제 13 회 학사학위 졸업논문
지도교수 홍현기 교수님

ConvLSTM-C2D Residual Network for Video Recognition.

박 철 우

School of Integrative Engineering
in Chung-Ang University

2020 . 06

제 13 회 학사학위 졸업논문
지도교수 홍현기 교수님

ConvLSTM-C2D Residual Network for Video Recognition.

이 논문을 학사학위 졸업논문으로 제출합니다.

중앙대학교 공과대학
융합공학부
박 철 우

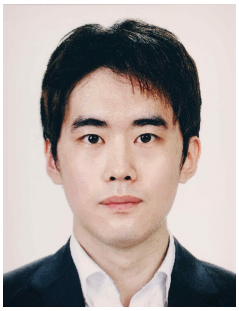
2020년 06월

중앙대학교 공과대학 융합공학과 박철우의 공학사 학위 취득을 위한
요구 사항 중 그 일부인 본 졸업논문을 인정함.

2020년 06월

지도교수 : _____ (인)

약 력



박철우는 1993년 대전시 대덕구 오정동에서 박봉희씨와 이영숙씨의 장남으로 태어났다. 2012년 명석고등학교를 졸업하고, 2015년도에 중앙대학교 공과대학 융합공학부로 입학하여 현재 4학년에 재학 중이며 2020년 08월에 졸업할 예정이다.

감사의 글

어느덧 짧지 않은 대학 생활을 마무리하며 지난 시간들을 돌이켜보니 많은 아쉬움과 후회가 남습니다. 학업적 성취에 있어서의 아쉬움만이 아닌, 고마운 많은 분들께 감사의 마음을 제대로 전하지 못했기에 더욱 그러한 것 같습니다. 제가 이렇게 성장하기까지 오랜 시간이 걸렸지만 그 세월 속에서 직·간접적으로 힘이 되고 방향을 잡아주셨던 많은 분들께 감사의 말씀을 전하고자 합니다.

먼저 본 논문이 완성되기까지 세심한 지도와 많은 격려로 이끌어 주신 홍현기 교수님께 진심으로 감사드립니다. 매 학기 마다 큰 열정으로 심도 있는 강의를 해주신 중앙대학교 융합공학부 디지털이미징공학과 교수님들께도 감사드립니다.

마지막으로 항상 사랑으로 키워주시고 부족한 자식을 믿어주신 부모님께 감사의 말씀을 드립니다. 언제나 제 편이 되어 힘을 주시고 바르게 생각하고 행동할 수 있도록 가르쳐주신 부모님께 누가 되지 않는 아들이 되기 위해 더욱 성장하도록 노력하겠습니다.

목 차

ConvLSTM-C2D Residual Network for Video Recognition.

1. Introduction

Studies to processing videos using the deep neural network have left its footprints in two ways – CNN based [1)2)5)] or CNN+RNN [3)4)5)] based networks. Although there are many pieces of evidence that CNN based networks perform well in video classification or anomaly detection tasks, I was motivated by a suspicion that these outcomes are still coming from a narrow understanding of each frame rather than integrating temporal domain. Because, on CNN, there is no structure designed for time series data.

Thanks to its excellent ability to understand images, the CNN based video representation model may seem to work well on formalized video datasets, even if it lacks understanding of time features, by grasping the characteristic features of each class in the image phase. If this shortcut truly has overwhelmed our evaluating process, When the parts of the network about the temporal domain are reinforced, this advanced model would do better at video comprehensions in terms of computing time and accuracy.

Under this assumption, I rebuilt a CNN based model by replacing CNN parts of layers that are responsible for the temporal domain with Convolutional LSTM. Because Convolutional LSTM cells have more structures to process time-series data. And then I compared the performance of this network with those of the original models in the aspects of video classification accuracy, computational cost, model size.

A computationally expensive model is usually considered hard to afford. But with more performance, it can reduce the iteration itself needed for training. In this way, it's possible to achieve a lower computational cost for real usage. So I didn't hesitate to adopt Convolutional LSTM cells.

