A general overview of SpatioTemporal Fusion Techniques for Human Activity Recognition

Sarosij Bose
Department of Computer Science and Engineering, University of Calcutta.

Introduction

Sarosij Bose: I am currently a 3rd year undergraduate pursuing my BTech in Computer Science and Engineering from University of Calcutta. My interests lie in Computer Vision and Machine Learning.

For the past 7 months, I have been studying and working on Video understanding and it’s applications in Human Action Recognition (HAR) under the guidance of **Professor Amlan Chakrabarti**.

Problem statement and Motivation

- **Problem**: Human Activity Recognition (HAR) from videos.
- **Challenges:-**

1. Image based approaches unable to capture temporal information.
2. 3D CNNs too bulky and expensive.
3. Exclusive temporal data extraction from videos may not be always needed.
4. No defined relation can be captured between space and time.

- **Approach:-**

1. Try to see till what extent the captured temporal information may be useful.
2. Adopt a simple late fusion based method between two different spatial streams.

Existing Drawbacks

Huge Computational Cost A simple convolution 2D net for classifying 101 classes has just ~5M parameters whereas the same architecture when inflated to a 3D structure results in ~33M parameters. It takes 3 to 4 days to train a 3DConvNet on UCF101 and about two months on Sports-1M, which makes extensive architecture search difficult and overfitting likely.

Capturing long context models which predicted from a single frame were trained on the huge and diverse ImageNet dataset performed reasonably well by identifying the actions such as bending, falling etc. However for some complex/extended actions such as walking vs running or bending vs falling, more local temporal information became necessary since such differentiations are virtually impossible to make from a single frame.

Lack of suitable Datasets The most popular and benchmark datasets have been UCF101 and HMDB-51 for quite some time. An architecture search on Sports1M can be extremely expensive. For UCF101, although the number of frames is comparable to ImageNet, the high spatial correlation among the videos makes the actual diversity in the training much lesser. Also, given the similar theme (sports) across both the datasets, generalization of benchmarked architectures to other tasks remained a problem. Other datasets such as THUMOS have limited classes and videos leading to overfitting. This has been solved to some extent lately with the introduction of the Kinetics dataset.

Development Timeline

Highlights of the timeline for the development of Two-stream related architectures:-

1. **Karpathy et al.** in 2014 first introduced 3D ConvNets as a network which could extract Spatiotemporal features directly from videos.

2. **Simoyan et al and Zissermann et al.** first introduced two stream networks in 2014.

3. Researchers at **Facebook** and **Dartmouth** released **c3d** in 2015.

4. **Wang et al** proposes some good practices for training very deep two-stream networks in 2015.

5. **Feichtenhofer et al.** introduces several new fusion techniques in 2016.

6. Researchers at **Facebook** and **Columbia** conducted an Neural Architecture Search (NAS) in 2017.

7. **Hara et al** talks about some augmentation techniques in his paper in 2017.

8. **Carriera et al.** introduces I3d and the Kinetics Dataset in 2018.

