvLSTM replaces the traditional LSTM which is good for extracting spatiotemporal information. High density of residual can make the network robust as well.

ACKNOWLEDGMENT

We are grateful to University of Strathclyde for providing us with appropriate software and computer system because of which we are able to present this work. Also, thanks to the deep learning group based in CeSIP for their valuable support and guidance.

REFERENCES

[1] Huang, Gao, et al. "Densely Connected Convolutional Networks." CVPR. Vol. 1. No. 2. 2017.
[2] Zolfaghari, M., Singh, K., & Brox, T. (2018). ECO: Efficient Convolutional Network for Online Video Understanding. arXiv preprint arXiv:1804.09066.
[3] Tran, Du, et al. "Learning Spatiotemporal Features with 3D Convolutional Networks." Computer Vision (ICCV), 2015 IEEE International Conference on. IEEE, 2015.
[4] Ng, Joe Yue-Hei, et al. "Beyond short snippets: Deep networks for video classification." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.
[5] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." IEEE transactions on pattern analysis and machine intelligence 35.1 (2013): 221-231.
[6] Taylor, Graham W., et al. "Convolutional learning of spatio-temporal features." Proceedings of the 11th European conference on Computer vision: Part VI. Springer-Verlag, 2010.
[7] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional Two-Stream Network Fusion for Video Action Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.
[8] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1. MIT Press, 2014.
[9] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, October). Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision* (pp. 20-36). Springer, Cham.
[10] Wang, L., Li, W., Li, W., & Van Gool, L. (2017). Appearance-and-relation networks for video classification. *arXiv preprint arXiv:1711.09125*.
[11] Soomro,K,Zamir,A.R.,Shah,M.:UCF101:A dataset of 101 human actions classes from video in the wild. CoRR abs/1212.0402(2012)
[12] Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR. (2015)
[13] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Largescale video classification with

convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. CVPR'14, Washington, DC, USA, IEEE Computer Society (2014) 1725–1732

[14] Zolfaghari, M., Oliveira, G.L., Sedaghat, N., Brox, T.: Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In: IEEE International Conference on Computer Vision (ICCV).

[15] Tran, D., Ray, J., Shou, Z., Chang, S., Paluri, M.: Convnet architecture search for spatiotemporal feature learning. CoRR abs/1708.05038 (2017)

[16] Lev, G., Sadeh, G., Klein, B.,Wolf, L.: Rnn fisher vectors for action recognition and image annotation. In Leibe, B., Matas, J., Sebe, N., Welling, M., eds.: Computer Vision – ECCV 2016, Cham, Springer International Publishing (2016) 833–850

[17] Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.: Videolstm convolves, attends and flows for action recognition. Comput. Vis. Image Underst. 166(C) (January 2018) 41–50

[18] Diba, A., Fayyaz, M., Sharma, V., Karami, A.H., Arzani, M.M., Yousefzadeh, R., Gool, L.V.: Temporal 3d convnets: New architecture and transfer learning for video classification. CoRR abs/1711.08200 (2017)

[19] Xingjian, S. H. I., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (pp. 802-810).

[20] H. Wang and C. Schmid. Action recognition with improved trajectories. In ICCV, pages 3551–3558, 2013.

[21] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. Long-term recurrent convolutional networks for visual recognition and description. In CVPR, pages 2625–2634, 2015.

[22] J. Carreira and A. Zisserman. Quo vadis, action recognition?A new model and the kinetics dataset. In CVPR, pages 6299–6308, 2017.

Authors' background

First Author's Name: Tian Jin	
Position (Prof., Assoc. Prof., Asst. prof., Dr., Mr., Ms.)	**Mr.**
Research Field	Vision, Recognition and Reconstruction
Personal Webpage	
Second Author's Name: Zhihao He	
Position (Prof., Assoc. Prof., Asst. prof., Dr., Mr., Ms.)	Mr.
Research Field	Vision, Recognition and Reconstruction
Personal Webpage	
Third Author's Name: Amlan Basu	
Position (Prof., Assoc. Prof., Asst. prof., Dr., Mr., Ms.)	Mr.
Research Field	Vision, Recognition and Reconstruction
Personal Webpage	https://pureportal.strath.ac.uk/en/persons/amlan-basu
Fourth Author's Name: John Soraghan	
Position (Prof., Assoc. Prof., Asst. prof., Dr., Mr., Ms.)	Prof.
Research Field	Image Processing, Vision, Recognition and Reconstruction
Personal Webpage	https://pureportal.strath.ac.uk/en/persons/john-soraghan
Fifth Author's Name: Gaetano Di Caterina	
Position (Prof., Assoc. Prof., Asst. prof., Dr., Mr., Ms.)	Dr.
Research Field	Image Processing, Vision, Recognition and Reconstruction
Personal Webpage	https://pureportal.strath.ac.uk/en/persons/gaetano-di-caterina
Sixth Author's Name: Lykourgos Petropoulakis	
Position (Prof., Assoc. Prof., Asst. prof., Dr., Mr., Ms.)	Dr.

Research Field	Image Processing, Vision, Recognition and Reconstruction
Personal Webpage	https://pureportal.strath.ac.uk/en/persons/lykourgos-petropoulakis