CVPR LONG BEACH CALIFORNIA June 16-20, 2019

FBK-HUPBA Submission to the EPIC-Kitchens 2019 Action Recognition Challenge

Swathikiran Sudhakaran, Sergio Escalera, Oswald Lanz

sudhakaran@fbk.eu sergio@maia.ub.es lanz@fbk.eu

https://github.com/swathikirans/LSTA

FONDAZIONE BRUNO KESSLER

UNIVERSITAT DE BARCELONA

CVC Centre de Visió per Computador

HuPBA

Introduction

Introduction

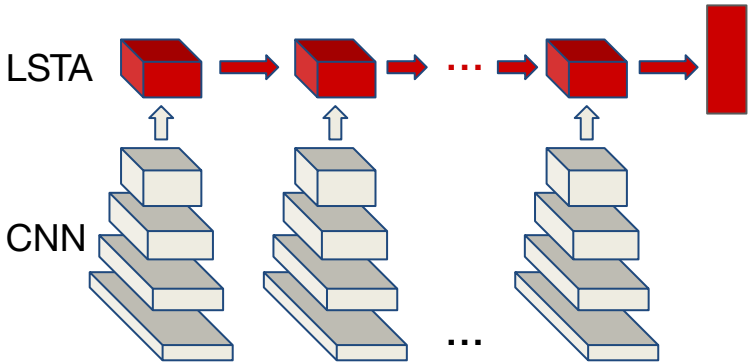


Fig. CNN-LSTA

S. Sudhakaran, S. Escalera, and O. Lanz. LSTA: Long Short-Term Attention for Egocentric Action Recognition. CVPR, 2019

Introduction



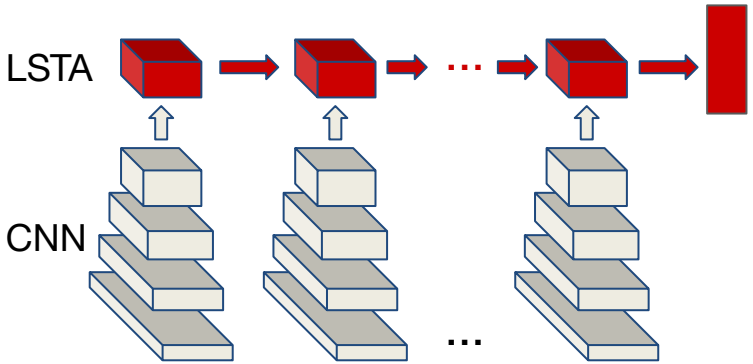


Fig. CNN-LSTA

S. Sudhakaran, S. Escalera, and O. Lanz. LSTA: Long Short-Term Attention for Egocentric Action Recognition. CVPR, 2019

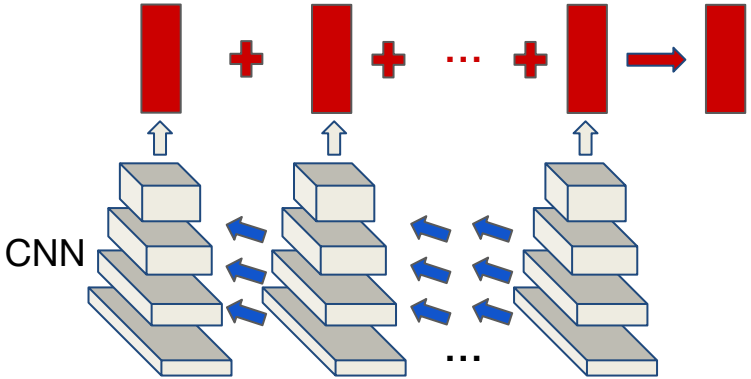


Fig. HF-TSN

S. Sudhakaran, S. Escalera, and O. Lanz. Hierarchical Feature Aggregation Networks for Video Action Recognition. arXiv:1905.12462

CNN-LSTA

Long Short-Term Attention (LSTA) enhances LSTM with:

- Built-in recurrent attention
- Novel output pooling

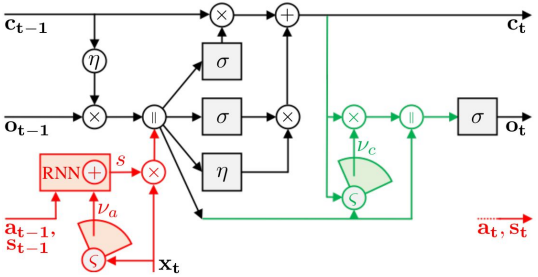
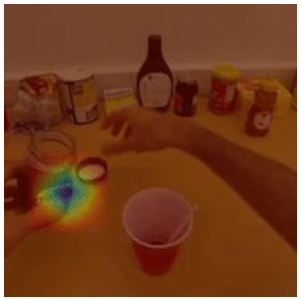


