Journal of The Korea Society of Computer and Information
Vol. 24 No. 6, pp. 21-28, June 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.06.021

ADD-Net: Attention Based 3D Dense Network for Action Recognition

Qiaoyue Man*, Young Im Cho*

Abstract

Recent years with the development of artificial intelligence and the success of the deep model, they have been deployed in all fields of computer vision. Action recognition, as an important branch of human perception and computer vision system research, has attracted more and more attention. Action recognition is a challenging task due to the special complexity of human movement, the same movement may exist between multiple individuals. The human action exists as a continuous image frame in the video, so action recognition requires more computational power than processing static images. And the simple use of the CNN network cannot achieve the desired results. Recently, the attention model has achieved good results in computer vision and natural language processing. In particular, for video action classification, after adding the attention model, it is more effective to focus on motion features and improve performance. It intuitively explains which part the model attends to when making a particular decision, which is very helpful in real applications. In this paper, we proposed a 3D dense convolutional network based on attention mechanism (ADD-Net) , recognition of human motion behavior in the video.

▸Keyword: Deep Learning, Action Recognition, Convolution Neural Network, Attention Mechanism

I. Introduction

with the development of deep learning models and a large number of applications, excellent performance has been achieved in object detection, image segmentation, image classification, and face recognition. For the processing of images in the video, from the previous manual mark processing method, starting to apply the deep learning method, processing. In the real world, there are many images based on the video. How to deal with images in the video is an important research direction in computer vision. Recently, the Convolutional Neural Network (CNN) [6] [23] has achieved success in still image processing. The CNN network began to be applied by researchers to process dynamic image [4] and video recognition systems.

In the field of computer vision, human action recognition in videos has become one of the most researched areas. And a lot of applications in real life, such as surveillance, video search, healthcare, human-computer interaction, etc. However, recognizing human actions from a video stream is a challenging task. First of all, video-based human behavior recognition has a somewhat complicated motion background. The high complexity and variability of human motion make it difficult to recognize movement correctly. Second, unlike image data, video data also contains temporal information, which is very important in video classification. In a dynamic and chaotic environment, human positioning

• First Author: Qiaoyue Man, Corresponding Author: Young Im Cho
*Qiaoyue Man (manqiaoyue@gmail.com), Gachon University, Korea
*Young Im Cho (yicho@gachon.ac.kr), Gachon University, Korea
• Received: 2019. 04. 30, Revised: 2019. 05. 30, Accepted: 2019. 05. 30.
• This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2019-2017-0-01630) supervised by the IITP and a project(number 10077965)by Ministry of Trade, Industry and Energy.

becomes more difficult, and the behavior of specific actions at a certain time interval is usually not trivial. In order to recognize human actions and changes in video, it is necessary to segment the video accurately, locate the time region where the actions occur, and recognize the actions. This is challenging work. In this paper, our job with how to accurately detect human and main moving body parts under challenging conditions and extract motion characteristics to achieve the purpose of action recognition. In the paper, we introduced the attention mechanism. Recently, the attention mechanism has played an important role in deep learning, especially in the Recurrent Neural Network (RNN). Attention mechanism originates from the study of human vision. In cognitive science, due to the bottleneck of information processing, human beings selectively pay attention to a part of all information while ignoring other visible information. In neural networks, neural attention mechanism enables the neural network to focus on its input (or feature) subset: select specific input. This is effective for video data processing with complex information. a proper attention model can better focus on where the problem is, to draw a classification decision. It intuitively explains which part the model attends to when Making a particular decision, which is very helpful in real applications, e.g, medical AI systems or Self-driving cars.

In this paper, we propose a novel 3D dense convolutional video recognition model based on the attention mechanism [10]. In order to better process and classify the human motion features in the video, we extract the motion features in the 3D dense convolution network and combine the attention mechanism to further focus the features to obtain the better-classified video. Our solution is based on the original DenseNet [21] framework and is extended under the original framework, by default has 2D filters and pooling kernels - to merge 3D Filters and pooling kernels, namely DenseNet3D. We used DenseNet because it is high parameter efficiency. We replaced the standard transition layer in the standard DenseNet architecture with an attention model to improve feature extraction. In the attention model, we use a dual attention efficient model structure combining space and channel to improve recognition performance. In addition, in the UCF101 and HMDB51 datasets, we validate our proposed model, and the experimental results demonstrate the power of our method, showing superiority or accuracy comparable to the most advanced methods with the same input.

