NORCE

Attention Boosted Deep Networks for Video Classification

Junyong You, Norwegian Research Centre (NORCE), Bergen, Norway

Jari Korhonen, Shenzhen University, Shenzhen, China

AbuDhabi ICIP 2020

IEEE International Conference on Image Processing
25-28 October 2020, United Arab Emirates

IEEE Signal Processing Society

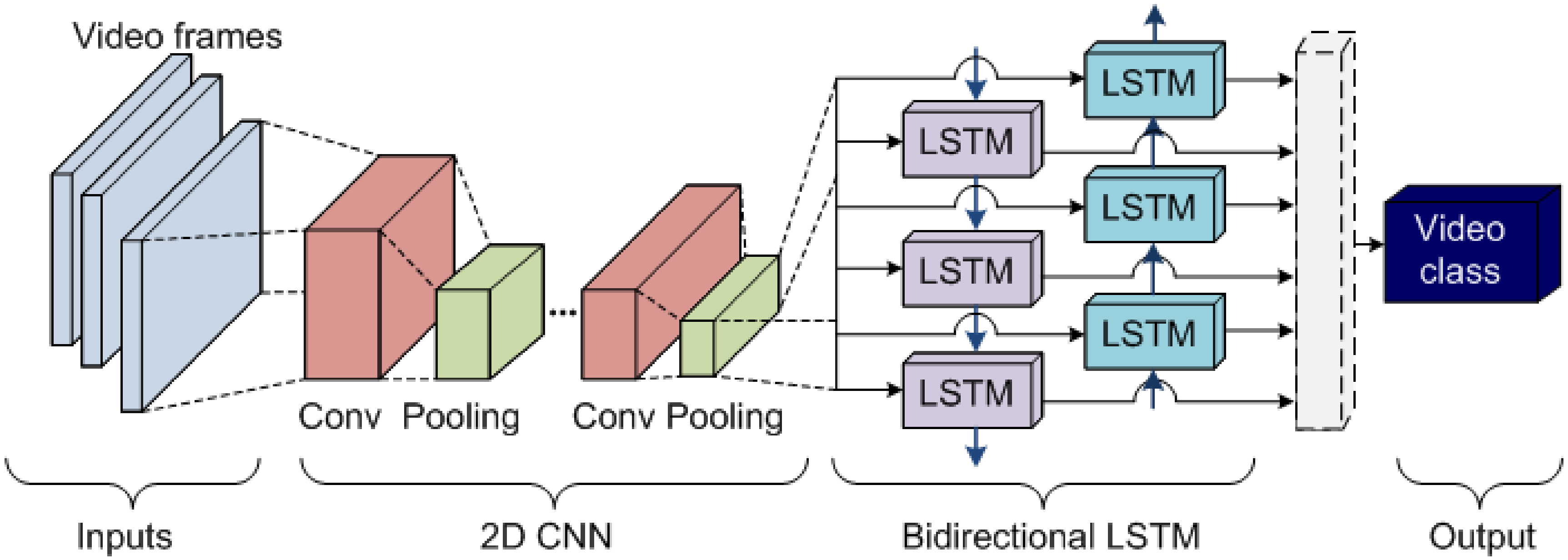
Attention mechanism

- HVS cannot process all visual information
- “Attend to” a certain part of visual stimuli while ignoring other perceptible information
- Attention in deep learning:
  - NLP (e.g., Transformer), visual signal process, etc.
  - Two commonly used attention functions:
    - Additive attention (D. Bahdanau et al.)
    - Dot-product

CNN and bi-LSTM for video classification (I)

- 2D-CNN serves as frame feature extractor
  - VGG / Inception / Resnet / Xception
  - ImageNet pretrained
- Video classes determined by frame contents and their relationships – modelled by LSTM
  - Viewers can retrospect the content in a reverse time order to obtain the full context when classifying video content – bidirectional LSTM
- **Main architecture: 2D-CNN + bi-LSTM**