2. Materials and Methods

2-1. Materials

2-1.a 3D Convolutions[1)2)5)]

In many precedent studies[1)2)5)], 3D convolutions are adopted to encode the spatio-temporal information. Let me suppose a clip has the size of l x h x w where l, h, w denotes the clip length, height, width respectively. 3D convolutional filter with d x k x k shape, where d means temporal depth and k means spacial kernel size, is stridden in l x h x w videos.

2-1.b Pseudo 3D block (P3D block) [3)]

In the P3D block, 3D CNN is replaced by 2D CNN and 1D CNN. The 2D filter encodes the spatial domain, the 1D filter encodes the temporal domain. P3D is more efficient than C3D thanks to its smaller filter size and computational cost. The author[3)] proposed three types of building blocks, as depicted in Figure 1.

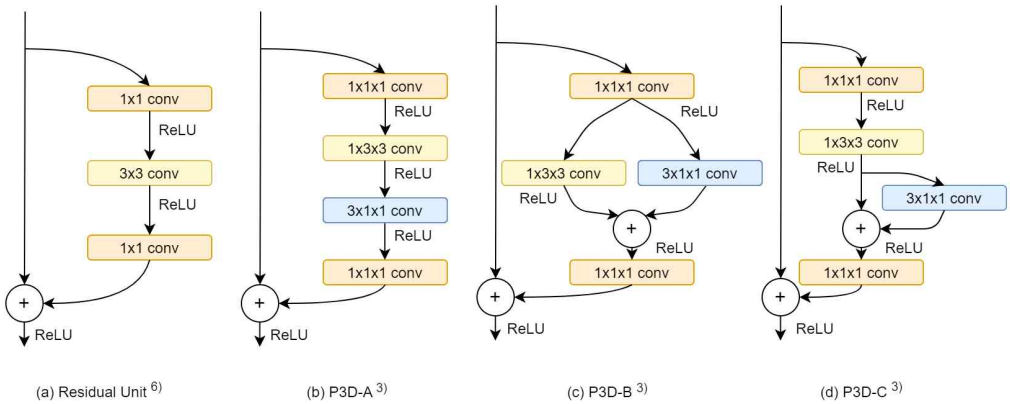


Figure 1. Three desighs of P3D blocks [3)] with Residual Unit [6)]

2-1.c ConvLSTM-C2D block

Inspired by ResNet and P3D[3] building block, I introduce a building block namely Convolutional LSTM - 2D CNN block (ConvLSTM-C2D block). In this structure, the 1D filters of the P3D block are replaced by Convolutional LSTM(ConvLSTM)[8]. The kernel size in ConvLSTM is 3x3 and step size is 1, that is an effective step. The plane sizes and channels of each ConvLSTM layers are the same as that of 2D CNN for the spatial domain in the same building block. The input of ConvLSTM is sliced and the sequentially fed. The output data of ConvLSTM is concatenated and returned to the next step.

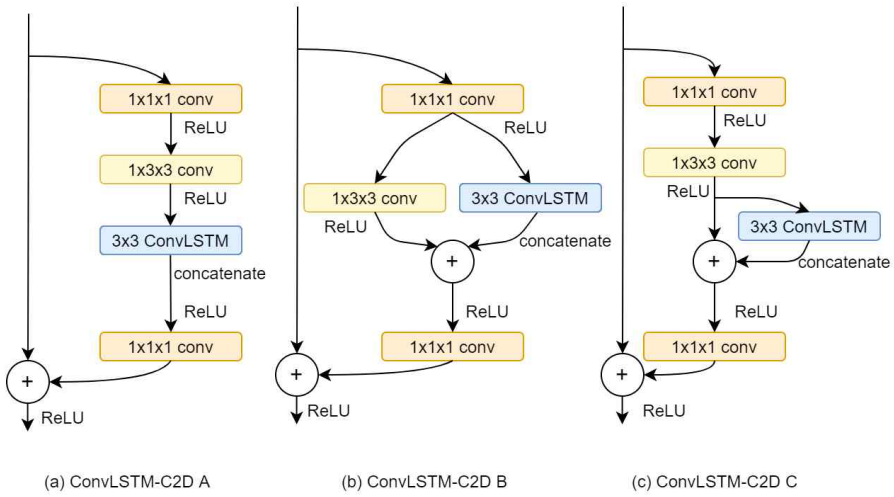


Figure 2. Three designs of ConvLSTM-C2D block.

