Human Activity Understanding in Videos Almost From Scratch

Yao Zhou

SenseTime Group Limited

zhouyao@sensetime.com

February 8, 2018

What is the action in video

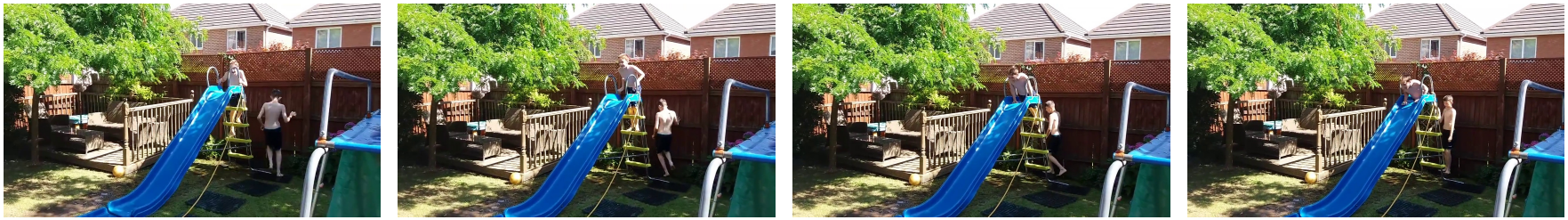


Figure : Frames are sampled from a video in the Kinetics dataset.

A boy is slipping down the water slide.
action: **Water Sliding**

Overview

1. Action Recognition in Videos
   - Untrimmed Videos (long videos, 100s+)
   - Trimmed Videos (short, $\leq$ 10s)
2. Temporal Action Proposal and Detection
   - Generating the Proposals
   - Classifying the Actions
3. Spatio-temporal Action Localization

Methods and results on recognizing untrimmed videos

Difficulties in recognizing untrimmed videos

- Untrimmed videos have long duration, means that recognition needs lots of computational cost.
- Model should be designed to capture the long-range temporal structure in untrimmed videos.
- Not only recognizing the action in long videos, but also object, scene.

Methods and results on recognizing untrimmed videos

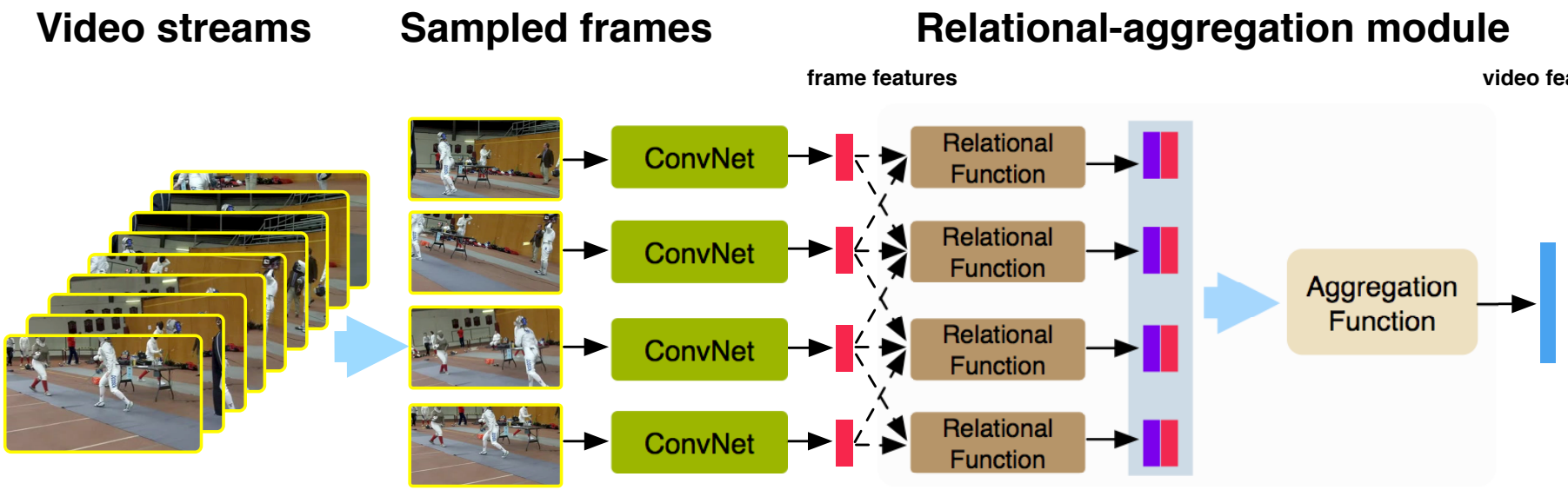
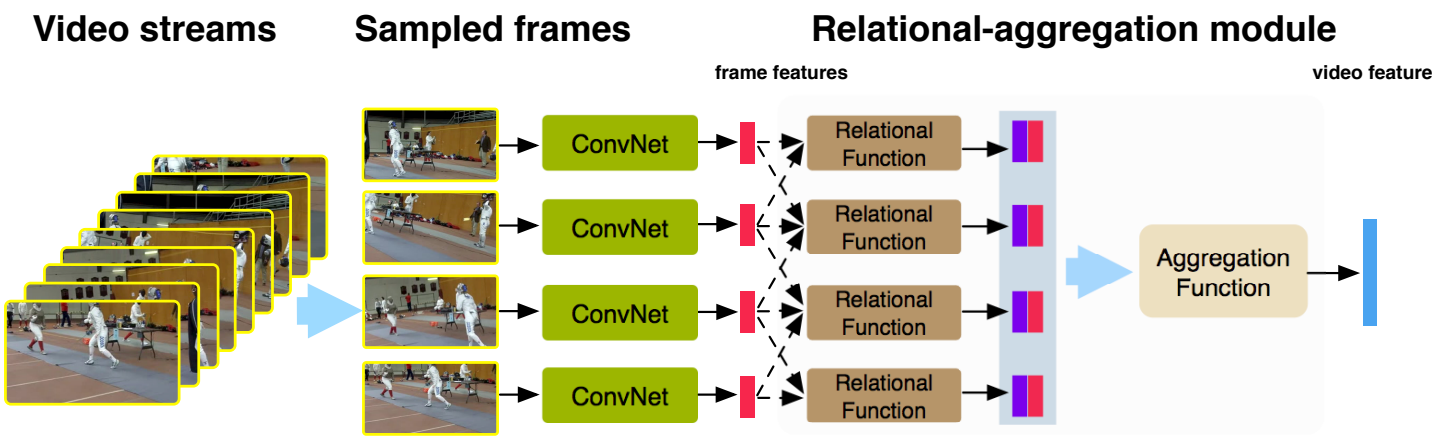


Figure : A graphical overview of the proposed model, named **temporal relation feature encoding networks**.

