Revisiting the Effectiveness of Off-the-shelf Temporal Modeling Approaches for Video Recognition

July, 2017

Chuang Gan

Tsinghua University

Our Baidu&Tsinghua Team

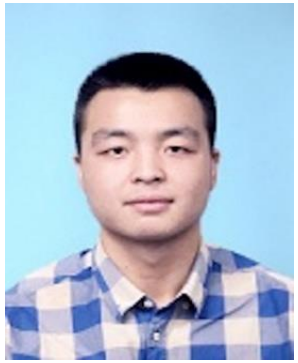
Chuang Gan

Xiao Liu

Fu Li

Yunlong Bian

Xiang Long



Heng Qi

Jie Zhou

Shilei Wen

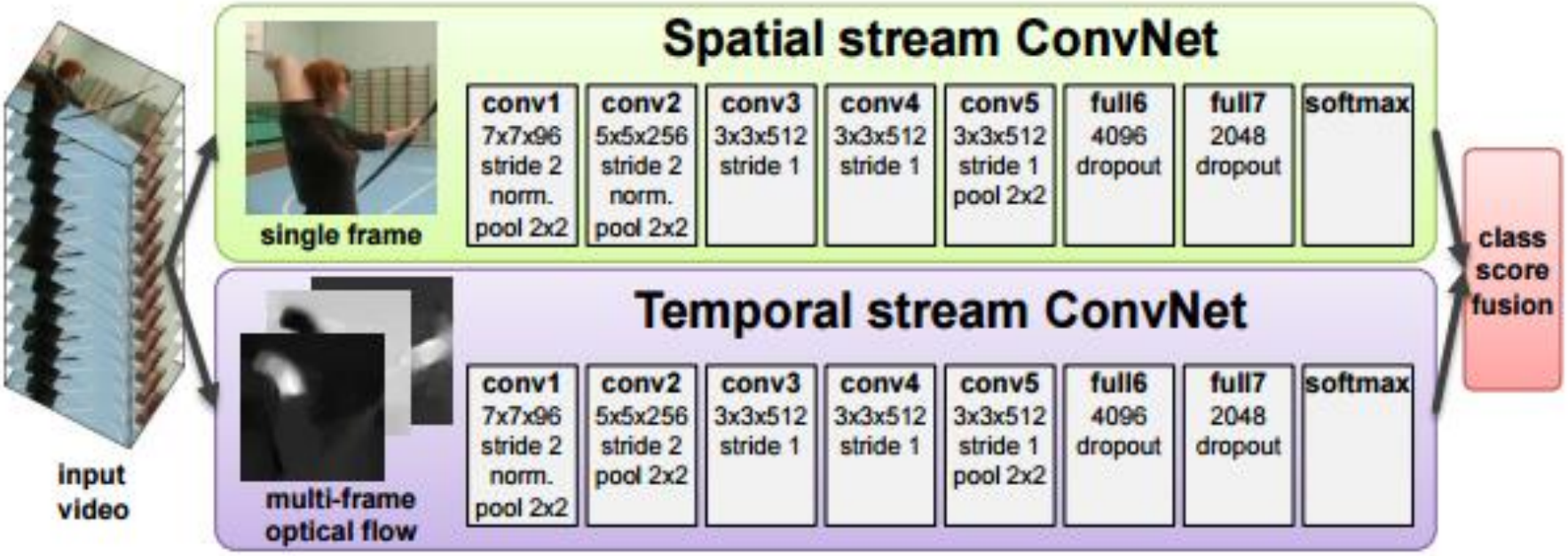
Outline

- **Temporal Modeling Approaches**
  - ✓ Background
  - ✓ Proposed approach
  - ✓ Experiment results
  - ✓ Conclusions and Discussion