CNN and bi-LSTM for video classification (II)

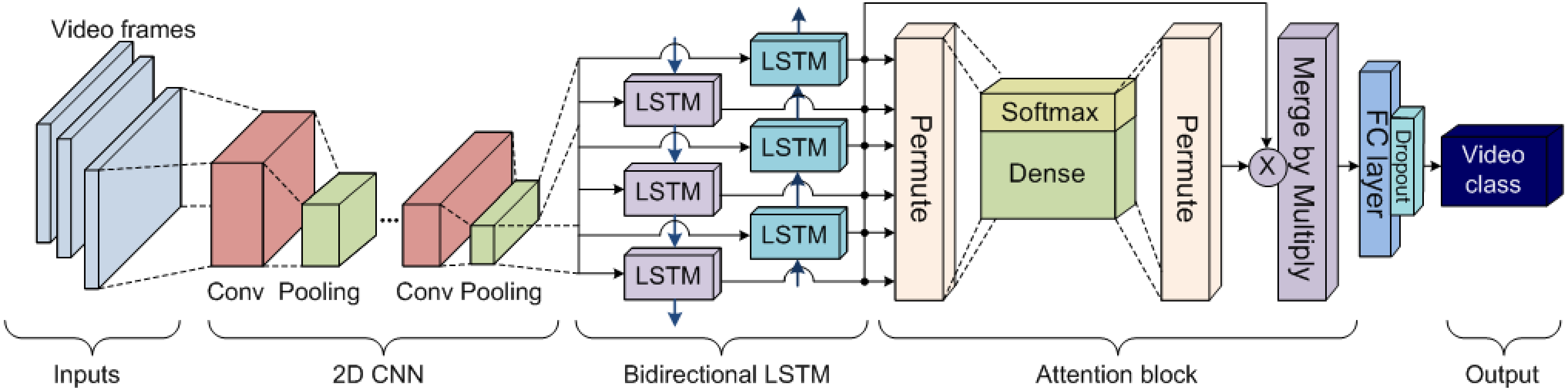


Simple attention block

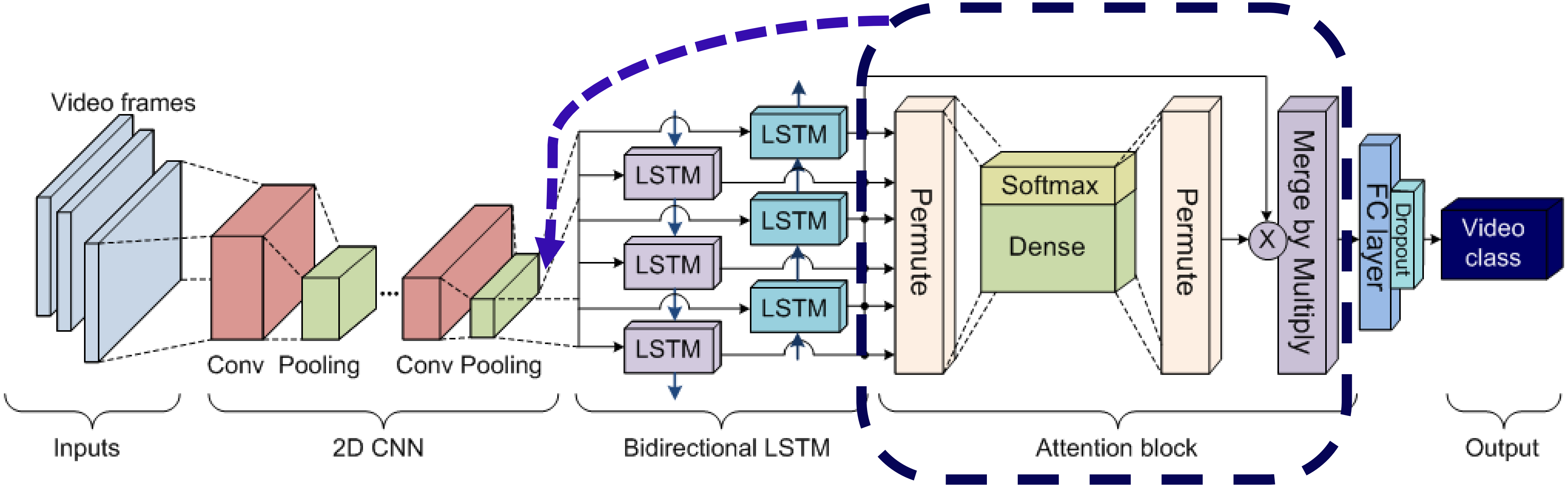
- Attention mechanism
  - Pay different attentions to different parts of input