II. Preliminaries

In recent years, deep learning models are widely used in real life, especially in a large number of applications of image recognition, showing its powerful performance, at the same time, video action recognition has also been rapidly developed. Such as surveillance, robotics, healthcare, video searching, virtual reality, and human-computer interaction. Different static single image understanding, video understanding requires a large number of images to determine dynamically changing content in the video.

According to the research, we briefly summarize human behavior recognition into two broad categories: Hand-crafted and deep learning-based methods [20]. When deep learning has not been widely used, hand-crafted representation learning methods dominate, usually detecting time and space points of interest, which are then represented by local information. Such as Space-Time Interest Points (STIP), Histogram of Gradient [22] and Histogram of Optical Flow, 3D Histogram of Gradient [29], and the Dense Trajectory (iDT) [8] of the state-of-the-art handcrafted method. Which explicitly considers the motion features by pooling rich descriptors along dense trajectories [17] and compensates camera motions. Then by the encoding method, the descriptors are aggregated into the video-level representation.

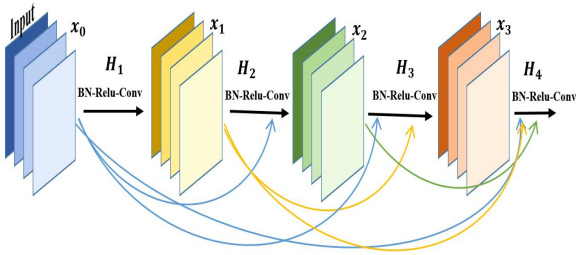

Fig. 1. The connection structure of composite function in a dense block

In recent years, with the rise of in depth learning, especially the emergence of CNN network, as well as powerful and excellent processing capabilities, more and more researchers began to use CNN network to conduct research on action recognition in the videos. For instance, 3D convolutional networks (C3D) [24] [25] use 3D convolution kernels to extract features from a sequence of dense RGB frames; Temporal Segment Networks (TSN) [15] sample frames and optical flow on different time segments to extract information for activity recognition;

I3D networks [28] use two stream CNNs with inflated 3D convolutions on both dense RGB and optical flow sequences to achieve state of the art performance on the Kinetics dataset. And the recent popular use of CNN networks considers the spatiotemporal effects [18] [19], combined with optical flow estimation [12] [14] [16], two-stream fusion [7] [11] [13] to classify video behavior recognition [1] [5]. And using transfer learning to add pre-trained network models, transferring knowledge within or across modalities is effective, and leads to significant improvements in performance.

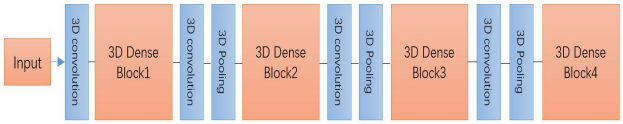

Fig. 2. The overall architecture of DenseNet with four blocks

Video motion recognition is a challenging research field. Due to a large amount of video content and the existence of much unnecessary information, the recognition effect is greatly hindered. In recent years, the attention mechanism model has been widely used by researchers in various aspects of deep learning, whether it is image processing [31] [32], speech recognition [33][34] or natural language processing [35][36] in various types of tasks. In large-scale image classification tasks, Wang et al. [38] propose an attention model based on encoder-decoder. By refining the feature map, the network becomes more robust and also suppresses noise input. Huet al. [37] Introducing the Squeeze-and-Excitation module based on the channel attention mechanism. In the module, they use global average-pooling to calculate channel attention. As we all know, attention is an important part of human visual perception. The visual attention mechanism is a brain signal processing mechanism unique to human vision, by quickly scanning the global image, human vision obtains the target area that needs to be focused on, which is the focus of attention and then invests more attention resources in this area to obtain more detailed information about the target. And suppress other useless information. For example, when people read, usually only a small number of words to be read will be noticed and processed. The attention mechanism has two main aspects: deciding which part of the input to focus on; and allocating limited information processing resources to the important part.