Outline

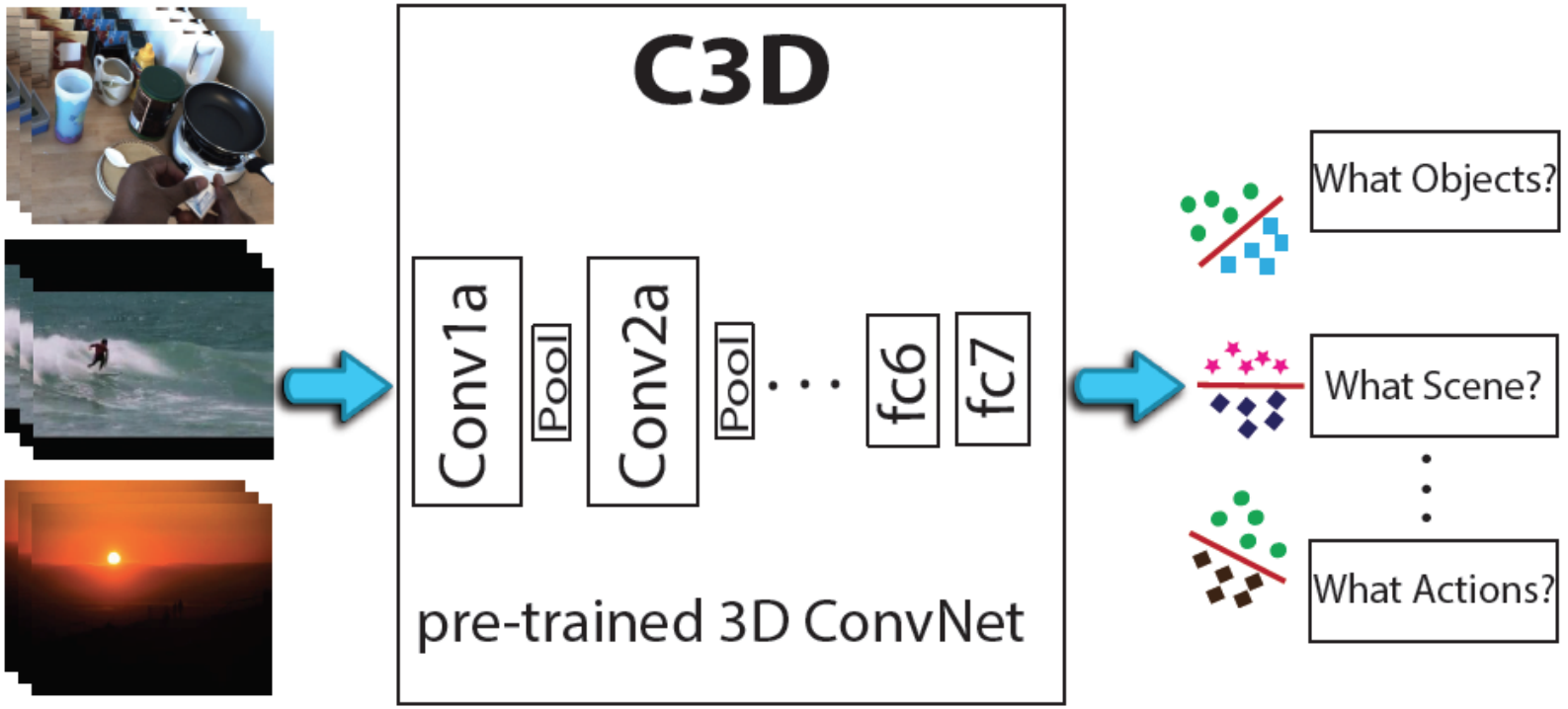
- **Temporal Modeling Approaches**
  - ✓ Background
  - ✓ Proposed approach
  - ✓ Experiment results
  - ✓ Conclusions and Discussion

Background



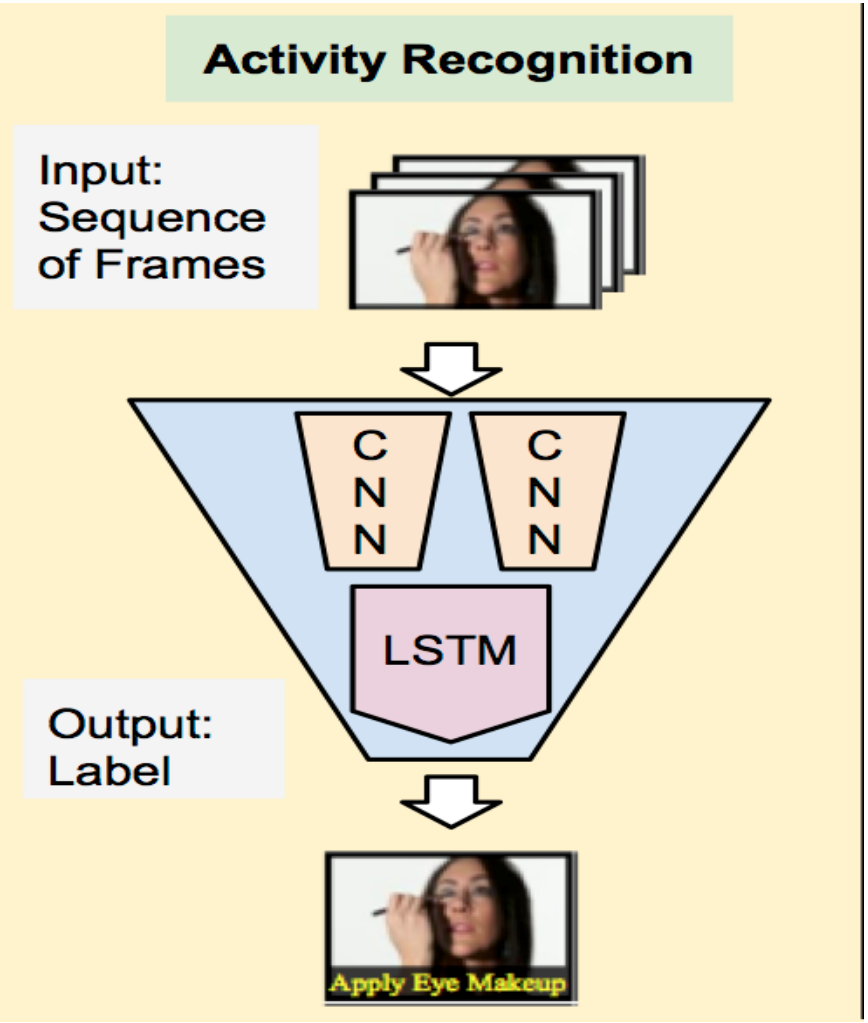
Two-Stream Convolutional Networks for Action Recognition in Videos. NIPS'14