Fig. LSTA block



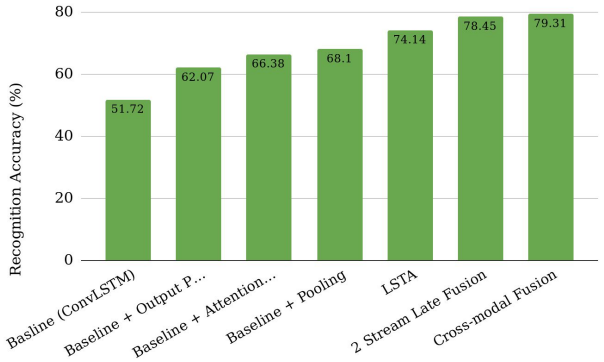
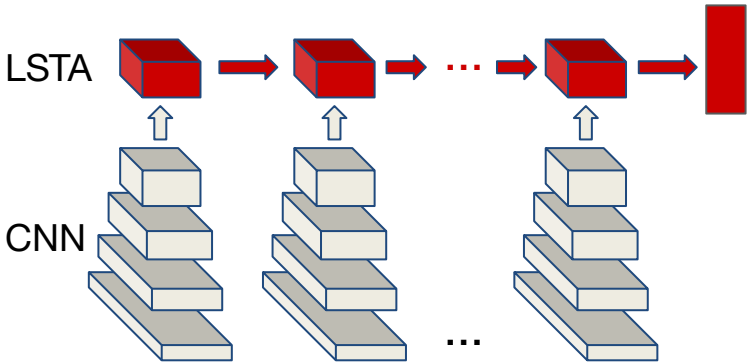


Fig. Ablation study on GTEA61 fixed split

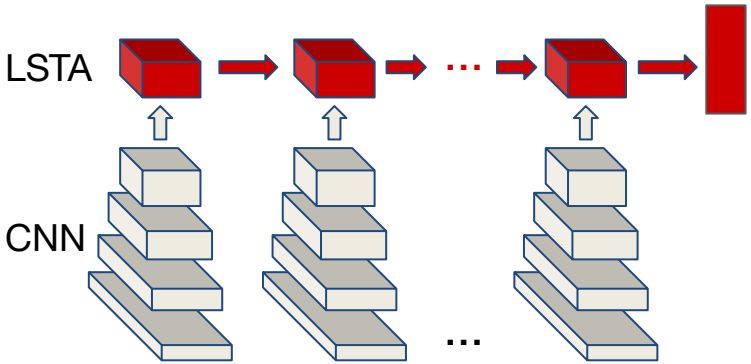
CNN-LSTA



CNN-LSTA

Variants:

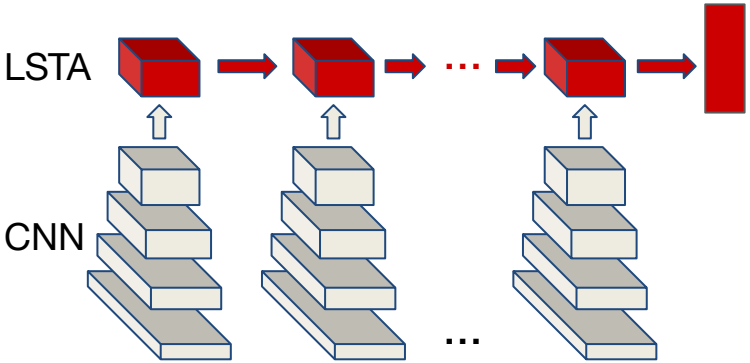
- Backbone: ResNet-34, ResNet-50, InceptionV3
- Pre-training: ImageNet, Kinetics