Methods and results on recognizing untrimmed videos

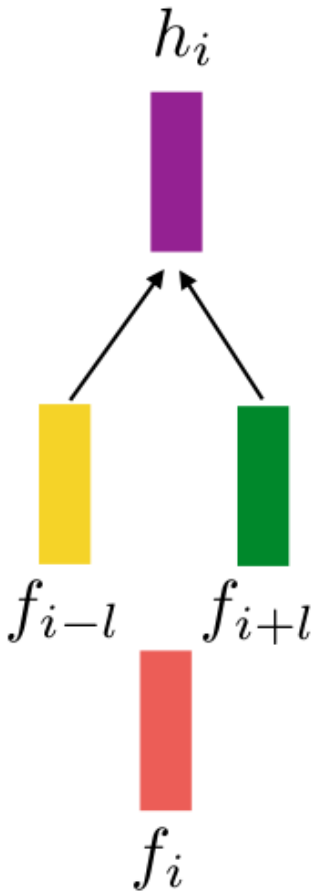


Model details

- Sparsely sampling frames using temporal segment-based sampling. (TSN, ECCV 2016)
- Extracting frame features by backbone ConvNets (Inception, ResNet).
- **Encoding the temporal relation and contextual information via a various of relation functions.**
- Aggregating the frame features by pooling, LSTM, CNN etc.

Temporal relation encoding

Capture the relationship of contextual frames



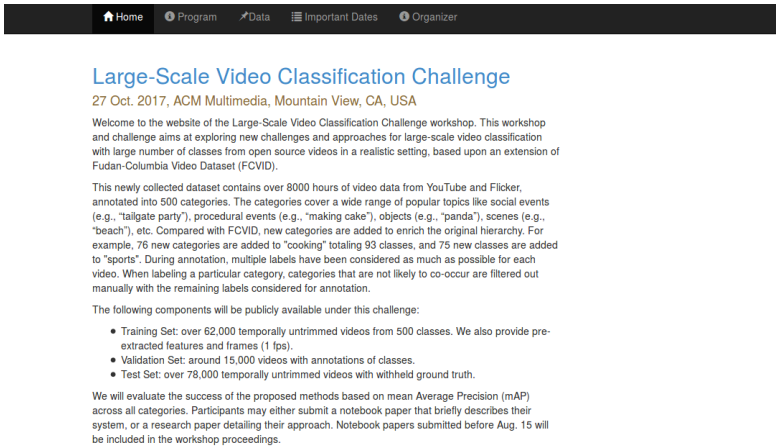
$$\textbf{Projection}: h_i^{proj} = CompFunc(f_{i-l}, f_{i+l}) = ReLU(W[f_{i-l}, f_{i+l}] + b)$$

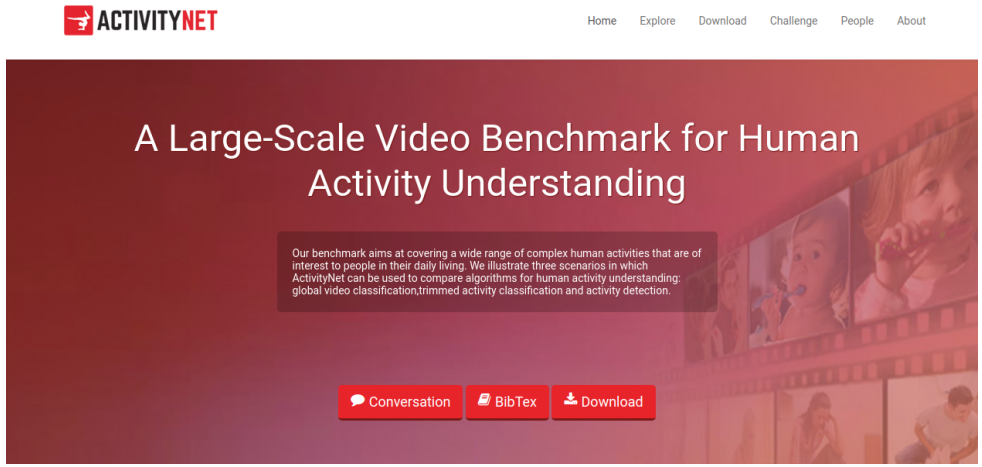
$$\textbf{Distance}: h_i^{dist} = CompFunc(f_{i-l}, f_{i+l}) = \frac{euc(f_{i-l}, f_{i+l})}{2} W[f_{i-l}, f_{i+l}] + \frac{\cos(f_{i-l}, f_{i+l})}{2} W[f_{i-l}, f_{i+l}]$$

$$\textbf{Subtraction}: h_i^{sub} = CompFunc(f_{i-l}, f_{i+l}) = f_{i+l} - f_{i-l}$$

$$\textbf{Multiplication}: h_i^{mul} = CompFunc(f_{i-l}, f_{i+l}) = f_{i-l} \odot f_{i+l}$$

Methods and results on recognizing untrimmed videos





Experiments on two benchmarks

- **FCVID** dataset at the Large Scale Video Classification Challenge(LSVC) of ACM MM 2017.
- **ActivityNet v1.3** at ActivityNet Large Scale Activity Recognition Challenge of CVPR 2017.

Experiment results on LSVC 2017 at MM

K	ConvNet	RelFunc	AggFunc	mAP
5	Inception-v3	w/o	avg pool	0.722
7	Inception-v3	w/o	avg pool	0.724
7	Inception-v3	projection	avg pool	0.748
7	Inception-v3	distance	avg pool	0.743
7	Inception-v3	submul	avg pool	**0.772**
7	Inception-v3	submul	LSTM	0.741
7	Inception-v3	submul	CNN	0.750

Table : Experiment results using Inception-v3 backbone with different settings.

Experiment results on LSVC 2017 at MM

K	ConvNet	RelFunc	AggFunc	mAP
5	Inception-ResNet-v2	w/o	avg pool	0.761
7	Inception-ResNet-v2	w/o	avg pool	0.766
7	Inception-ResNet-v2	projection	avg pool	0.792
7	Inception-ResNet-v2	distance	avg pool	0.795
7	Inception-ResNet-v2	submul	avg pool	**0.814**
7	Inception-ResNet-v2	submul	LSTM	0.801
7	Inception-ResNet-v2	submul	CNN	0.785

Table : Experiment results using Inception-ResNet-v2 backbone.

Experiment results on LSVC 2017 at MM

Results: 3rd place at LSVC of ACMMM 2017.

Methods and benchmarks for trimmed videos

Key to recognizing the trimmed videos.

