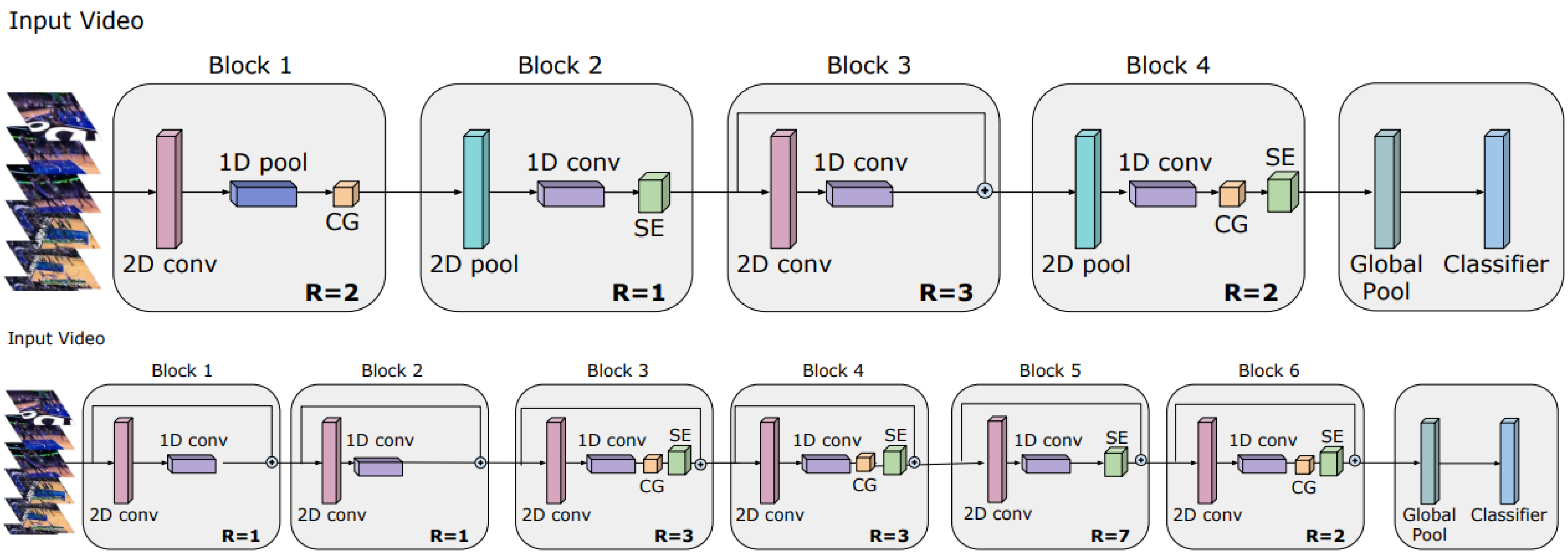
Tiny Video Networks on UCF50 dataset

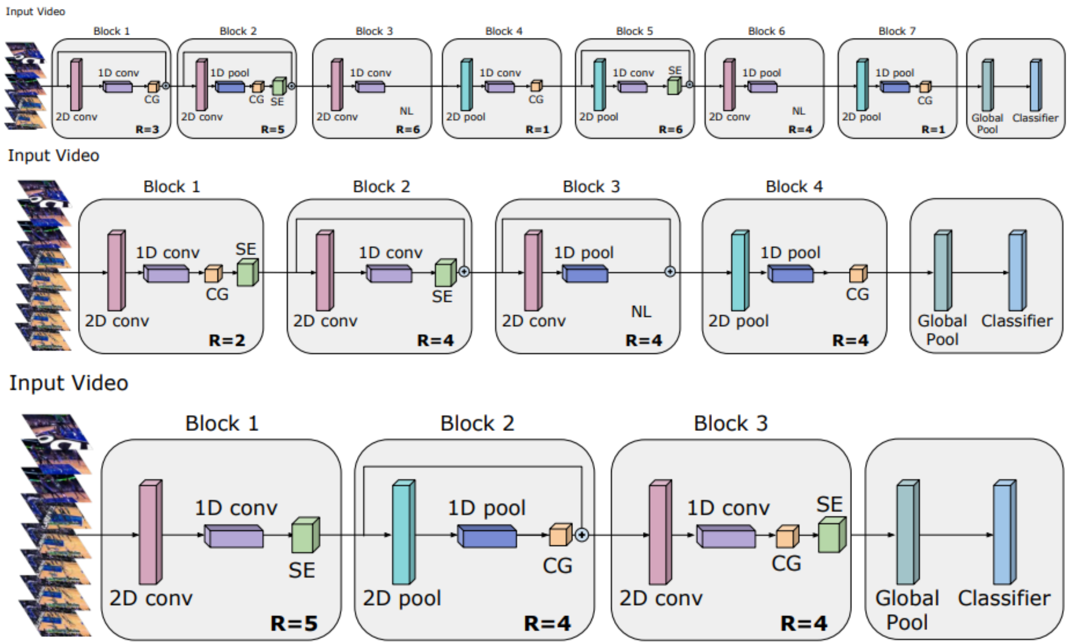


Выполнила Чувиляева Ирина

Постановка задачи

- Воспроизвести одну из сетей, описанный в статье Tiny Video Networks, AJ Piergiovanni Anelia Angelova Michael Ryoo
- Обучить её на датасете UCF50
- Сравнить с известными результатами