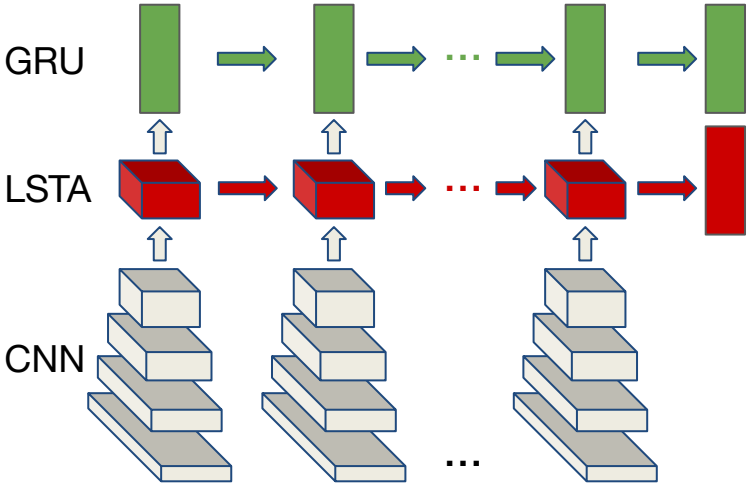


CNN-LSTA

Variants:

- Backbone: ResNet-34, ResNet-50, InceptionV3
- Pre-training: ImageNet, Kinetics
- Aggregation

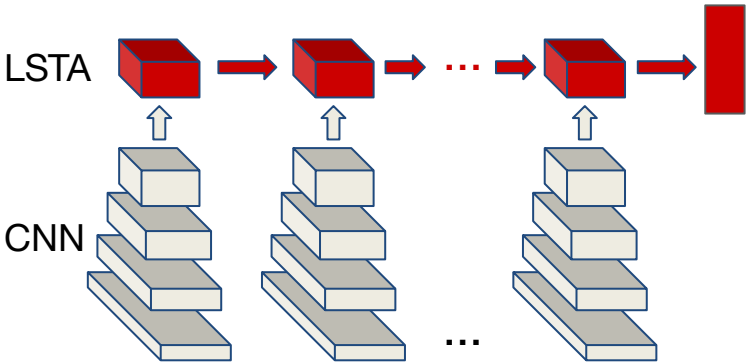


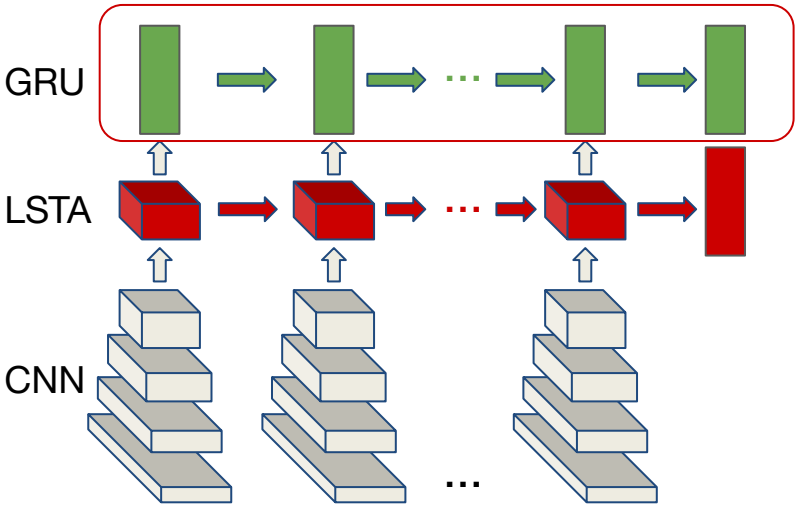


CNN-LSTA

Variants:

- Backbone: ResNet-34, ResNet-50, InceptionV3
- Pre-training: ImageNet, Kinetics
- Aggregation
- Training variations