- Trimmed videos have short durations, recent popular benchmarks are annotated with 10s or 3s.
- Both appearance and motion are important for recognizing action in trimmed videos, result in multi-modalities (e.g RGB, Flow etc)
- Frame feature can be extracted by 2D ConvNets and short video volume are natural for 3D convolution operation, leading to multi-models.(C2D, C3D models)

Methods and benchmarks for trimmed videos

Modalities: Two/Multi-streams

- Appearance: Sparsely or densely sampled **RGB** frames.
- Motion: Stacked **optical flows** by TVL1 algorithm(OpenCV).
- Audio, RGB difference, Pose, Object detection etc.

Backbones: 2D ConvNets vs. 3D ConvNets

- Popular ConvNets: VGG, Inception, ResNet, DenseNet etc.
- Inflated 3D ConvNets: **I3D**(CVPR17), **R3D**(CVPR17), P3D(CVPR17), S3D(arXiv), Non-local(arXiv) etc.

Fusion: Combine the modalities and models

- Fuse multi-modalities by (weighted) averaging.
- Mixed 2D and 3D convolution operation in one model.

Methods and benchmarks for trimmed videos

Modalities: Two/Multi-streams

- Appearance: Sparsely or densely sampled **RGB** frames.
- Motion: Stacked **optical flows** by TVL1 algorithm(OpenCV).
- Audio, RGB difference, Pose, Object detection etc.

Modalities: Two-stream networks(NIPS14, VGG Oxford)

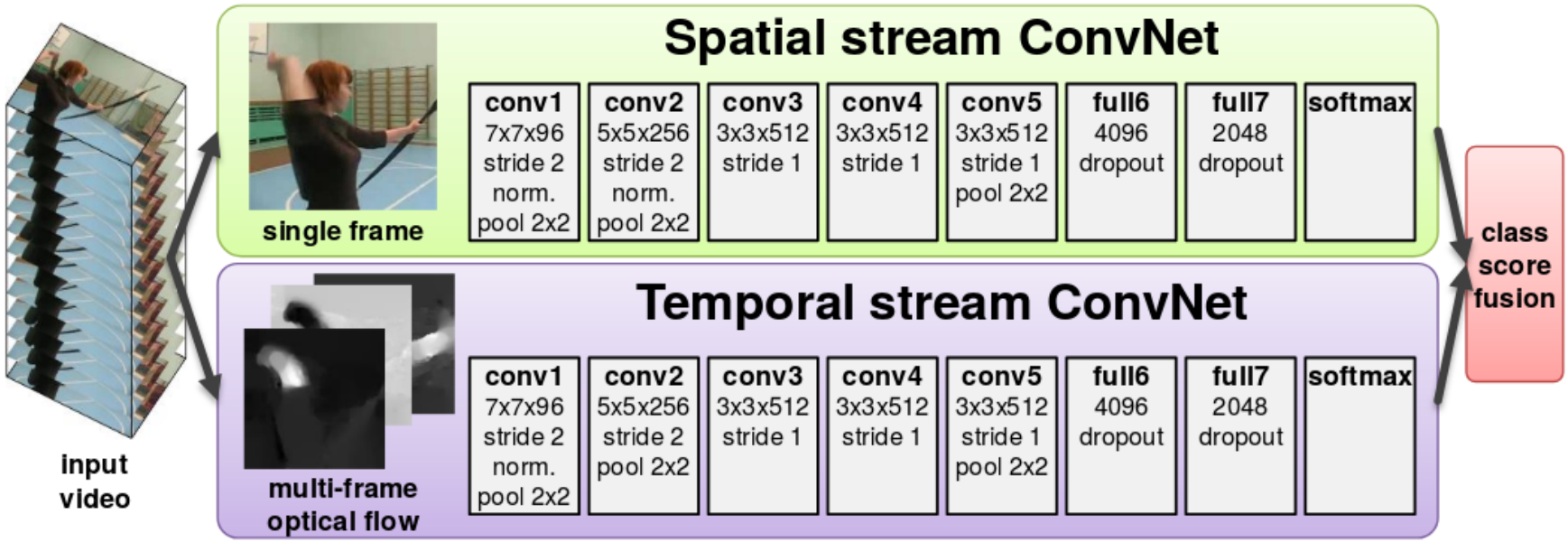
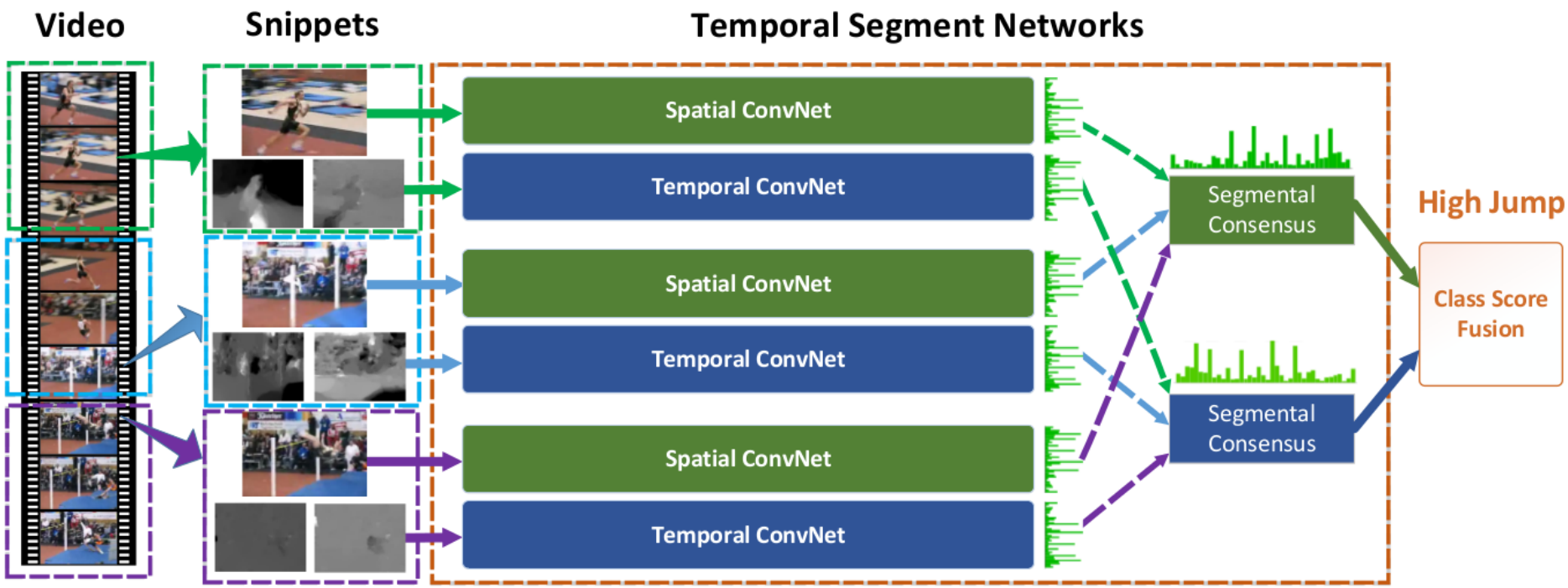


Figure 1: **Two-stream architecture for video classification.**

Modalities: Temporal segment networks(ECCV16, CUHK)



Methods and benchmarks for trimmed videos

Backbones: 2D ConvNets v.s. 3D ConvNets

- Popular ConvNets: VGG, Inception, ResNet, DenseNet etc.
- Inflated 3D ConvNets: **I3D**(CVPR17), **R3D**(CVPR17), P3D(CVPR17), S3D(arXiv), Non-local(arXiv) etc.