- Can be modelled by a dense (fully-connected) layer using "softmax" as activation
  - Dense layer with same length as the input (output of bi-LSTM) length
  - Softmax limits the weights within (0, 1) with sum = 1

Attention integrated networks for video classification



Attention integrated networks for video classification

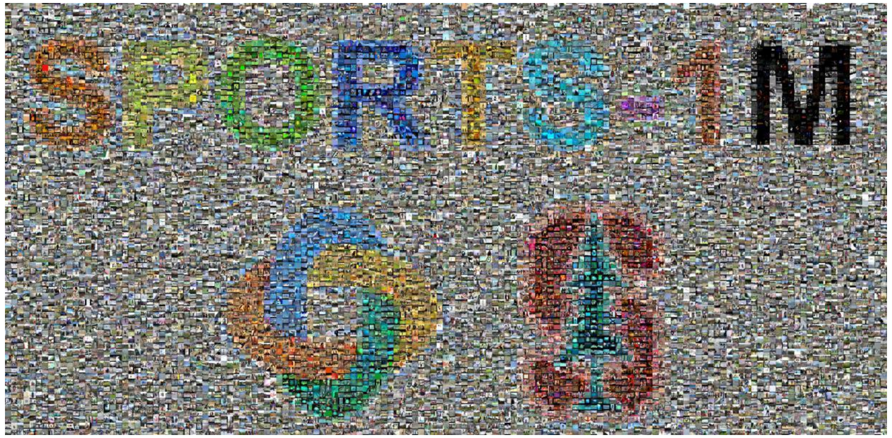


- Attention block can also apply prior to bi-LSTM layer

Network hyper-parameters

- A single bi-LSTM layer
- Unit number 256 chosen from selections [64, 128, 256, 512] in the experiments
- One frame / second employed due to frame redundancy
- Unit number in the dense layer in attention block is the average of all video frame numbers (temporal dimension of input)
- Unit number 512 for the last FC layer
- Dropout rate = 0.5