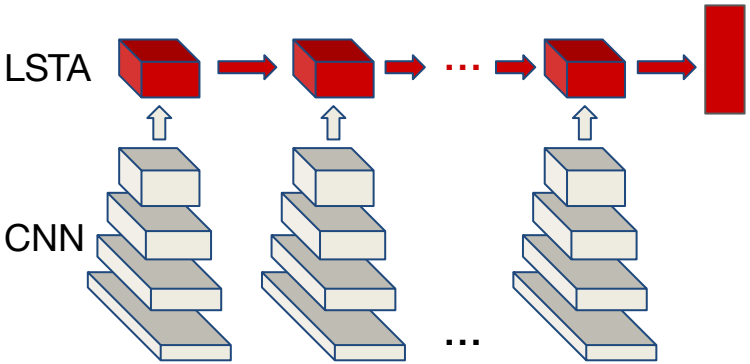


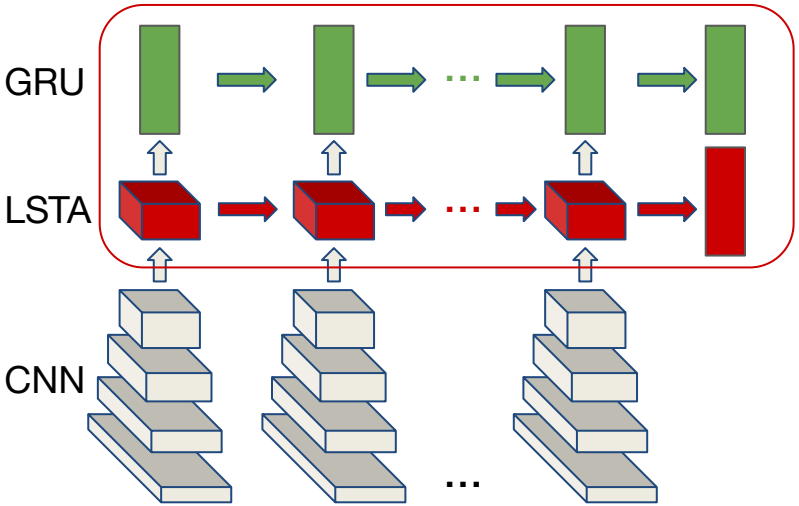


CNN-LSTA

Variants:

- Backbone: ResNet-34, ResNet-50, InceptionV3
- Pre-training: ImageNet, Kinetics
- Aggregation
- Training variations

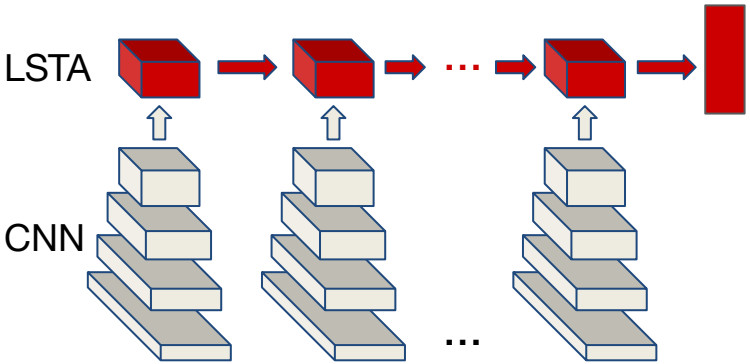


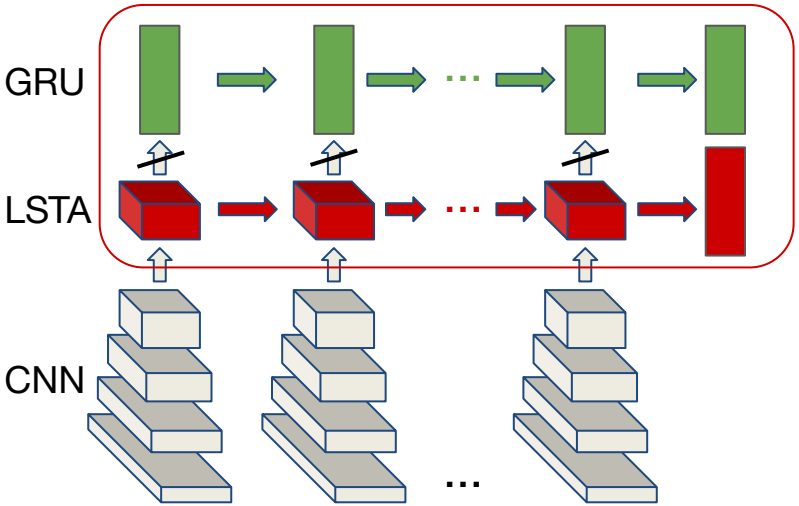


CNN-LSTA

Variants:

- Backbone: ResNet-34, ResNet-50, InceptionV3
- Pre-training: ImageNet, Kinetics
- Aggregation
- Training variations





CNN-LSTA

Two stream with Cross-modal fusion

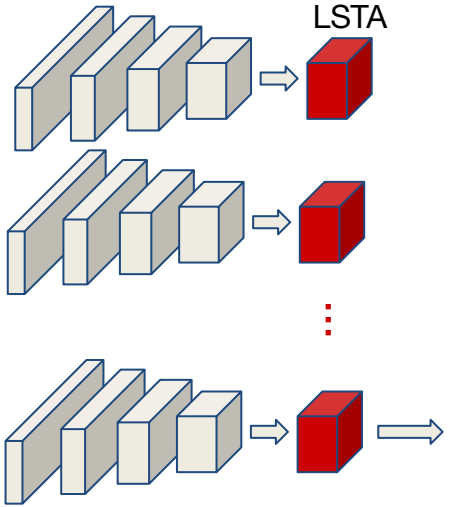
Features of one stream controls the bias of the LSTA gates of other stream