Backbones: Inception-v1 inflated 3D (CVPR17, DeepMind)

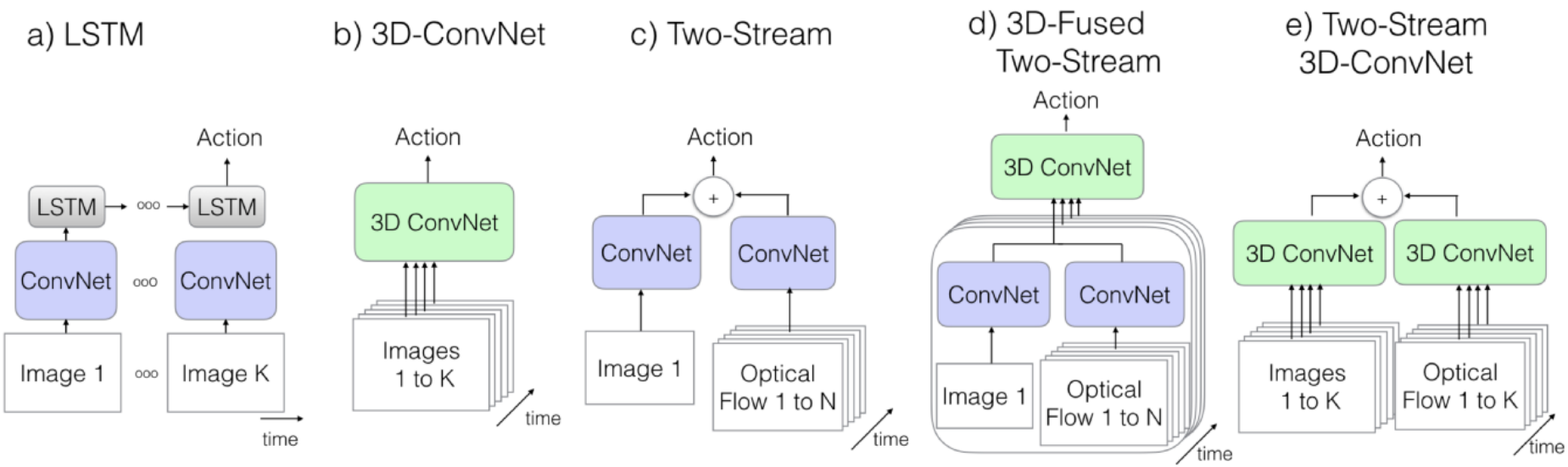
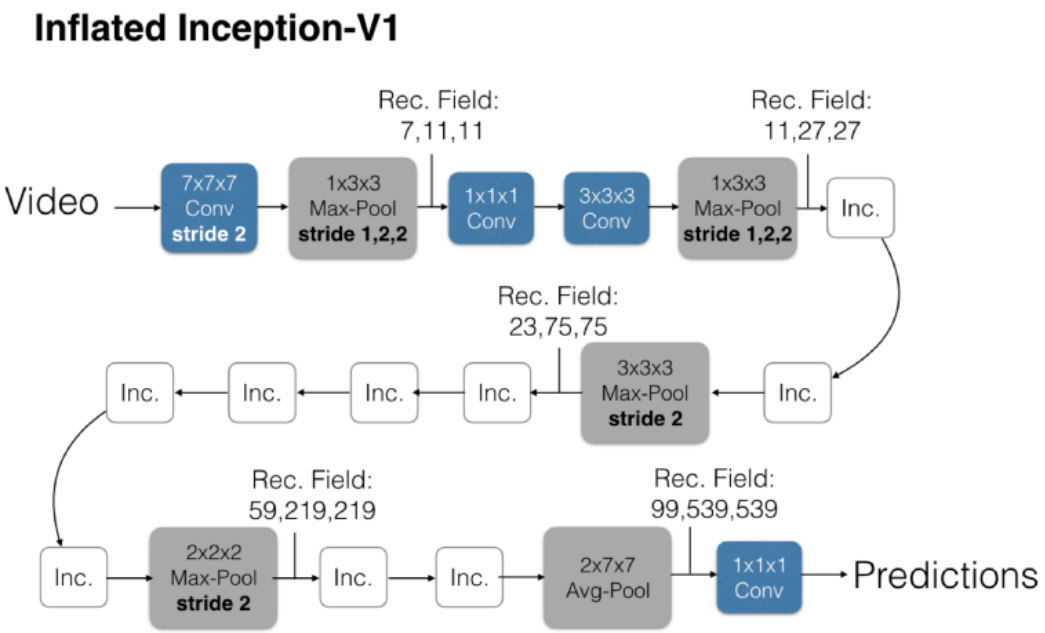


Figure : Architecture of 2D and 3D models.

Backbones: Inception-v1 inflated 3D (CVPR17, DeepMind)



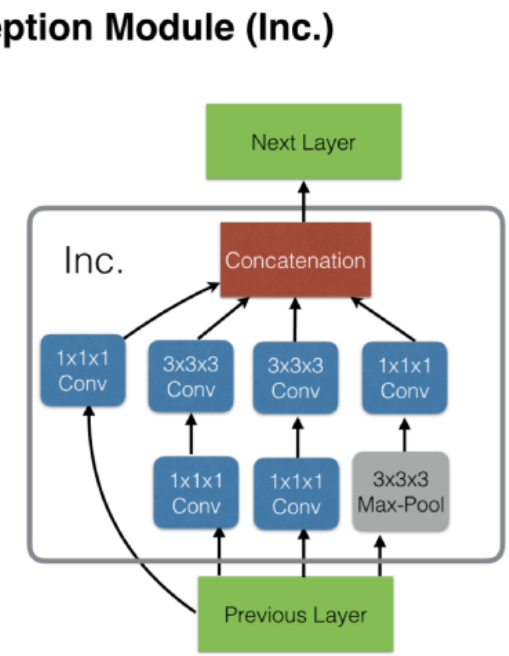


Figure : **I3D**: Inflating kxk kernel to $k_t xkxk$ kernel

Backbones: Inception-v1 inflated 3D (CVPR17, DeepMind)

Comparison between 2D and 3D inception block.