Background



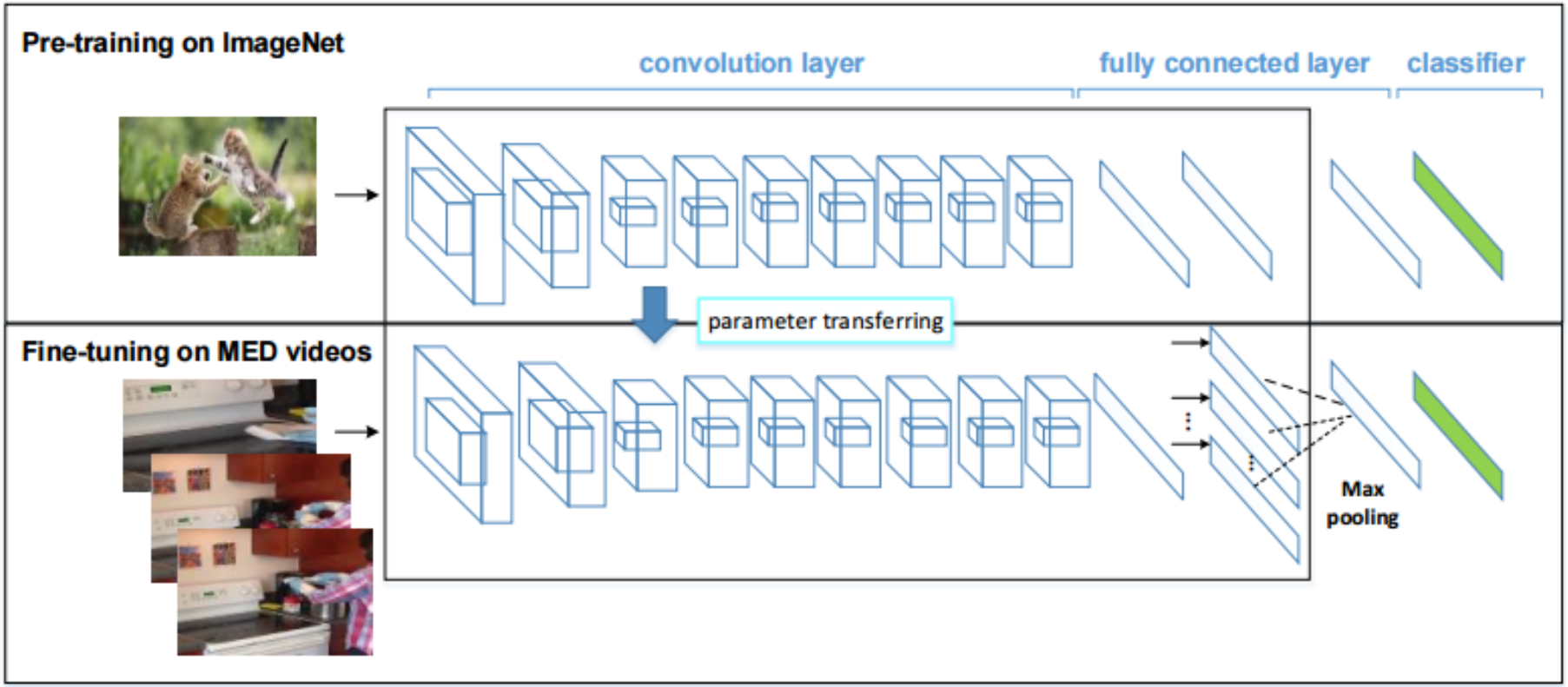
Learning Spatiotemporal Features With 3D Convolutional Networks. ICCV' 15