CNN-LSTA

Two stream with Cross-modal fusion

Features of one stream controls the bias of the LSTA gates of other stream

Train the appearance stream with RGB frames



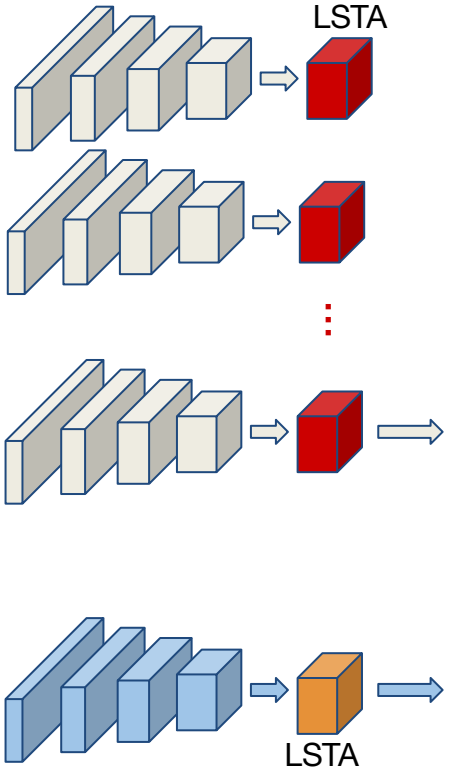
CNN-LSTA

Two stream with Cross-modal fusion

Features of one stream controls the bias of the LSTA gates of other stream

Train the appearance stream with RGB frames

Train the motion stream with stacked optical flow



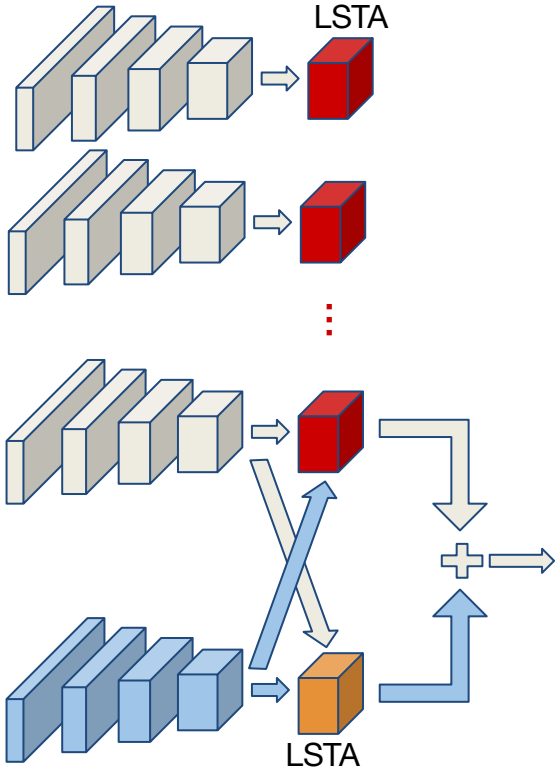
CNN-LSTA

Two stream with Cross-modal fusion

Features of one stream controls the bias of the LSTA gates of other stream

Train the appearance stream with RGB frames

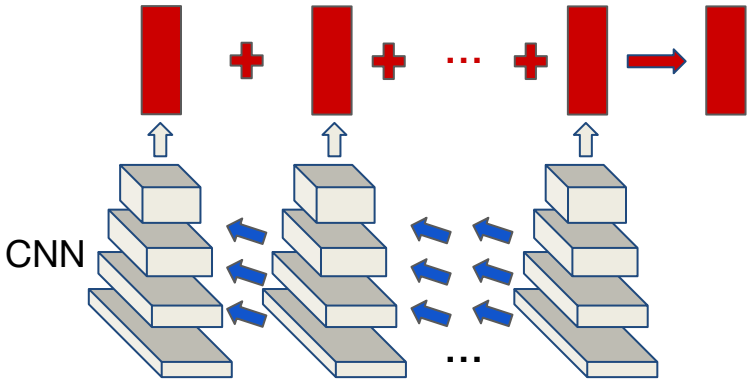
Train the motion stream with stacked optical flow