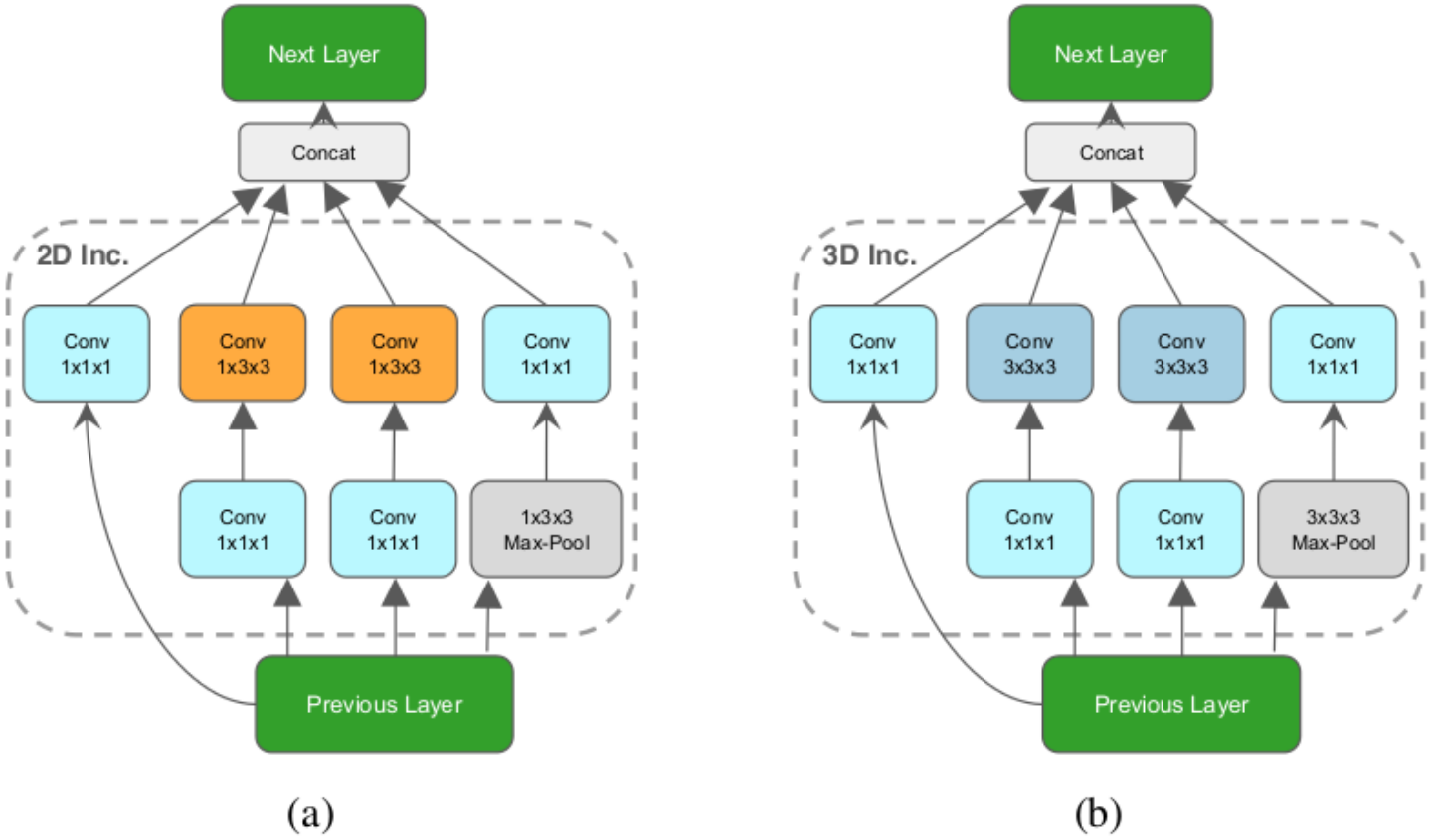
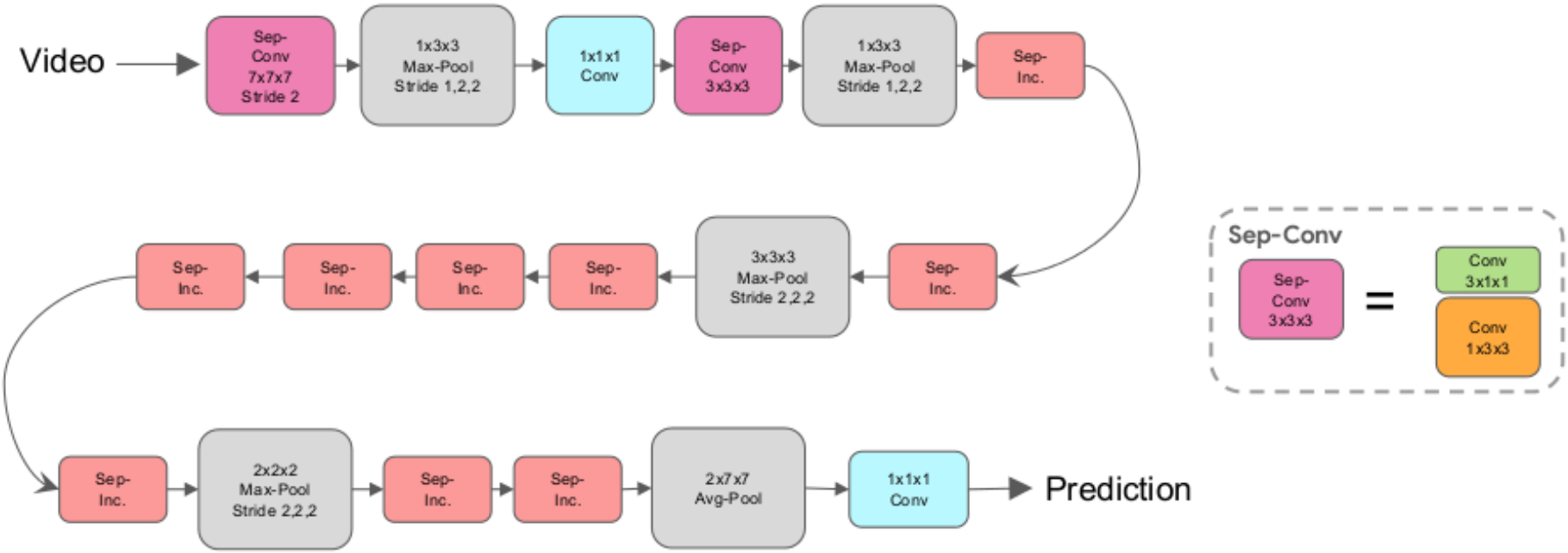


Figure 2. (a) 2D Inception block; (b) 3D Inception block.