In the deep learning network, with the deepening of the network depth, the preamble and gradient signals of the network during training may gradually disappear after passing through many layers. In this paper, we use dense convolutional networks as the underlying network. DenseNet (Fig. 2) directly connects all the layers in the network using the same size feature maps to ensure the maximum information flow between the layers. Each layer requires information from all previous layers as input, and then it passes its feature maps. The structure can be found in Fig. 1. In traditional convolutional neural networks, if there are L layers, there will be L connections, but in DenseNet, there will be L (L+ 1)/2 connections. Each layer in the network can directly access the gradient through the original input signal and the loss function, leading to implicit deep supervision. The shorter the connection between the layer near the input and the layer near the output, the convolutional neural network can be deeper, more accurate and more effective. Due to the dense block layer, DenseNet has a narrower network and fewer parameters than ResNet [30]. At the same time, this connection makes the transmission of features and gradients more efficient, and the network is easier to train. And, we add an attention model to replace the standard transition layer in the original DenseNet framework, We efficient attention model combines channel and spatial attention, channel attention focuses on ‘what’ is meaning given an input image, the spatial attention is a supplement to channel attention, focus on the 'where' in the information section. After testing and verifying, we found that using channel space attention is better than using channel only, focus on the feature area after convolution, and reduce the network training parameters and have better results.

III. The Proposed Scheme

1. Architecture Overview

We used the 3d dense convolution network as our base network. And in the original dense convolutional network, replace the transition layer with our attention model. Adopting the model framework combined with the channel and spatial attention, the feature refinement extraction, and achieving considerable performance improvement, while maintaining a small system overhead. We use the

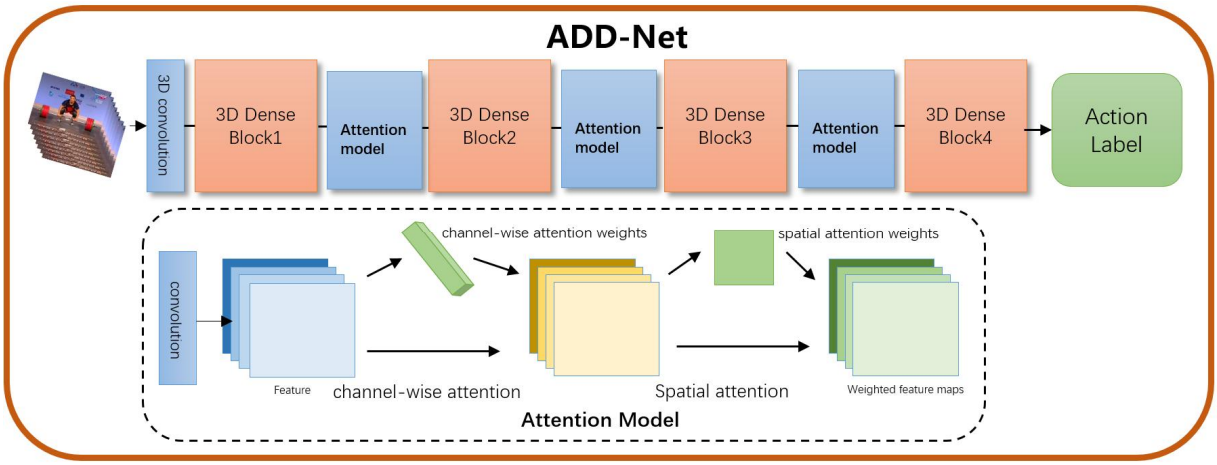

Fig. 3. The overall architecture of the based on the attention model 3D dense convolutional network (ADD-Net). We are based on a 3D dense convolutional network and modify the original network to add our attention model. In the attention model, we use an efficient attention mechanism based on a combination of channel and space attention

DenseNet architecture for several reasons, such as a deeper architecture with simpler and higher parameters, intensive knowledge propagation, and optimal performance for image classification tasks. We modified 2D DenseNet to replace the 2D core with a 3D core in the standard DenseNet architecture, which we present as DenseNet3D. Our attention-based 3D DenseNet model (ADD-Net) Fig. 3 is formed by deploying a dual channel attention module in DenseNet (channel-spatial attention module) instead of a transition layer in a dense network.

2. 3D Dense Networks

A large body of literature suggests that to make convolutional network performance better, either make the network wider or make the network deeper. In our model, we use the DenseNet as a major network branch and add the corresponding 3D module to form the 3D DenseNet, Based on the original 3D dense network, in the transition layer between the dense blocks, we added an attention module to improve the feature recognition effect. And take the multiple densely connected dense blocks. More dense blocks will cause the network to become deeper and better but this will increase the parameters and complexity of the network. On the contrary, the use of a few dense blocks will affect the accuracy of the model due to the small number of network layers. Inthis paper, we use four dense blocks. Each dense block contains several composite functions that are connected in a feed-forward manner. The structure can be found in Figure 3. An attention module is added between two adjacent dense blocks to enhance the feature recognition effect. In our network, we use 3D dense blocks similar 2D dense networks. That directly connects the 3D output of any layer to all subsequent layers in the 3D dense block. The output feature-map of H_l in the l^{th} layer is given by:

$$x_l = H_l([x_0, x_1, \ldots x_{l-1}]) \qquad (1)$$

Where $[x_0, x_1, \ldots x_{l-1}]$ denotes that the features maps are concatenated. The spatial sizes of the x_i features maps are the same. The $H_l(.)$ is a composite function of BN-ReLU-3DConv operations.