2-1.d ConvLSTM-C2D Residual Network.

I built ConvLSTM-C2D Residual Network(ConvLSTM-C2D ResNet) that is modified from ResNet-50[6] using ConvLSTM-C2D building block.

Same as P3D ResNet[3], A, B, C type building blocks are placed sequentially. Before every ReLu activation functions, except between building blocks, batch normalization is adopted. I set batch normalization configuration following the configuration of P3D ResNet[3].

There is one dropout layer with a dropout rate of 0.5 after the fully connected layer. Compared to P3D ResNet, the size of ConvLSTM-C2D ResNet has decreased from 261 MB to 236 MB.

	Architecture	
Layer	P3D ResNet(50)	ConvLSTM-C2D ResNet(63)
conv1	7x7,64,stride 2	1x7x7,64,stride (1,2,2) [3]
	3x3 max pool, stride 2	2x3x3 max pool, stride 2 [3]
conv2_x	[1x1,64 3x3,64 1x1,256] x3	1x1x1,64 [1x3x3,64 ConvLSTM 3x3,64 (concatenate)] x3 1x1x1,256
conv3_x	[1x1,128 3x3,128 1x1,512] x4	1x1x1,128 [1x3x3,128 ConvLSTM 3x3,128 (concatenate)] x4 1x1x1,512
conv4_x	[1x1,256 3x3,256 1x1,1024] x6	1x1x1,256 [1x3x3,256 ConvLSTM 3x3,256 (concatenate)] x6 1x1x1,1024
conv5_x	[1x1,512 3x3,512 1x1,2048] x3	[1x1x1,512 1x3x3,512 1x1x1,2048] x3
	average pool, 1000-d fc, softmax	average pool, 101-d fc, softmax

2-1.e DataSet[7]

Training, evaluation, and comparison are conducted on UCF101[7] video action recognition dataset. It consists of 101 action classes, over 13k clips, and 27houres of video data. Three training/test splits are provided by the organizer and each split includes about 9.5k training data and 3.7K test videos. During the training and testing phase, Three splits of training/test set are used as the organizer provide.