Backbones: Separable 3D (arXiv1712, UCSD and Google)

Parameters increases rapidly when network going deeper.

$\Rightarrow$Reducing parameters by splitting k_txkxk to 1xkxk and k_tx1x1.



Backbones: Separable 3D (arXiv1712, UCSD and Google)

Parameters increases rapidly when network going deeper. $\Rightarrow$Reducing parameters by replacing k_txkxk to 1xkxk and k_tx1x1.

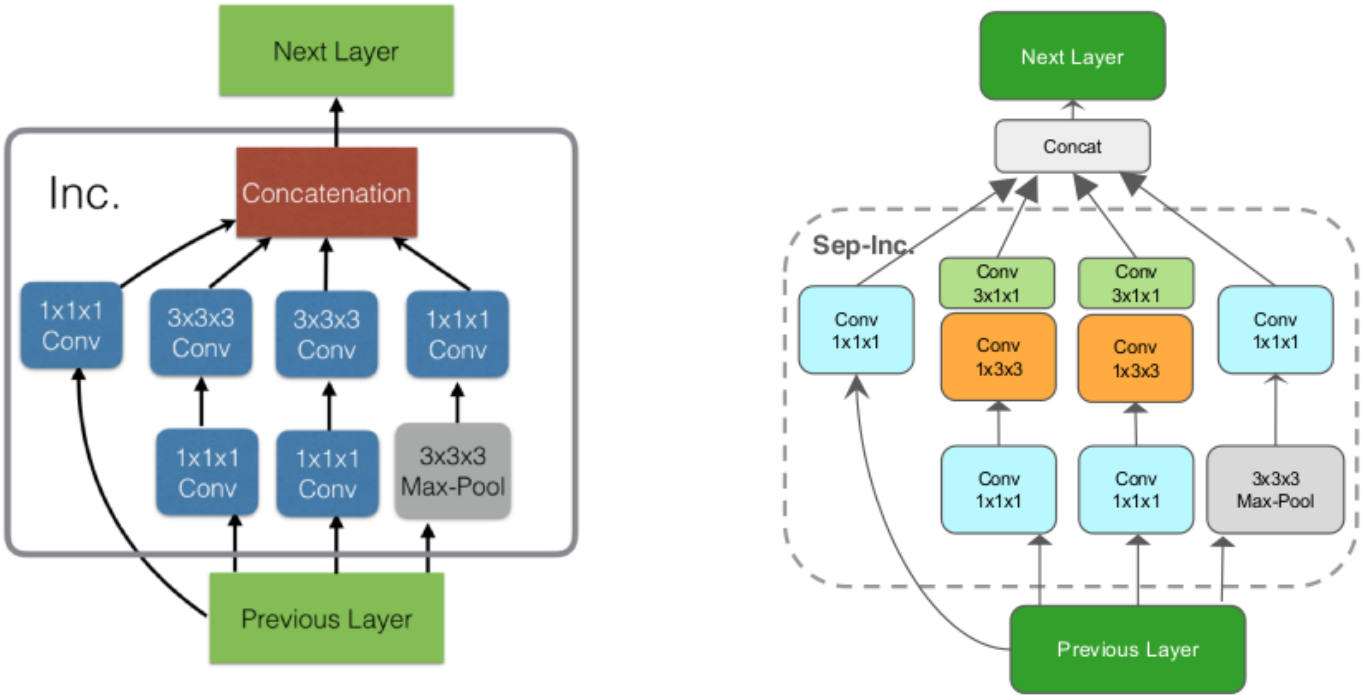


Figure : Comparison of I3D inception block and S3D inception block.

Benchmarks for recognizing trimmed videos

Dataset	Category	Duration	Classes	Examples	Organizer
UCF101	human action	10s	101	13,320	UCF
Kinetics	human action	10s	400	300,000	DeepMind
Moments in Time	action or activity	3s	339	1 million	MIT and IBM
SoA	scene object action	-	-	-	Facebook

Table : We focused on four benchmarks, three will be used at ActivityNet 2018!

Results for trimmed videos

Models	RGB		Flow		RGB+Flow(1:1)	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Two-Stream with ResNet-50	56.0%	77.3%	49.5%	71.9%	61.0%	81.3%
TSN with Inception-v3	72.5%	90.2%	62.8%	82.4%	76.6%	92.4%
TSN with Inception-v3(Ours)	72.6%	90.2%	60.4%	82.0%	75.7%	91.9%
+ Similarity	72.6%	90.4%	-	-	-	-
+ Subtraction	72.9%	90.6%	-	-	-	-
+ Multiplication	73.0%	90.6%	-	-	-	-
+ Subtraction and Multiplication	73.2%	90.8%	-	-	-	-
Ensemble	73.5%	91.0%				

Table : Ablation study of different modalities and the relation functions on Kinetics validation set. (2D models)

Results for trimmed videos

Models	RGB		Flow		RGB+Flow(1:1)	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Two-Stream I3D	71.4%	89.3%	62.5%	83.9%	74.0%	91.3%
Two-Stream R3D	69.8%	86.2%	60.7%	81.4%	73.0%	89.6%
Two-Stream S3D	-	-	-	-	-	-