3. Channel-Spatial Attention Model

The attention model is embedded in the deep network model, especially in dealing with image recognition and has achieved outstanding results. Since the attention mechanism uses selective focus on specific areas to better capture the desired results, we try to apply the attention mechanism to the video recognition task. In our attention model, we use the attention model which combines channel attention and spatial attention.

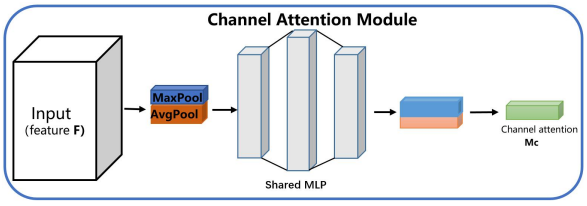

Fig. 4. Channel attention module

Channel attention module: More feature information for efficient attention, we use a continuity between feature channels to establish a channel attention mechanism model. Channel attention focused on image features,

similar to the process of selecting semantic attributes, focusing on "what" is meaningful. The structure can be found in Fig. 4. We use a combination of max-pooling and average-pooling to compress the spatial dimensions of the input feature map and improve the efficiency of the channel attention. After the two pooling layers, a hidden layer of multilayer perceptron is added to reduce the parameter overhead. The two descriptors are sent to the hidden layer, and then produce channel attention map Mc. In short, the channel attention is computed as:

$$M_c(F) = \sigma(MLP(F^c_{avg})) + (MLP(F^c_{\max})) \quad (2)$$

where σ denotes the sigmoid function.

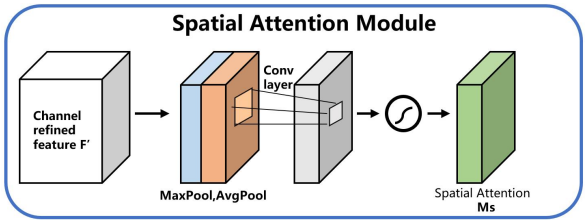

Fig. 5. Spatial attention module

Spatial attention module: As the channel attention cannot focus on detailed attention to features, we further focus on features to form a spatial attention model. Different from channel attention, spatial attention focuses on the "where" of feature information, and further focuses on the features of channel attention. The structure can be found in Fig. 5. After channel attention, the combination of max-pooling and average-pooling is used to generate more efficient feature descriptors. Then, we use the convolutional layer and the convolution produces a spatial attention map $M_s(F)$. use average-pooling and max-pooling to get two pooling features(F^s_{avg}, $F^s_{\max}$), and two pooling aggregate the channel information operations of the feature map to generate two 2D maps. After connecting through the convolutional layer, form our 3D attention map. In short, spatial attention is computed as:

$$M_s(F) = \sigma(f^{7*7}(F^s_{avg}; F^s_{\max})) \quad (3)$$

Where σ denotes the sigmoid function and f^{7*7} represents a convolution operation with the filter size of 7×7.

Channel attention and spatial attention, two attention modules, compute complementary attention, More exact focusing on the human behavior features in images. In the framework of this paper, we use a concatenation approach to combine channel attention and space attention, which has a better effect than the parallel.

IV. Experiments

First, we introduce the motion datasets used and the specific implementation details of the proposed model. We test and compare our proposed methods with Baselines and other state-of-the-art methods. Experiments verify the efficiency of our model and outperform other advanced models on large data sets.

1. Datasets

Experiments are conducted on two challenging video action datasets: UCF101 [3] and HMDB51 [2].

UCF101 dataset: The UCF101 dataset is a human motion video data, from the 101 types of real-world human motion videos clipped on YouTube, containing 13,320 videos. The videos in the 101 action categories are divided into 25 groups, each group containing multiple action videos. Videos from the same group have a similar background. The action mainly includes 5 types: people and objects interact, only body movements, people interact, play music equipment, all kinds of sports.

HMDB51 dataset: The HMDB51 database contains 6849 samples, divided into 51 categories, most of which come from movies, and some from public databases and online video libraries such as YouTube. These videos include human body movements, facial movements, general body movements, facial manipulations, and interactions with objects, etc.