Figure 3. 101Classes in UCF101 Dataset

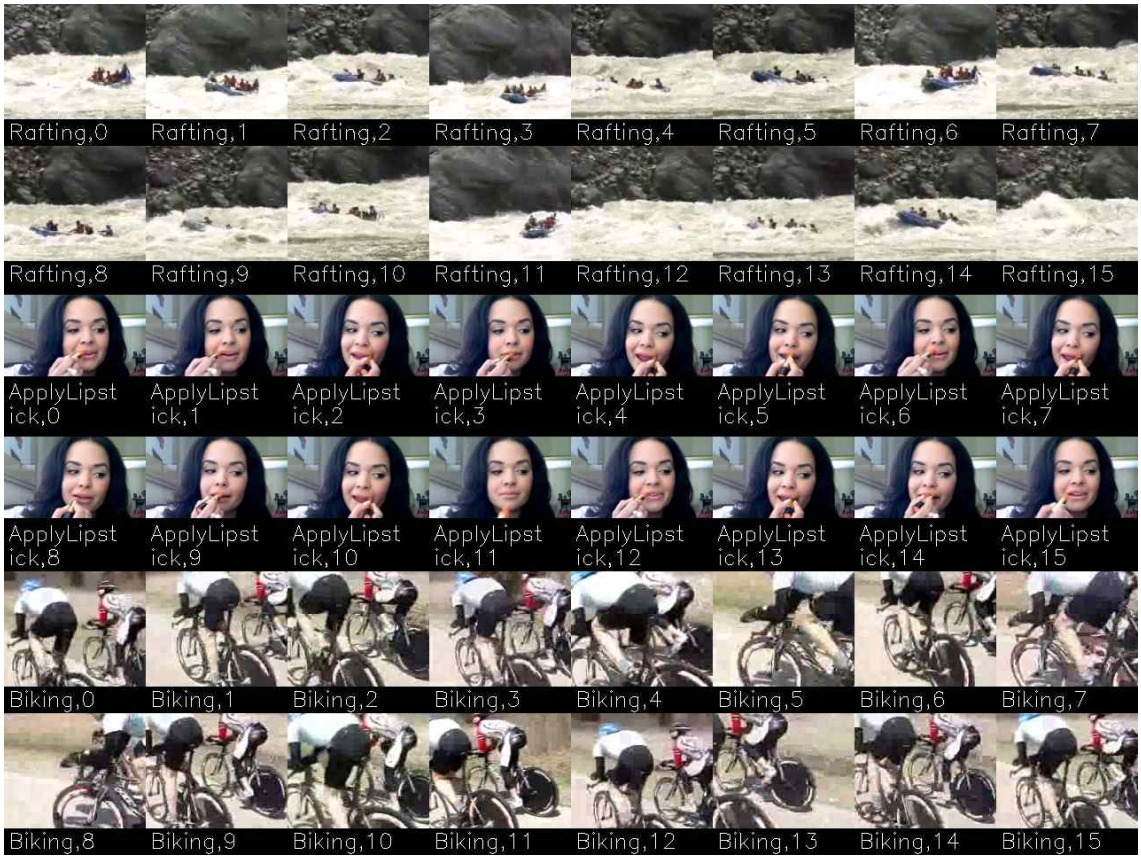


Figure 4. After preprocessing

2-2. Methods

2-2.a Training

I trained the ConvLSTM-C2D ResNet using 6 GTX1080ti GPUs with a batch size of 48. Fully Connected Layer is used for the classification task then cross-entropy loss is calculated. Fitting networks is performed using root mean square propagation optimizer. Due to the ConvLSTM cells, this model needs more VRAM than P3D ResNet. So that I needed to train the P3D ResNet in the same training condition for the comparison of the performances of each one. And then the performance of ConvLSTM-C2D ResNet is Compared with existing networks.

The dataset is preprocessed in 16frames per video. Because the running times are different, the frame rate is varied. I cropped the center 192x192 area from 240x320 frames. Then I resized it to

160x160. Data augmentation is not used. The learning rate is 0.0001 and I didn't use the learning rate scheduling. I didn't pretrain both models.

2-2.b Training Phase

ConvLSTM-C2D ResNet(63) took 8 hours 56minutes to finish the training phase and one epoch took about 39min. P3D ResNet(199) took 15 hours 25 minutes for training and one epoch took about 13min.