HF-TSN

Hierarchical feature aggregation

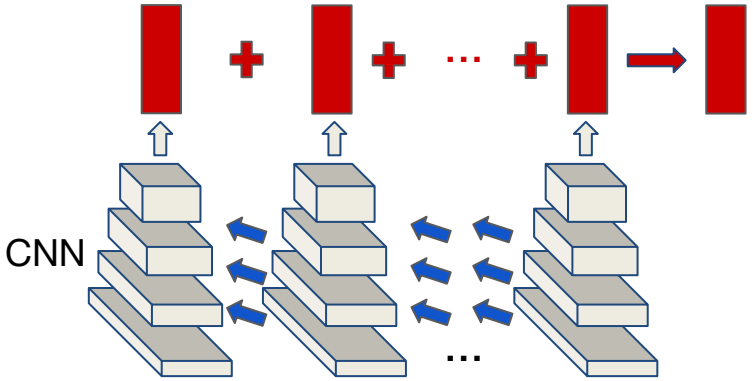
Plug and play module on any CNN backbone

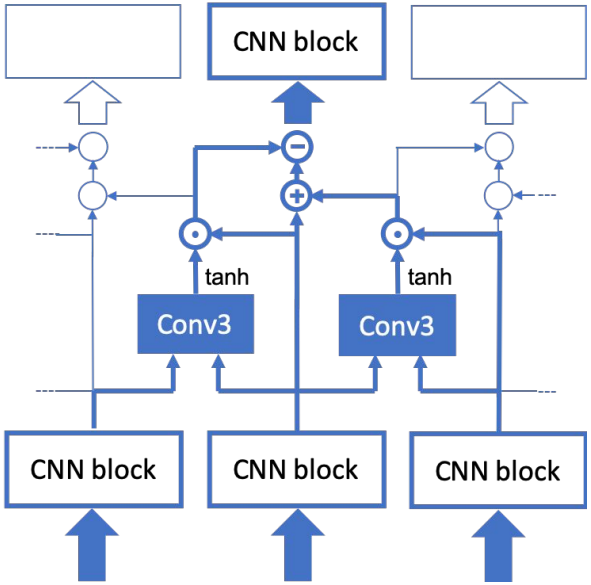


HF-TSN

Hierarchical feature aggregation

Plug and play module on any CNN backbone



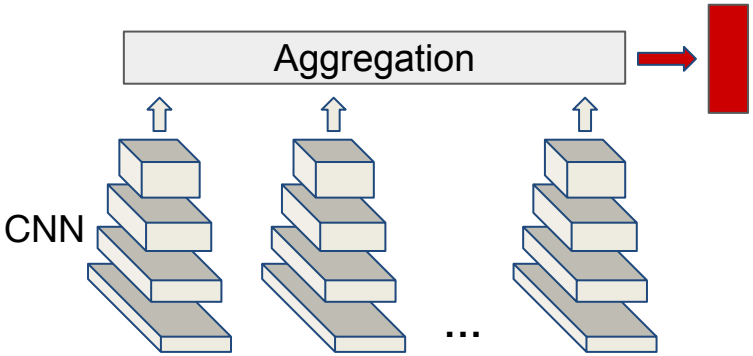


Huge performance gain over TSN baseline (17% Vs 41%) on 20BN-something-v1 dataset with an overhead of 1% increase in both parameters and FLOPs