2. Training

In our experiments, all the dataset subjects divided into training set and test set. The classifiers trained on a training set. We extract image data (25fps) frame by

frame from the input video and apply it to the video clip by using the center crop to unify the image size 224 × 224. Using stochastic gradient descent(SGD) [26] trained the networks. In the network settings, the initial learning rate set to 0.1 and decreased to 0.01 at 1/2 epochs, which 0.001 at 3/4 epochs. In our model, we used a weight decay of 10-4 and a Nesterov momentum [27] of 0.9 without dampening for all the weights. The network

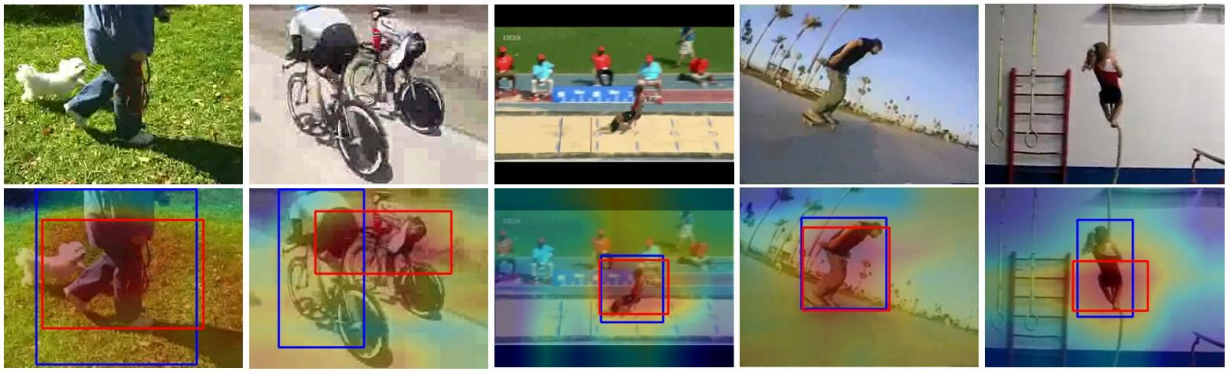

Fig. 6. Examples of attention. (Best viewed in color.) A frame from a video of action in UCF101. The top is the original image, spatial attention is shown as heatmap (Blue bounding boxes represent ground truth while the red ones are predictions from our learned spatial attention) in the bottom row. a: walking with dog, b: biking, c: long jump, d: skate boarding, e: rope climbing.

growth rate was set to 12 or 24. The batch size was set to 8 or 10. We trained this model for 100 epochs. Due to GPU memory constraints, the maximum crop size of the video (which is the height and width of the image) was set to 224 × 224.

Table 1. Exploration of ADD-Net(ours) and other 3D ConvNets on the UCF-101 dataset (split1)

Method	Accuracy %
ResNet3D-50	59.2
Inception3D	69.5
DenseNet3D	69.3
ADD-Net	71.5

Table 2. Accuracy (%) performance comparison of our method with other methods over all three splits of UCF101 and HMDB51

Method	UCF 101	HMDB 51
iDT+FV	85.9	57.2
C3D	82.3	56.8
Conv Fusion	82.6	56.8
Two-Stream	88.6	-
TSN-RGB	85.7	-
ResNet3D	86.1	55.6
DenseNet3D	88.9	57.8
Ours	91.3	60.5

3. Performance

Experiments show that our model has considerable experimental accuracy. In Fig. 6, our proposed attention-based 3D dense convolution model captures human motion features more efficiently in the ucf101 dataset. Table 1. show we verified in UCF101-split1 dataset that a 3D DenseNet using the attention model is more efficient than the original 3D dense convolutional network and other 3D networks. Table 2.shows the results on UCF101 and HMDB51 datasets for our model with other model-based action recognition methods. Our based on attention model 3D dense network method is significantly better than the 3D dense network method and has obvious advantages over other approaches. On both UCF101 and HMDB51 by 91.3% and 60.5% respectively.

We have not used the optical flow maps and still achieved good performance, which shows that our model is effective in feature capture and also reduces computational overhead.

V. Conclusion

In this paper, we discuss video-based human behavior recognition. In order to better improve the recognition effect, we have proposed a 3D densely concatenated convolutional network model based on attention mechanisms. Our models tested on the HMDB51 dataset and the UCF101 dataset. The experimental results show that our model is effective and still has advantages over other excellent models without using the two-flow model. Limited com-puting power due to excessive calculations, we regret that we do not use a two-stream 3D convolution network. If the optical flow maps data combined with the RGB data to form the dual-stream network, the recognition result will greatly be improved.