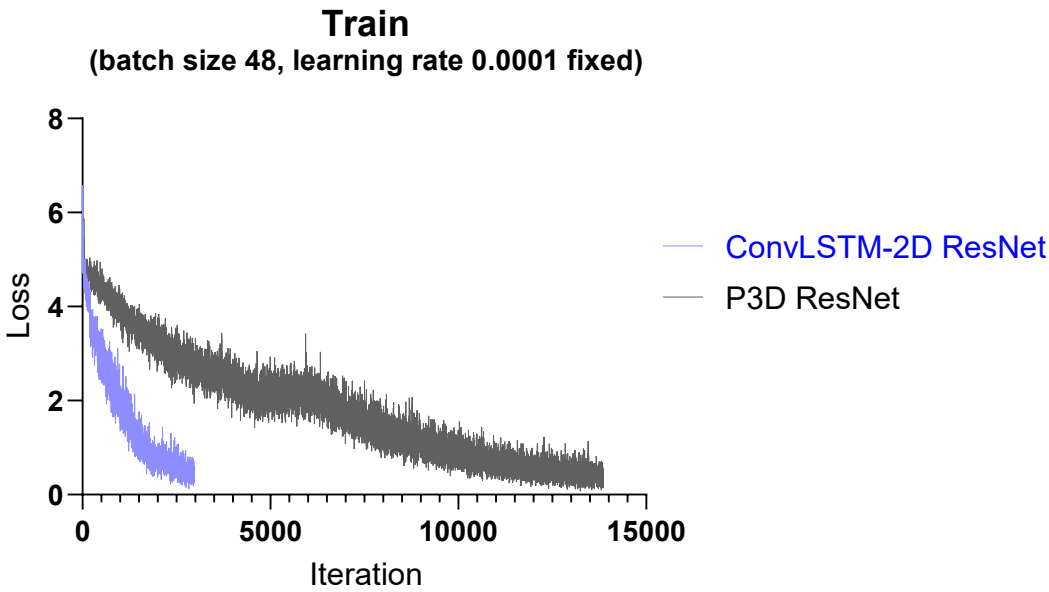


Figure 5. Comparision of the training phase of two types of building blocks

2-2.c Testing - Comparision with P3D ResNet.

I evaluated ConvLSTM-C2D ResNet adopting a single clip that is cropped from the center of the video clip.

Method	Classifier	Accuracy
C3D [2]	linear(SVM, Pretrained) [2]	82.30% [2]
ResNet-152 [3]	linear(SVM, Pretrained) [3]	83.50% [3]
P3D ResNet [3]	linear(SVM, Pretrained) [3]	88.60% [3]
P3D ResNet	linear(Cross Entropy, Not pretrained)	88.79%
ConvLSTM-C2D ResNet	**linear(Cross Entropy, Not pretrained)**	**95.57**

2-2.d Inferencing

2-2.d Inferencing

ConvLSTM-C2D ResNet(63) process 36.9~37 videos per a second on one Nvidia GTX1080ti and Intel I7 6950X. P3D-ResNet was 194.1~195 video clips per second under the same circumstance. Even though ConvLSTM-C2D takes more time to process clips, considering P3D ResNet(199) needs to be launched 21[3)] times for one video clip, The actual capacity of P3D-ResNet is 9.2 clips per second.

3. Result and Disccusion

ConvLSTM-C2D ResNet(63) with Linear classifier and Cross entropy loss function took 8 hours 56 minutes to finish the training. P3D ResNet(199), on the other hand, took 15 hours 25 minutes under the same conditions. Meanwhile, the accuracy of P3D ResNet(199) was 6.97% higher than that of P3D ResNet.

The inference time per video of ConvLSTM-C2D ResNet(63) was about 5 times longer than that of P3D ResNet(199). However, considering that B requires 21 inferences per video for the above accuracy, in practical circumstances, ConvLSTM-C2D ResNet(63) took only 24% of the P3D ResNet(199) inferencing time.

In terms of accuracy and efficiency, ConvLSTM-C2D ResNet, which processes time series data with RNNs, is significantly improved over P3D ResNet.

4. Conclusion

Because RNN is computationally intensive, it is rarely used in the process of improving model efficiency. However, I got successful results in terms of efficiency and accuracy by entrusting the temporal domain

to RNNs than CNNs. The amount of computation increases in one inference and one learning epoch, but it is possible to achieve higher accuracy with less computation in the overall learning phase and actual prediction.

Applying a computationally intensive but more suitable model to the right place was beneficial not only for accuracy but also for reducing computation time in practical.

5. References

1) Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei (2014) Large-scale Video Classification with Convolutional Neural Networks. CVPR, 2014

2) D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. (2015) Learning spatiotemporal features with 3d convolutional networks. In ICCV, 2015.

3)Zhaofan Qiu , Ting Yao , and Tao Mei, University of Science and Technology of China, Hefei, Microsoft Research Beijing China (2017) Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, arXiv:1711.10305v1 [cs.CV]

4) Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang, Shanshan Hao (2019) I3D-LSTM: A New Model for Human Action Recognition. IOP Conf. Ser.: Mater. Sci. Eng. 569 032035 2019

5) Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh. (2018) Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. CVPR 2018

6) HE, Kaiming, et al. (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

7) K. Soomro, A. R. Zamir, and M. Shah. (2012) UCF101: A dataset of 101 human action classes from ideos in the wild. CRCVTR-12-01, 2012.

8) Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, Wang-chun Woo. (2015)Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. nips 2015