Structured Prediction

Labels are verb-noun pairs

Generate action label from verb-noun pairs

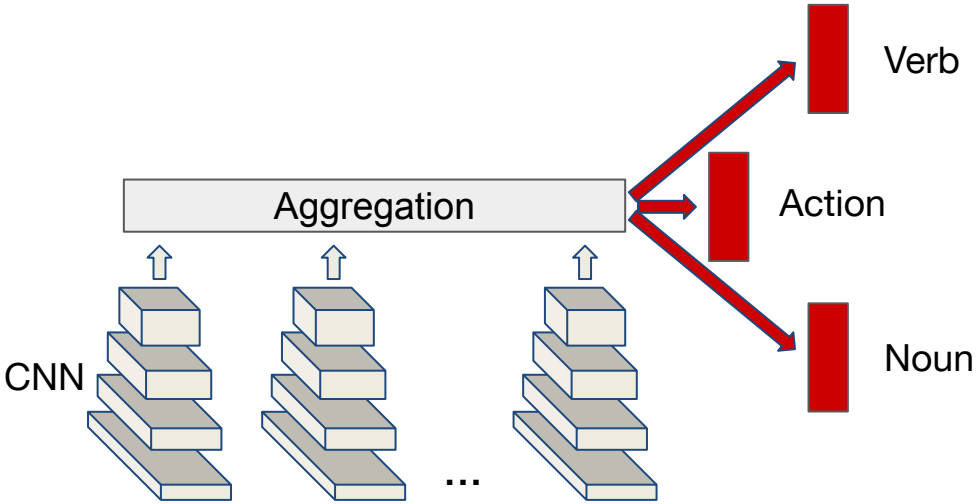


Structured Prediction

Labels are verb-noun pairs

Generate action label from verb-noun pairs

Multi-task learning



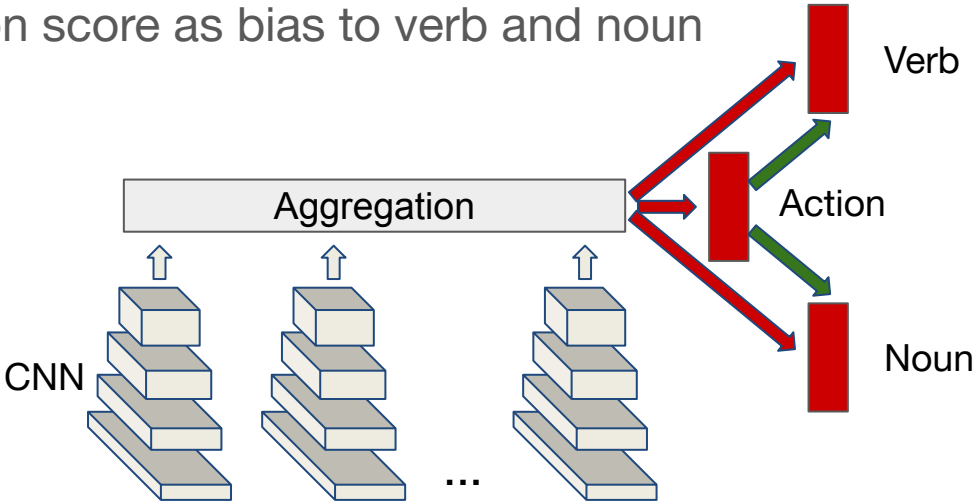
Structured Prediction

Labels are verb-noun pairs

Generate action label from verb-noun pairs

Multi-task learning

Apply action score as bias to verb and noun



Results

Method	Backbone	S1			S2		
		Verb	Noun	Action	Verb	Noun	Action
LSTA	Res-34	58.25	38.93	30.16	45.51	23.46	15.88
	Res-50	57.81	37.84	29.54	44.38	22.53	15.98
	Res-50†	57.69	39.36	29.79	43.53	22.98	16.25
	IncV3	57.28	39.32	29.35	44.66	23.76	17.31

†:Kinetics pre-trained

Results

Method	Backbone	S1			S2		
		Verb	Noun	Action	Verb	Noun	Action
LSTA	Res-34	58.25	38.93	30.16	45.51	23.46	15.88
	Res-50	57.81	37.84	29.54	44.38	22.53	15.98
	Res-50†	57.69	39.36	29.79	43.53	22.98	16.25
	IncV3	57.28	39.32	29.35	44.66	23.76	17.31
LSTA-GRU	Res-50*	57.30	37.59	29.17	43.94	22.16	15.94
	Res-34**	60.61	40.84	32.04	45.37	23.49	16.59
	Res-34***	61.31	40.93	32.14	44.90	22.60	16.25

†:Kinetics pre-trained; *: Fine-tuned layers: GRU; **: Fine-tuned layers: GRU+LSTA; ***: Fine-tuned layers: GRU+LSTA+Conv5_3

Results

Method	Backbone	S1			S2		
		Verb	Noun	Action	Verb	Noun	Action
LSTA	Res-34	58.25	38.93	30.16	45.51	23.46	15.88
	Res-50	57.81	37.84	29.54	44.38	22.53	15.98
	Res-50†	57.69	39.36	29.79	43.53	22.98	16.25
	IncV3	57.28	39.32	29.35	44.66	23.76	17.31
LSTA-GRU	Res-50*	57.30	37.59	29.17	43.94	22.16	15.94
	Res-34**	60.61	40.84	32.04	45.37	23.49	16.59
	Res-34***	61.31	40.93	32.14	44.90	22.60	16.25
LSTA-2S	Res-34	62.12	40.41	32.60	48.89	24.27	18.71

†:Kinetics pre-trained; *: Fine-tuned layers: GRU; **: Fine-tuned layers: GRU+LSTA; ***: Fine-tuned layers: GRU+LSTA+Conv5_3