REFERENCES

[1] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.

[2] H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre.Hmdb 51: A large video database for human motion recognition. In High Performance Computing in Science and Engineering. 2013.

[3] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild.arXiv:1212.0402, 2012

[4] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In Proc. CVPR, 2016.

[5] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In Proc. CVPR, 2016.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolu-tional neural networks. In NIPS, 2012.

[7] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.

[8] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In P with improved trajectories. In Proc. ICCV, 2013.

[9] J. Ba, V. Mnih, and K. Kavukcuoglu, “Multiple object recognition with visual attention,” in Proc. Int. Conf. Learn. Represent, 2015.

[10] S. Sharma, R. Kiros, and R. Salakhutdinov, “Action recognition using visual attention,” in Proc. Int. Conf. Learn. Represent. Workshop, 2016.

[11] A. Diba, A. M. Pazandeh, and L. Van Gool. Efficient two stream motion and appearance 3d cnns for video clas-sification. In ECCV Workshops, 2016.

[12] A. Diba, V. Sharma, and L. Van Gool. Deep temporal linear encoding networks. In CVPR, 2017.

[13] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatio-temporal residual networks for video ac-tion recognition. In Advances in Neural Information Processing Systems, pages 3468-3476, 2016.

[14] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutio nal two-stream network fusion for video action recognition. In CVPR, 2016.

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In ECCV, 2016.

[16] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In CVPR, 2015.

[17] H. Wang and C. Schmid. Action recognition with improved trajectories. In ICCV, 2013.

[18] D. Tran, J. Ray, Z. Shou, S.F. Chang, and M. Paluri. Convnet architecture search for spatio-temporal feature learning. arXiv:1708.05038, 2017.

[19] L. Sun, K. Jia, D.Y. Yeung, and B. E. Shi. Human action recognition using factorized spa-tio-temporal convolutional networks. In ICCV, 2015.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.

[21] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolu-tional networks. In CVPR, 2017.

[22] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In ECCV, 2006.

[23] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L.Fei-Fei. Imagenet: A large-scale hierar-chical image database. In CVPR, 2009.

[24] Yang, H.; Yuan, C.; Li, B.; Du, Y.; Xing, J.; Hu, W.; Maybank, S.J. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2012.

[25] Tran, D.; Bourdev, L.D.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal fea-tures with 3d convolutional networks. In Proceedings of the 2015 International Conference on Computer Vision, Las Condes, Chile, 11-18 December 2015.

[26] Zinkevich, M.; Weimer, M.; Li, L.; Smola, A. Parallelized stochastic gradient descent. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6-9 December 2010; pp. 2595-2603.

[27] Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and mo-mentum in deep learning. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16-21 June 2013; pp. 1139-1147.

[28] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR,2017.

[29] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In BMVC, 2008.

[30] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

[31] K. Xu et al., “Show, attend and tell: Neural image caption

generation with visual attention," in Proc. ICML, 2015, pp. 2048-2057.
[32] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question an-swering," in Proc. ICCV, vol. 3, Oct. 2017, pp. 1300-1309.
[33] W. Zou, D. Jiang, S. Zhao, and X. Li, "A comparable study of modeling units for end-to-end mandarin speech recognition," arXiv preprint arXiv:1805.03832, 2018.
[34] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based endto-end speech recognition using multi-task learning," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 4835-4839.
[35] Rush, A. M. & Weston, J. A Neural Attention Model for Abstractive Sentence Summariza-tion. EMNLP , 2015.
[36] Kadlec, R., Schmid, M., Bajgar, O. & Kleindienst, J. Text Understanding with the Attention Sum Reader Network. arXiv:1603.01547v1 [cs.CL] , 2016.
[37] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507, 2017
[38] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X.,Tang, X.: Residual atten-tion network for image classification. arXiv preprint arXiv:1704.06904, 2017

Authors

Qiaoyue Man received his B.S. in computer engineering from Gachon University. He is currently pursuing his M.S. in computer engineering at Gachon University. His current re-search interests include Big Data analysis, Machine Learning and Smart City.

Young Im Cho received her B.S., M.Sc., and Ph.D. from the Department of Computer Science, Korea University, Korea, in 1988, 1990, and 1994, respectively. She is a professor at Gachon University. Her research interest includes AI, big data, information retrieval, smart city, etc.