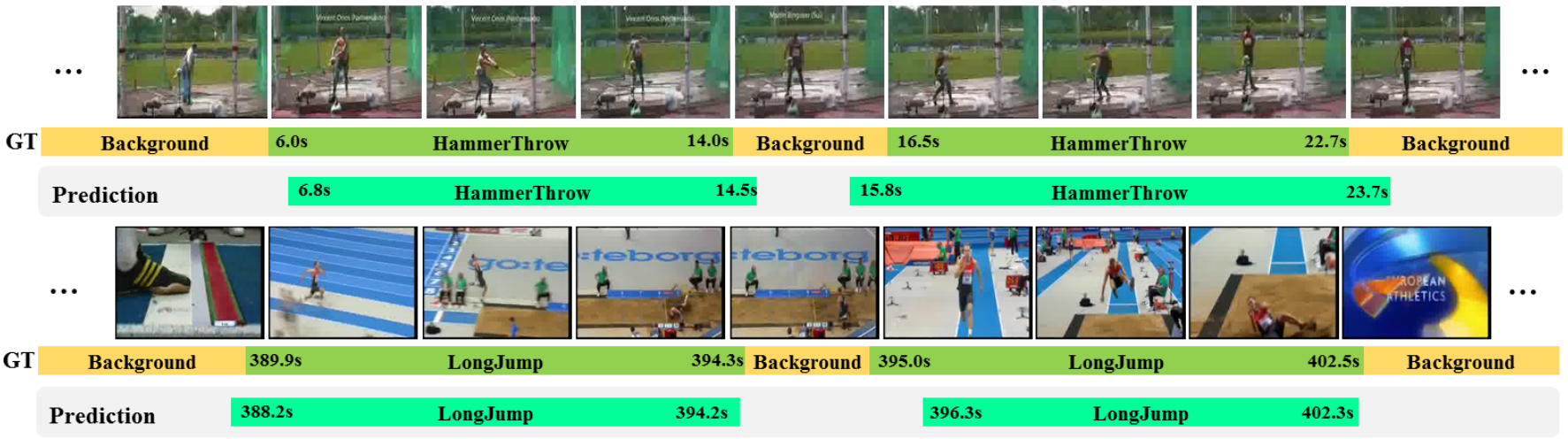
Table : Performance of different 3D models on Kinetics validation set.

Results for trimmed videos

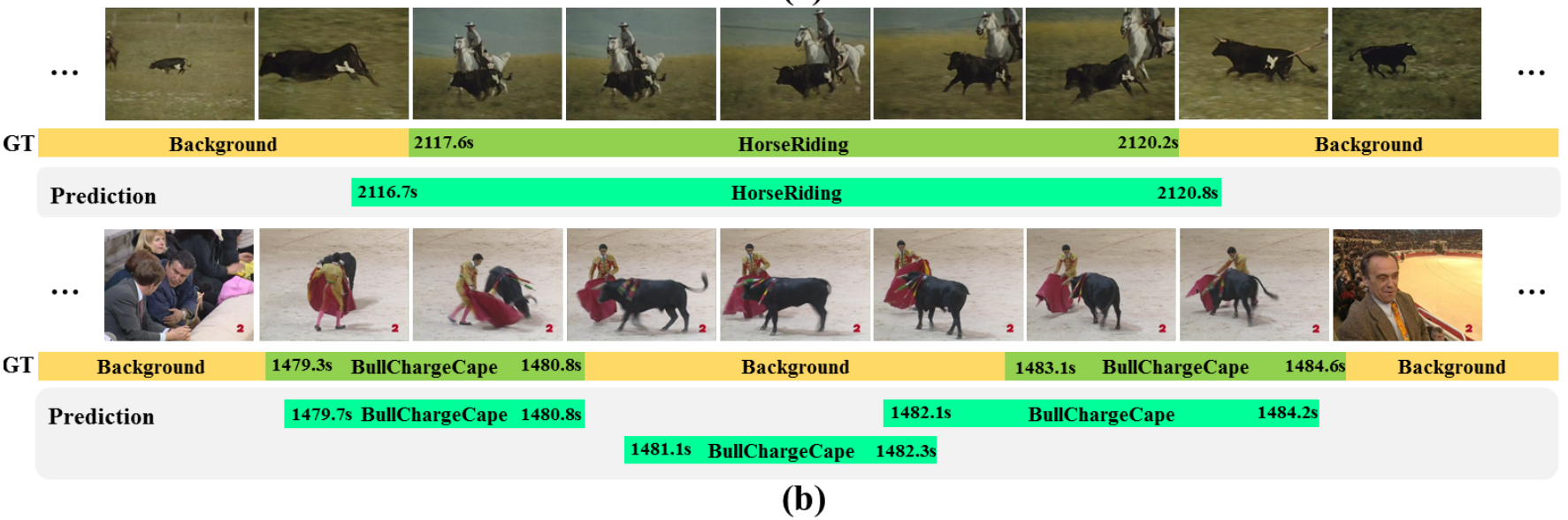
Models	RGB	Flow	RGB+Flow
TSN (Inception-v3)	93.2%	95.3%	97.3%
R3D (ResNet-50)	93.4%	94.2%	96.1%
I3D (Inception-v1)	**95.6%**	**96.7%**	**98.0%**

Table : Performance on UCF101 validation set. All models implemented by TensorFlow framework with Kinetics pre-training.

Temporal Action Proposal and Detection



(a)



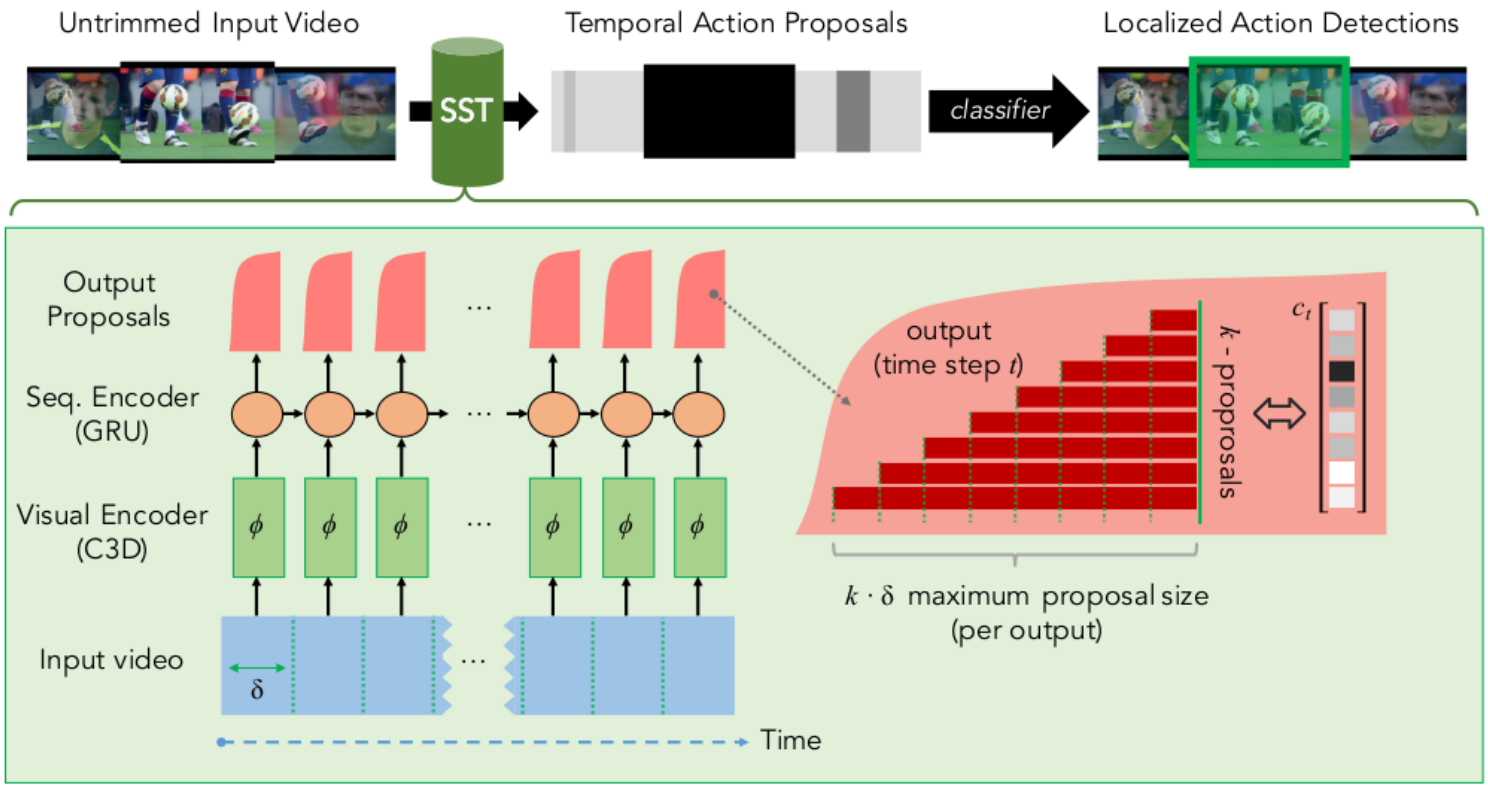
(b)