Diagrams of some backbone architectures

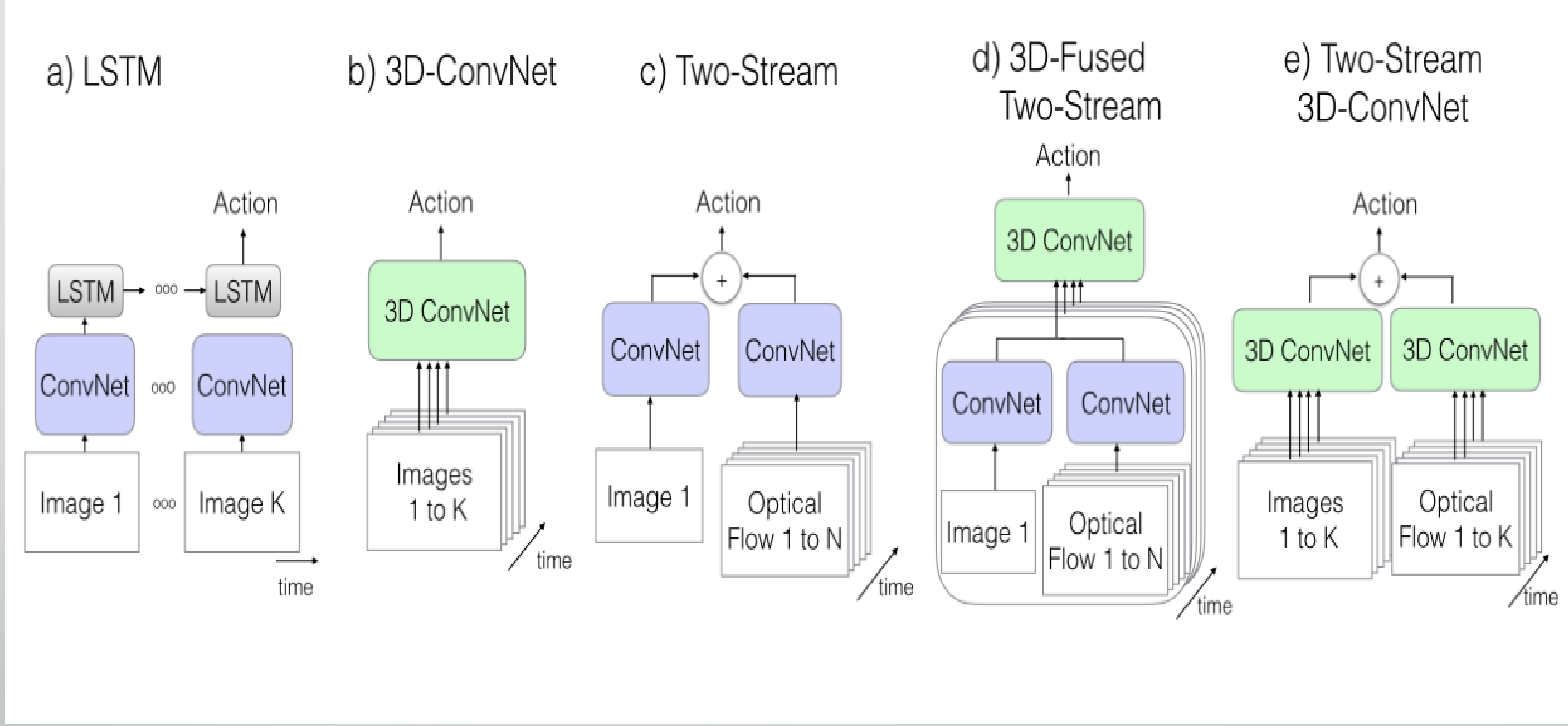

Fig 1. Evolution of various Architectures

Proposed Model Architecture

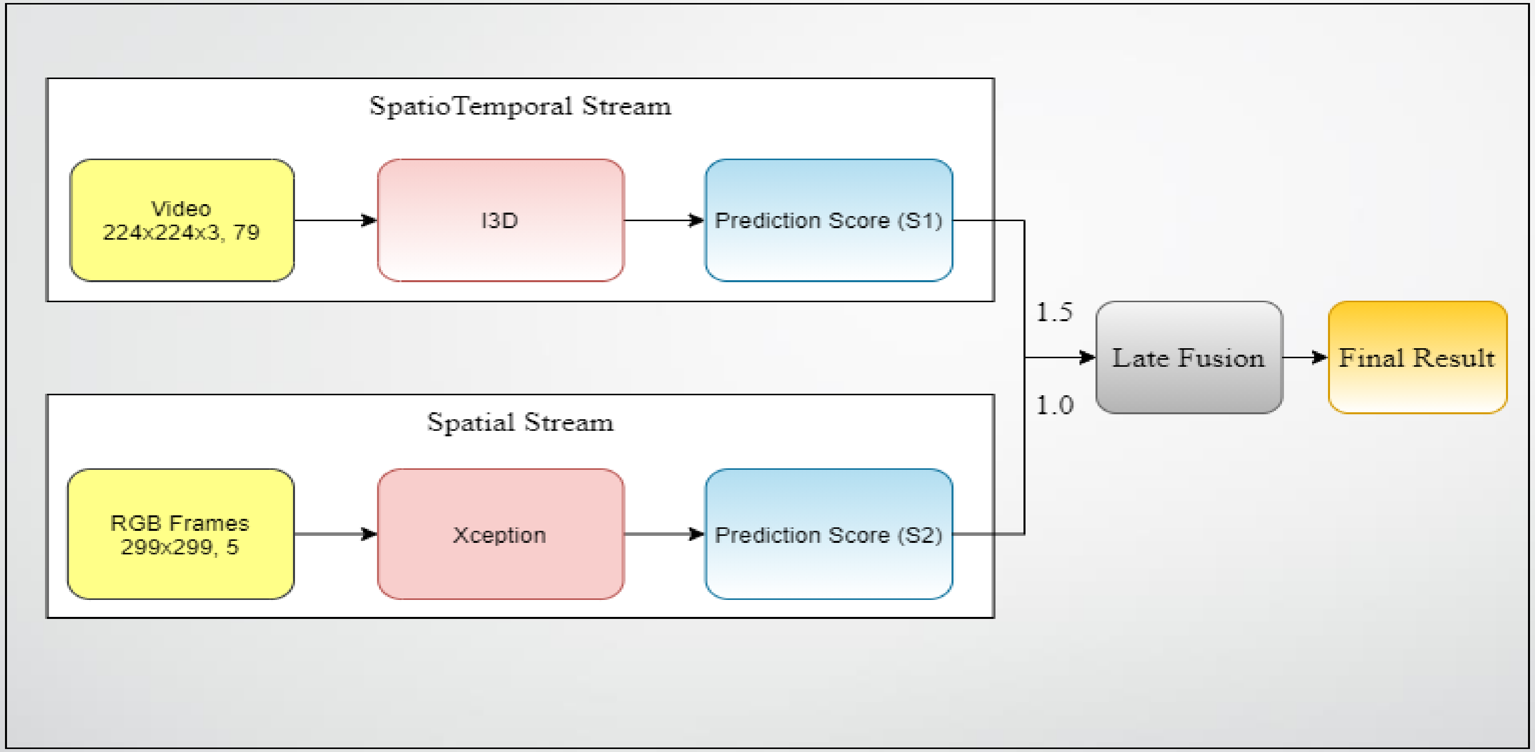

Fig 2. Model Architecture

The architecture primarily consists of the softmax score fusion between I3D and Xception. I3D is a 3D CNN model pre-trained on the Kinetics-600 dataset and Xception is another 2D CNN based model pre-trained on the ImageNet dataset. Final score is calculated based on the averaged late fusion between the individual streams.

Results

SL No.	True Label	Predicted Label	Prediction %	Top 1/Top-5
1.	v_Bowling_g22_c04	Bowling	99.9 (Highest)	Yes/yes
2.	V _ CricketBowling_g02_ c01	Playing squash or racketball	71.2	No/yes
3.	v_BabyCrawling_g18_ C06	Crawling baby	98.1	Yes/yes
4.	v_HammerThrow_g23_ C05	Hammer throw	99.5	Yes/yes
5.	v_BrushingTeeth_g17_ C02	Brushing Teeth	97.6	Yes/yes

Table 1

SL No.	Model Name	Parameters	Top-1 Accuracy (RGB)
1.	LSTM	9 Million	68.2%*
2.	3D CNN	79 Million	65.4%
3.	Two Stream	12 Million	86.9%*^
4.	C3D	73 Million	82.3%
5.	Res3D	33 Million	85.8%
6.	T3D	25 Million**	71.4%
7.	I3D	25 Million	88.8%*^^
8.	Ours	31 Million	87.5%

Table 2

Obtained results based on our fusion model. Table 1 shows the prediction accuracies obtained on a few sample videos from UCF-101. Table 2 shows the parameters and Top-1 accuracies of our model compared to some other architectures.

Advantages and Applications

Edge deployment friendly Our model requires only 6 GB of secondary storage and is therefore not bulky. The advantage of using such a model which uses less secondary storage is that it can be readily deployed into various real-time systems and edge-devices where resources are constrained and storage space on device is very limited. The use of transfer learning also means that productionizing the model is easy and maintainable.

Numerous Applications Video understanding is probably the biggest application of SpatioTemporal fusion. Human action recognition, scene understanding, real-time detection and several other applications exist. In other areas such as CT scan diagnosis and Medical Imaging, it is useful to observe changes in patterns in the infected area over a certain period of time (abnormality detection) or Surgical workflow modelling and monitoring. Other areas include robotics (autonomous driving, 3D mapping in drones) and manufacturing (Quality control).

Future Direction

- Our next targets are broadly summarized below:-

1. Designing a system which can be deployed in real-time. This can be done with the reduction in parameters and Inference time.
2. Improvement of the model for better video understanding. This can be done keeping in mind multiple approaches: Localization instead of brute-force classification, improving the temporal component using other architectures (self-attention, transformers etc).
3. If possible, find a specific use case (domain adaptation) in which this model can be applied. For Ex: HAR in an indoor environment, security applications etc, Tracking and detection(body, objects etc.)

References

- https://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review
- https://openaccess.thecvf.com/content_cvpr_2017/papers/Carreira_Quo_Vadis_Action_CVPR_2017_paper.pdf
- https://paperswithcode.com/task/3d-human-action-recognition
- https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf

Thank You!