Experiment: Datasets



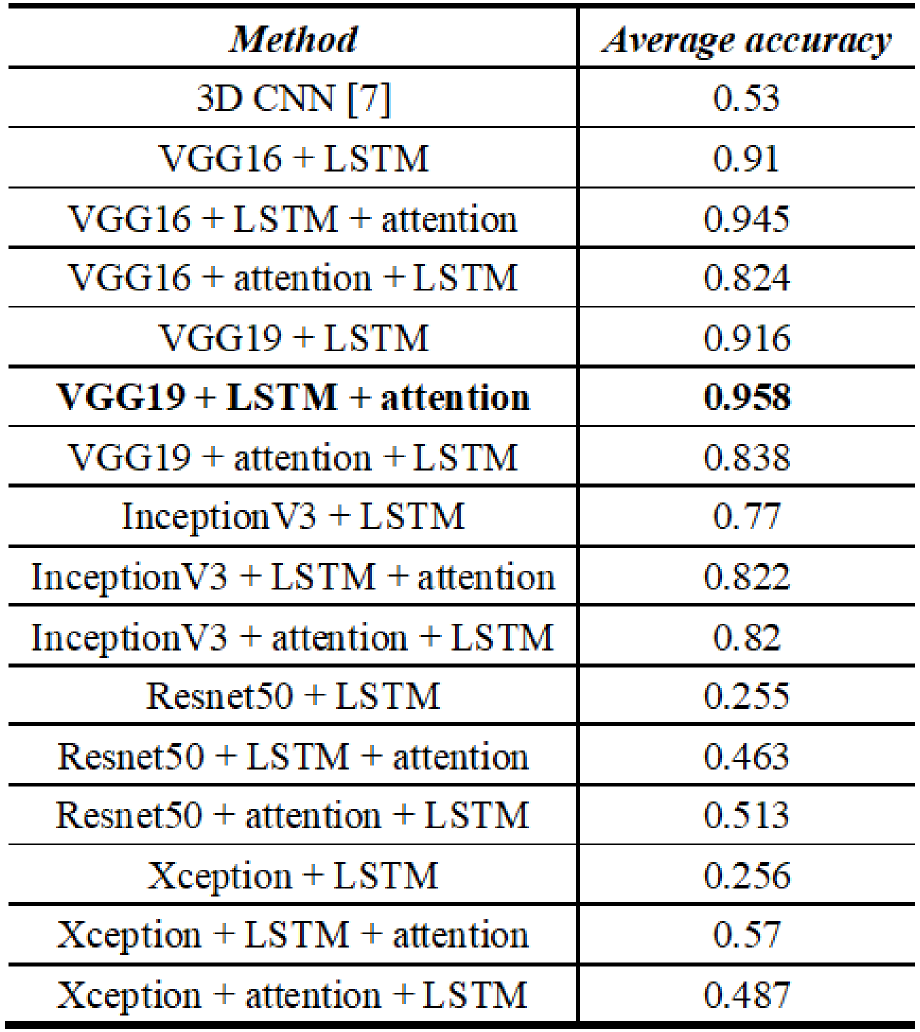
- UCF-101 action recognition dataset
  - 13,320 videos with 101 categories
- Subset of Sports-1M dataset (Sports-1M-99)
  - Video shorter than 20s in the first 99 categories from original 202 categories
  - Each category contains more than 100 videos
  - In total 18,319 videos

Experiments: Other models

- 3D-CNN model (S. Ji, W. Xu, M. Yang, and K. Yu, “3d Convolutional Neural Networks for human action recognition,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013)
- Variant CNNs: VGG16, VGG19, InceptionV3, Resnet50, Xception
  - CNN + attention + LSTM: attention prior to bi-LSTM
  - CNN + LSTM + attention: attention after bi-LSTM