Temporal Action Proposal and Detection

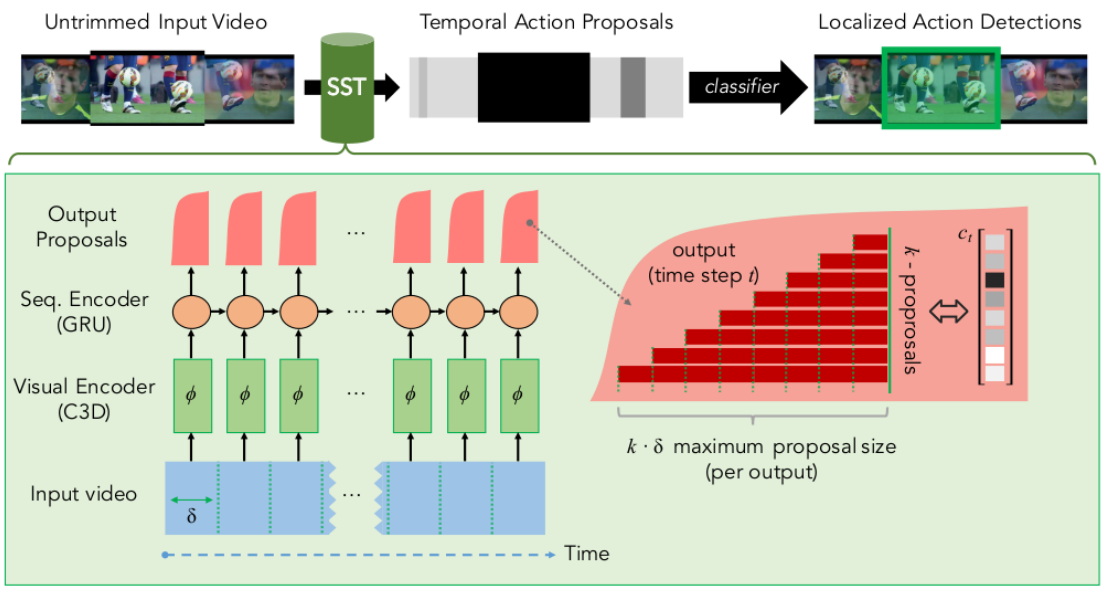
Problem definition:

- Given a long untrimmed video, generating segments (short video clips) which contains an action
- and then classify which class the segment belong to. Analogous to finding bounding boxes in a image which contains a object in object detection task.

SST: Single-Stream Temporal Action Proposals (CVPR17, Stanford)



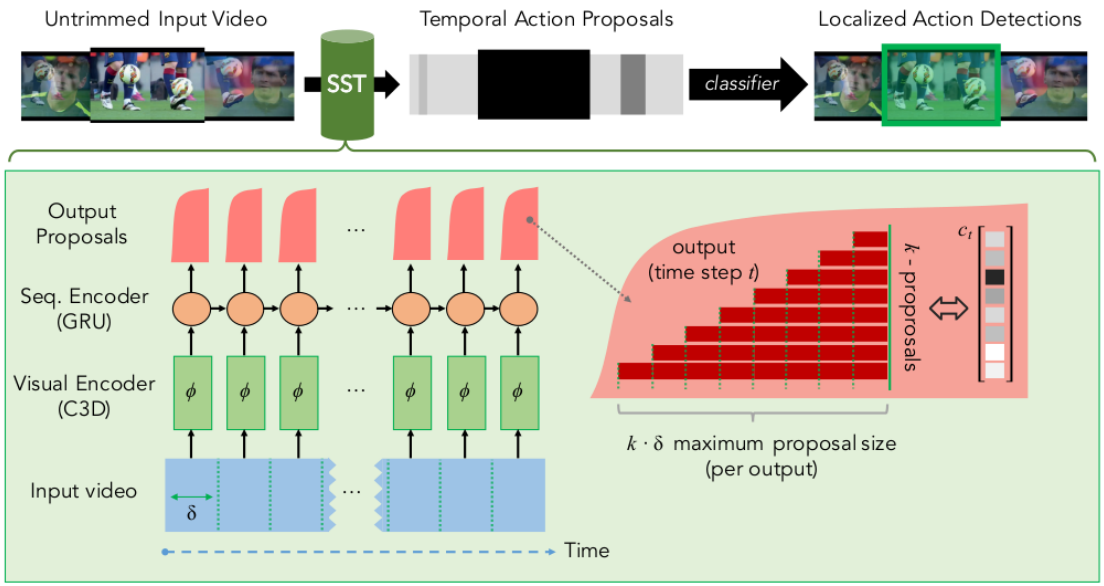
SST and improved version



Details

- Extracted visual feature by C3D with 16fps resolution.
- Encoding temporal information by a GRU.
- Generating K proposals at each time step.
- Setting positive if the tiou of proposal and ground truth more than 0.5, otherwise negative.

SST and improved version



Details

- Extracted visual feature by C3D with 16fps resolution.
- Encoding temporal information by a GRU.
- Generating K proposals at each time step.
- Setting positive if the tiou of proposal and ground truth more than 0.5, otherwise negative.

The End