Background



Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR'15

Background

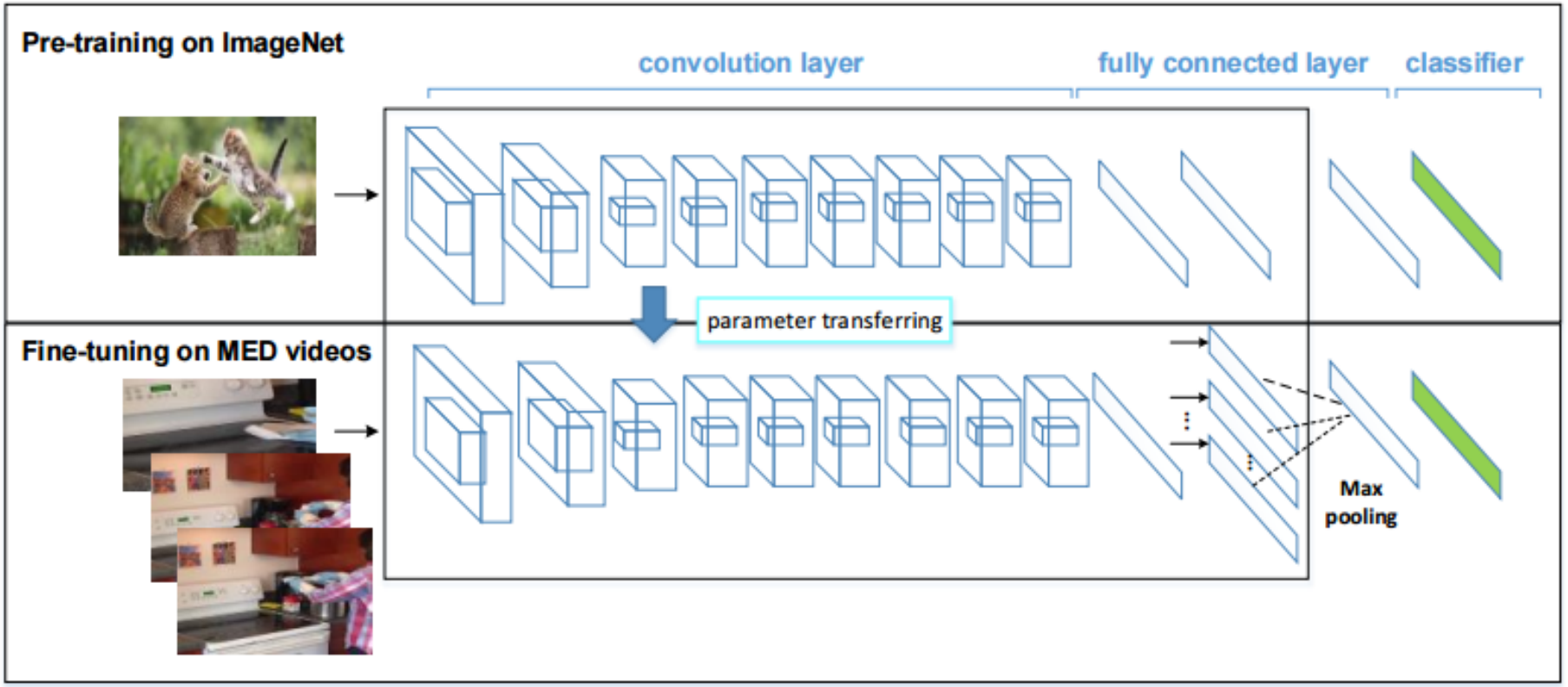


DevNet: A Deep Event Network for Multimedia Event Detection and Evidence Recounting. CVPR2015