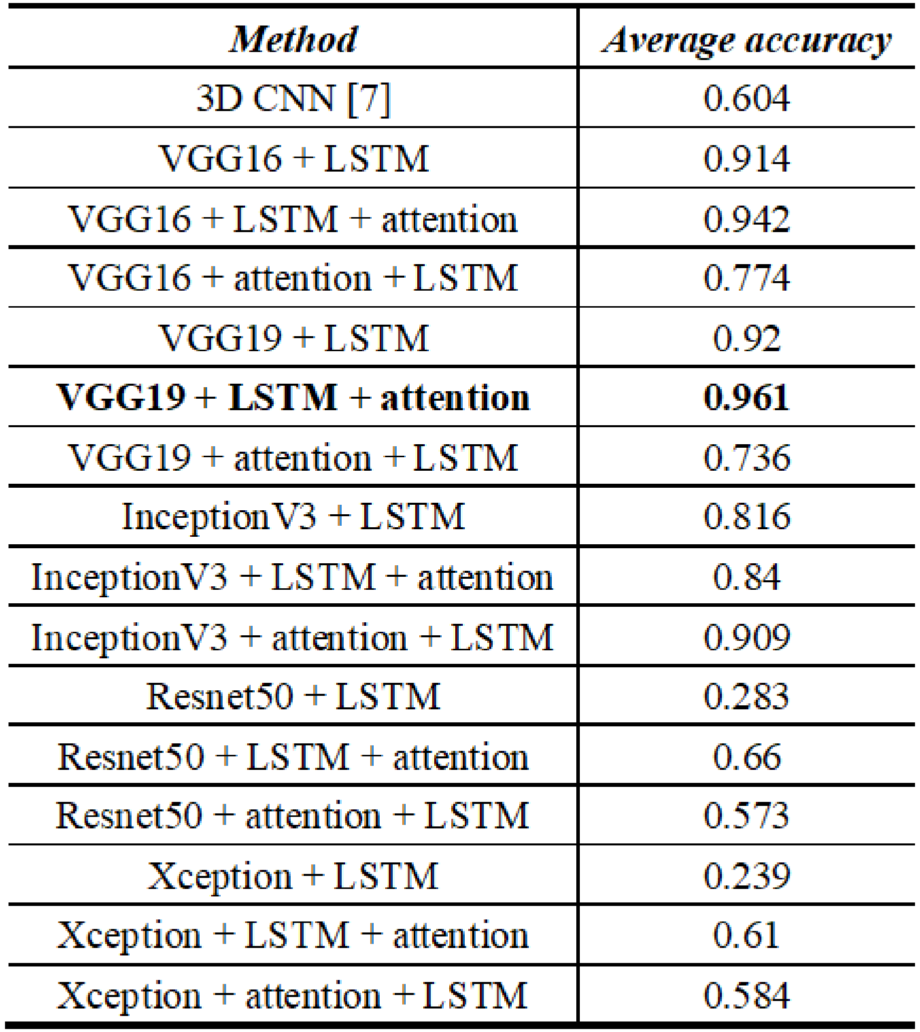
Experiments: Evaluation results

UCF-101



Method	*Average accuracy*
3D CNN [7]	0.53
VGG16 + LSTM	0.91
VGG16 + LSTM + attention	0.945
VGG16 + attention + LSTM	0.824
VGG19 + LSTM	0.916
VGG19 + LSTM + attention	**0.958**
VGG19 + attention + LSTM	0.838
InceptionV3 + LSTM	0.77
InceptionV3 + LSTM + attention	0.822
InceptionV3 + attention + LSTM	0.82
Resnet50 + LSTM	0.255
Resnet50 + LSTM + attention	0.463
Resnet50 + attention + LSTM	0.513
Xception + LSTM	0.256
Xception + LSTM + attention	0.57
Xception + attention + LSTM	0.487

Sports-1M-99



Method	*Average accuracy*
3D CNN [7]	0.604
VGG16 + LSTM	0.914
VGG16 + LSTM + attention	0.942
VGG16 + attention + LSTM	0.774
VGG19 + LSTM	0.92
VGG19 + LSTM + attention	**0.961**
VGG19 + attention + LSTM	0.736
InceptionV3 + LSTM	0.816
InceptionV3 + LSTM + attention	0.84
InceptionV3 + attention + LSTM	0.909
Resnet50 + LSTM	0.283
Resnet50 + LSTM + attention	0.66
Resnet50 + attention + LSTM	0.573
Xception + LSTM	0.239
Xception + LSTM + attention	0.61
Xception + attention + LSTM	0.584

Analysis and Conclusion

- Integrating attention can generally boost CNN + LSTM for video classification
  - Attention after LSTM better before
  - Probably due to dimension difference of input for attention block
- VGG16/19 > InceptionV3 > Resnet50 > Xception
- Attention before LSTM reduce accuracy on VGG16/19
  - Suspect due to late selection theory of attention mechanism
- CNN + LSTM better than 3D-CNN
  - No pretrain of 3D CNN
  - LSTM might better than 3D CNN on capturing long-term connections of frames

Source code published

N O R C E

SHENZHEN UNIVERSITY 深圳大學 1983

https://github.com/junyongyou/Attention-boosted-deep-networks-for-video-classificaton

Welcome to download and use.

Thank you for your attention!