Results

Method	Backbone	S1			S2		
		Verb	Noun	Action	Verb	Noun	Action
LSTA	Res-34	58.25	38.93	30.16	45.51	23.46	15.88
	Res-50	57.81	37.84	29.54	44.38	22.53	15.98
	Res-50†	57.69	39.36	29.79	43.53	22.98	16.25
	IncV3	57.28	39.32	29.35	44.66	23.76	17.31
LSTA-GRU	Res-50*	57.30	37.59	29.17	43.94	22.16	15.94
	Res-34**	60.61	40.84	32.04	45.37	23.49	16.59
	Res-34***	61.31	40.93	32.14	44.90	22.60	16.25
LSTA-2S	Res-34	62.12	40.41	32.60	48.89	24.27	18.71
HF-TSN	BNInc	57.57	39.90	28.09	42.40	25.23	16.93
	Res-50	56.69	40.70	29.38	45.48	24.55	17.38

†:Kinetics pre-trained; *: Fine-tuned layers: GRU; **: Fine-tuned layers: GRU+LSTA; ***: Fine-tuned layers: GRU+LSTA+Conv5_3

Results

Method	Backbone	S1			S2		
		Verb	Noun	Action	Verb	Noun	Action
LSTA	Res-34	58.25	38.93	30.16	45.51	23.46	15.88
	Res-50	57.81	37.84	29.54	44.38	22.53	15.98
	Res-50†	57.69	39.36	29.79	43.53	22.98	16.25
	IncV3	57.28	39.32	29.35	44.66	23.76	17.31
LSTA-GRU	Res-50*	57.30	37.59	29.17	43.94	22.16	15.94
	Res-34**	60.61	40.84	32.04	45.37	23.49	16.59
	Res-34***	61.31	40.93	32.14	44.90	22.60	16.25
LSTA-2S	Res-34	62.12	40.41	32.60	48.89	24.27	18.71
HF-TSN	BNInc	57.57	39.90	28.09	42.40	25.23	16.93
	Res-50	56.69	40.70	29.38	45.48	24.55	17.38
Ensemble1		63.2	43.28	33.45	48.62	26.8	20.14

†:Kinetics pre-trained; *: Fine-tuned layers: GRU; **: Fine-tuned layers: GRU+LSTA; ***: Fine-tuned layers: GRU+LSTA+Conv5_3

Results

Method	Backbone	S1			S2		
		Verb	Noun	Action	Verb	Noun	Action
LSTA	Res-34	58.25	38.93	30.16	45.51	23.46	15.88
	Res-50	57.81	37.84	29.54	44.38	22.53	15.98
	Res-50†	57.69	39.36	29.79	43.53	22.98	16.25
	IncV3	57.28	39.32	29.35	44.66	23.76	17.31
LSTA-GRU	Res-50*	57.30	37.59	29.17	43.94	22.16	15.94
	Res-34**	60.61	40.84	32.04	45.37	23.49	16.59
	Res-34***	61.31	40.93	32.14	44.90	22.60	16.25
LSTA-2S	Res-34	62.12	40.41	32.60	48.89	24.27	18.71
HF-TSN	BNInc	57.57	39.90	28.09	42.40	25.23	16.93
	Res-50	56.69	40.70	29.38	45.48	24.55	17.38
Ensemble1		63.2	43.28	33.45	48.62	26.8	20.14
Ensemble2		**63.34**	**44.75**	**35.54**	**49.37**	**27.11**	**20.25**

†:Kinetics pre-trained; *: Fine-tuned layers: GRU; **: Fine-tuned layers: GRU+LSTA; ***: Fine-tuned layers: GRU+LSTA+Conv5_3

Results

LSTA

- Trained on 1 GPU with batch size 32
- Parameter tuning done on GTEA61 dataset

Results

LSTA

- Trained on 1 GPU with batch size 32
- Parameter tuning done on GTEA61 dataset

HF-TSN

- Trained on 2 GPUs with batch size 32
- Parameter tuning done on 20BN something-v1 dataset

Key Takeaways

Two model families encode complementary features

Ensemble resulted in **35.54%** and **20.25%** action recognition accuracy

References

[1] S. Sudhakaran, S. Escalera, and O. Lanz. LSTA: Long Short-Term Attention for Egocentric Action Recognition. CVPR, 2019

[2] S. Sudhakaran, S. Escalera, and O. Lanz. Hierarchical Feature Aggregation Networks for Video Action Recognition. arXiv:1905.12462

https://github.com/swathikirans/LSTA