Outline

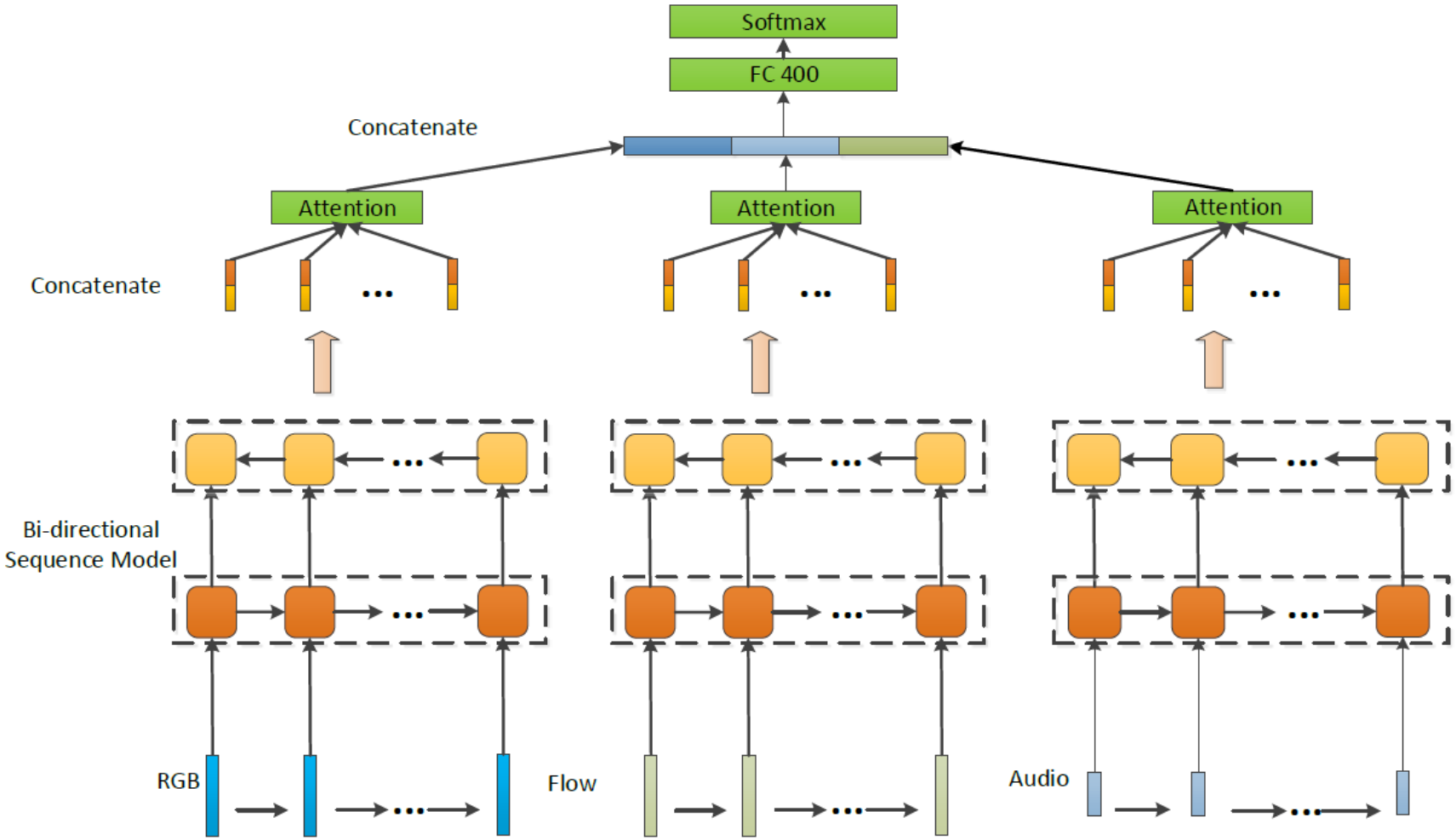
- **Temporal Modeling Approaches**
  - ✓ Background
  - ✓ Proposed approach
  - ✓ Experiment results
  - ✓ Conclusions and Discussion

Using the Fine-tuning features!!

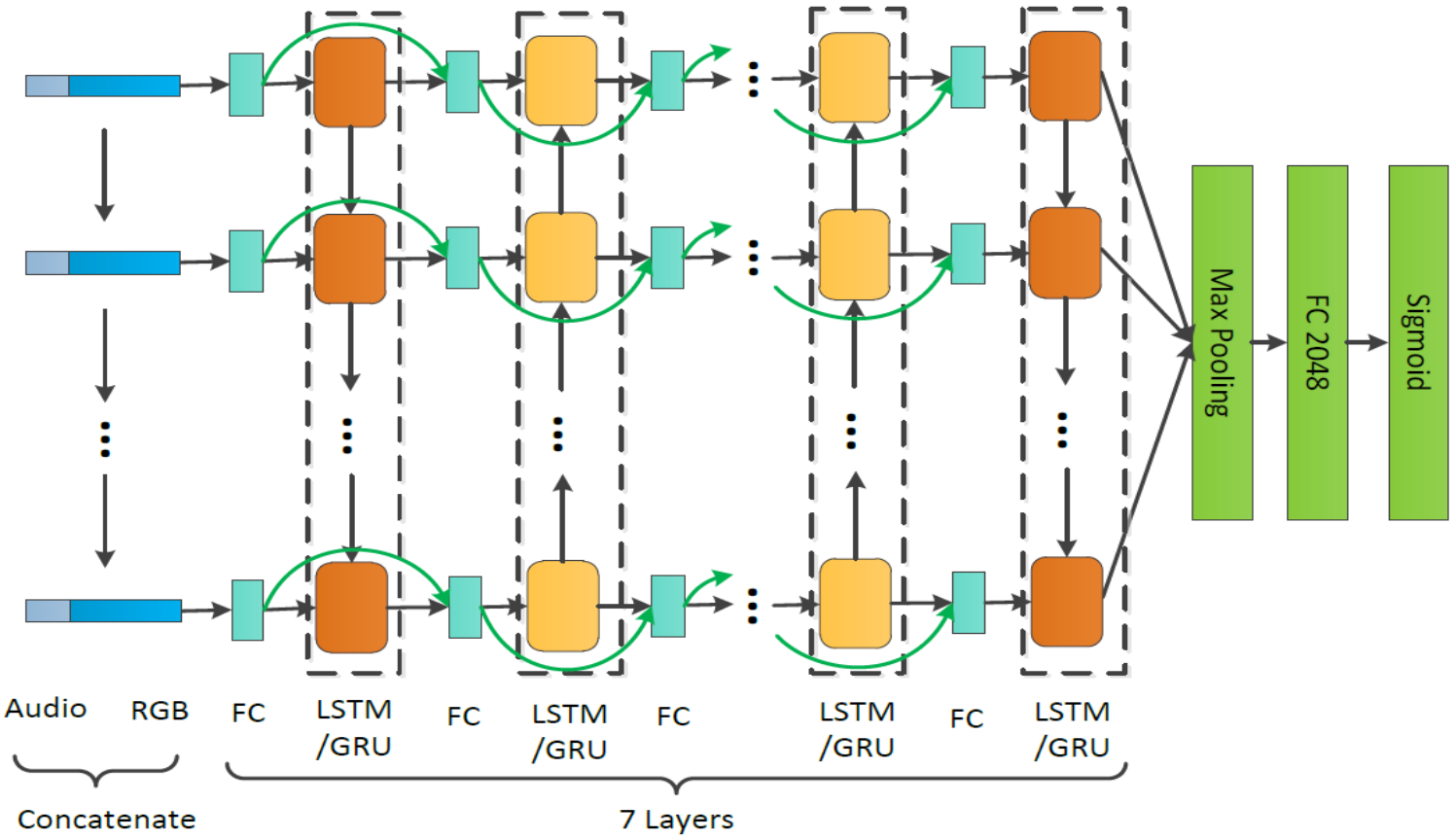


DevNet: A Deep Event Network for Multimedia Event Detection and Evidence Recounting. CVPR2015

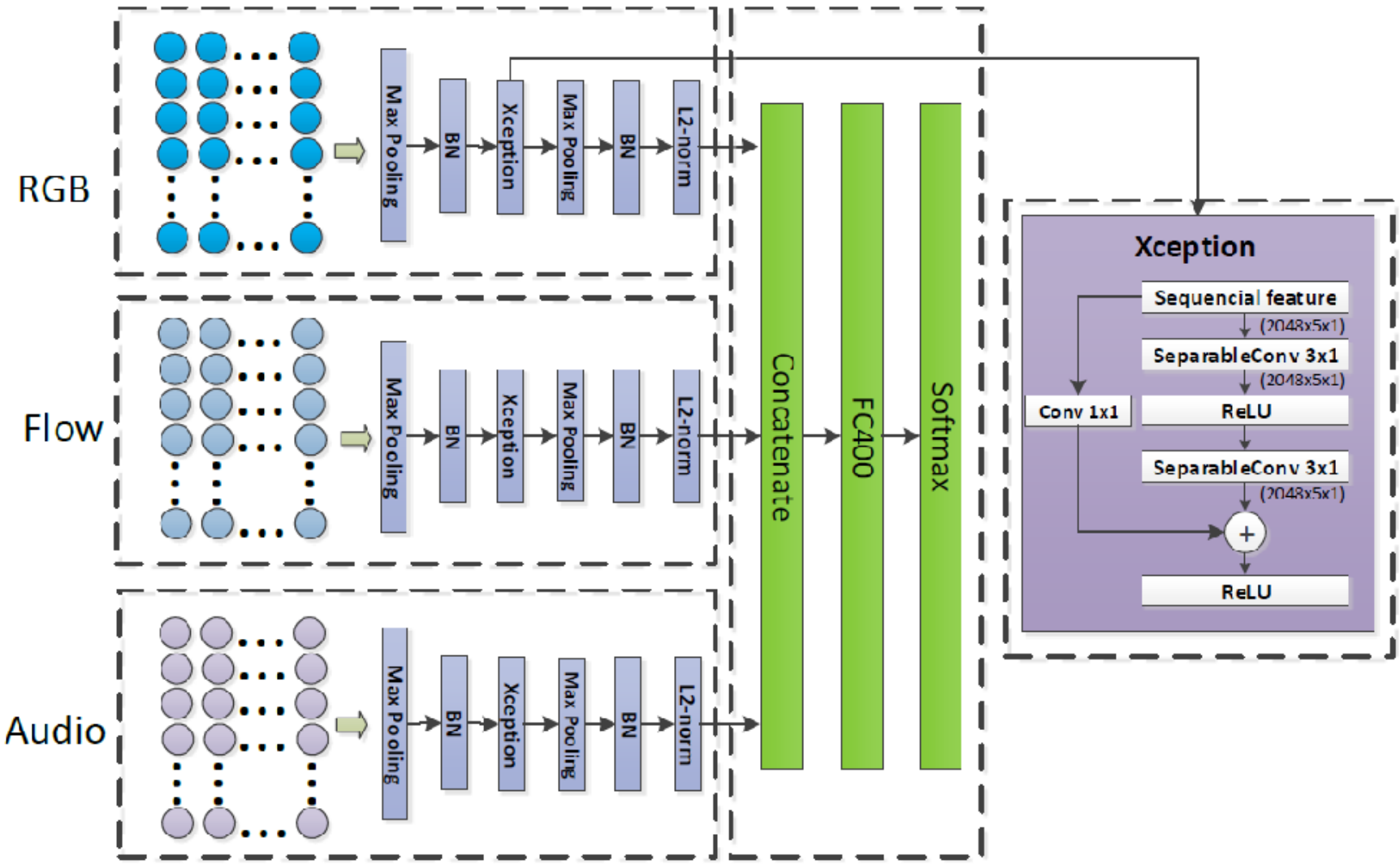
Multi-stream Sequence Model



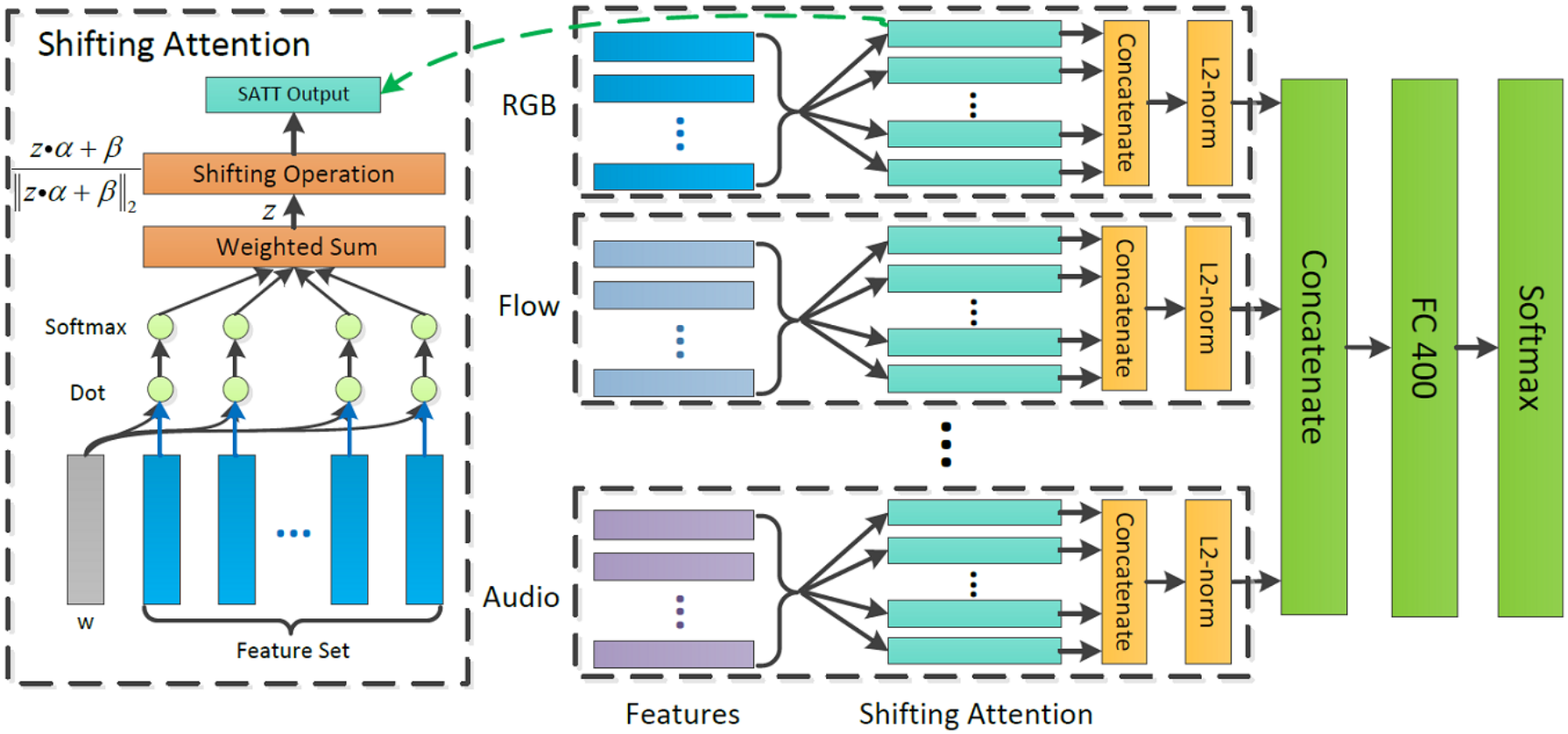
Fast-forward Sequence Model



Temporal Xception Network

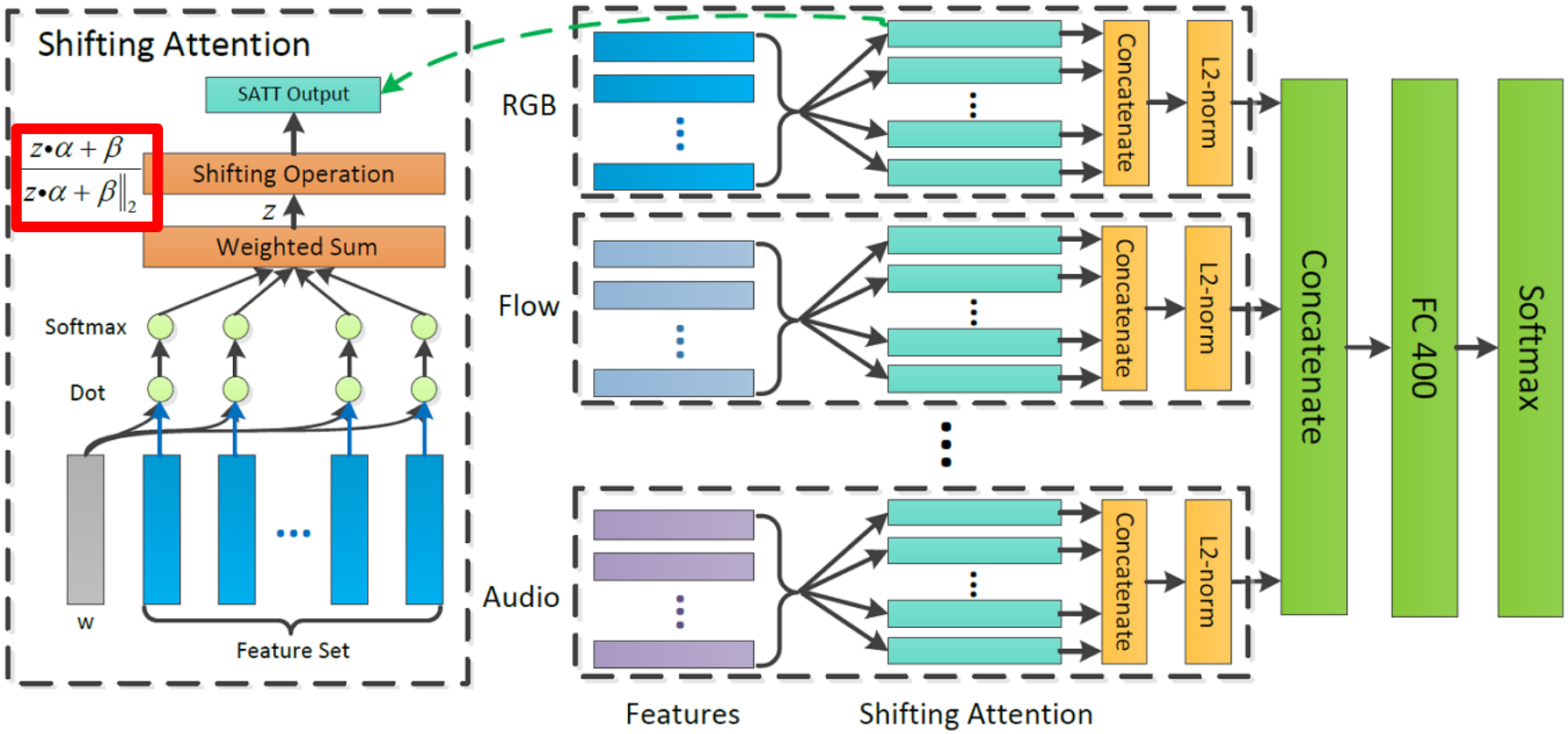


Shifting Attention Network



Secret Weapon for our winner solution on ActivityNet 2017

Shifting Attention Network



Secret Weapon for our winner solution on ActivityNet 2017

Results on the Validation Set

Approach	Top1 Acc. (%)	Top5 Acc. (%)
RGB	73.0	90.9
Flow	54.5	75.9
Audio	21.6	39.4
Three-stream fusion	74.9	91.6
Multi-stream LSTM	77.0	93.2
Fast-forward LSTM (Depth 7)	77.1	93.2
Temporal Xception	77.2	93.4
Shifted Attention	77.7	93.2
Ensemble	81.5	95.6

Results on the Validation Set

Approach	Top1 Acc. (%)	Top5 Acc. (%)
RGB	73.0	90.9
Flow	54.5	75.9
Audio	21.6	39.4
Three-stream fusion	74.9	91.6
Multi-stream LSTM	77.0	93.2
Fast-forward LSTM (Depth 7)	77.1	93.2
Temporal Xception	77.2	93.4
Shifting Attention	77.7	93.2
Ensemble	81.5	95.6

Take-home Message

- ✓ Multi-stream sequence model is an effective way to leverage multi-modality features.
- ✓ The fast forward connections is important to increase the depth of LSTM.
- ✓ Temporal convolution is an alternative approach for temporal modeling.
- ✓ Shifted attention is a very effective temporal modeling approach.

Acknowledgement

Thanks for the support from Baidu IDL team!

Thank You!