О статье

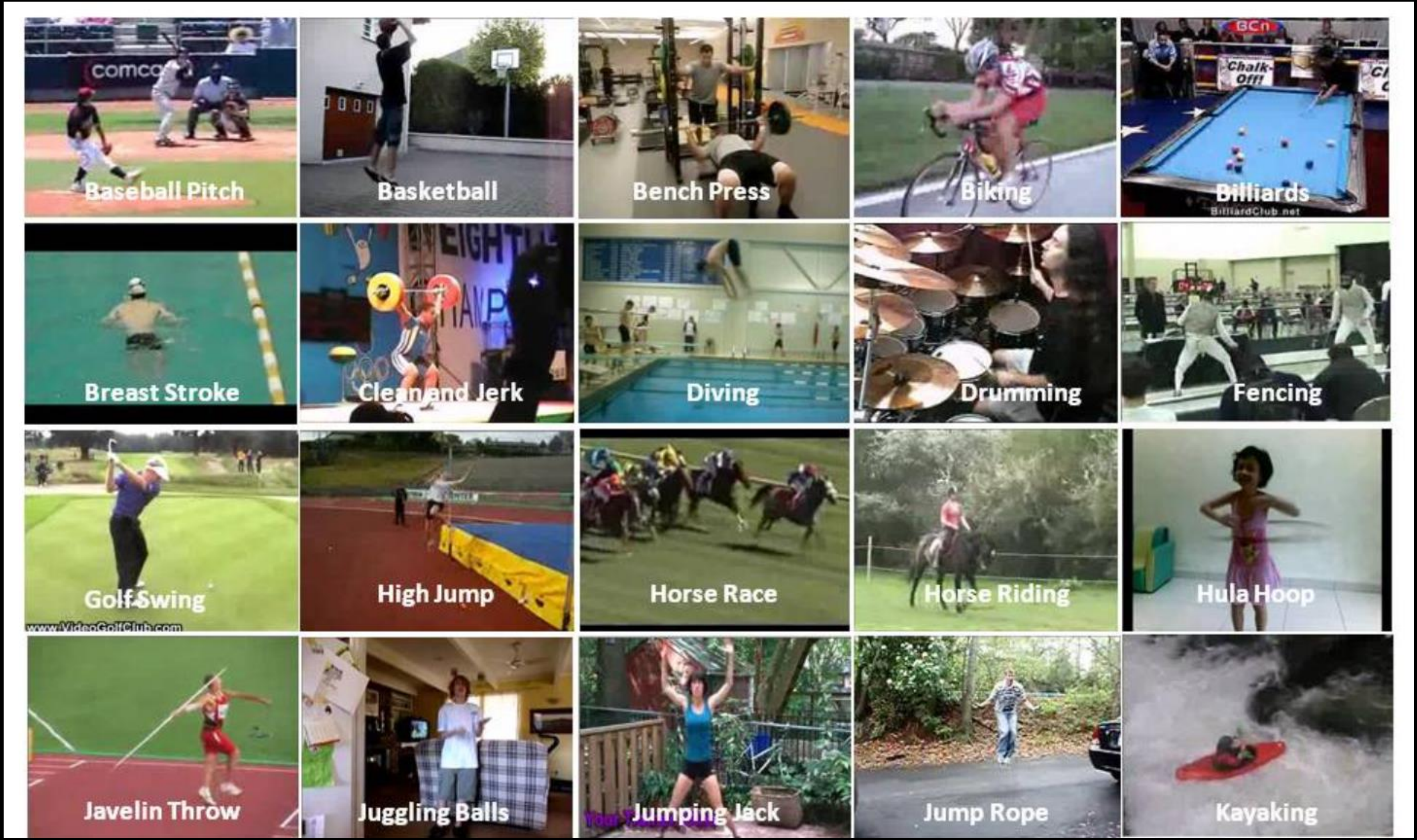


3: Example Tiny Video Networks found using architecture evolution showing several blocks with different configurations. A Tiny Video Net has multiple blocks, each repeated R times. Each block has a different configuration with spatial and temporal convolution, pooling, non-local layers, context gating and squeeze-excitation layers. It can select the image resolution and frame rate. From top to bottom: TVN-2, TVN-3, TVN-4. TVN-1 is shown in Figure 1.

Method	Runtime CPU(ms)	Runtime GPU(ms)	GFlops	Acc. (%)
ResNet-18	2120	105	38	21.1
ResNet-34	2256	110	50	24.2
ResNet-50	3022	125	124	28.1
ResNet-101	3750	140	245	30.2
TSN [47]	-	-	-	24.1
2DResNet50 [26]	-	-	-	27.1
bLVNet-TAM [7]	-	-	-	***31.4***
TVN-1	37	10	13	23.1
TVN-2	65	13	17	24.2
TVN-3	85	16	69	25.9
TVN-4	402	19	106	28.0
TVN-5	86	16	52	**29.8**
TVN-6	142	18	93	**30.7**

Table 3. Results on the MiT dataset comparing different Tiny Networks to baselines and state-of-the-art (which are all RGB-only). TVN models perform competitively and are also much faster. No runtime was reported in prior work.

О датасете UCF50



О реализации

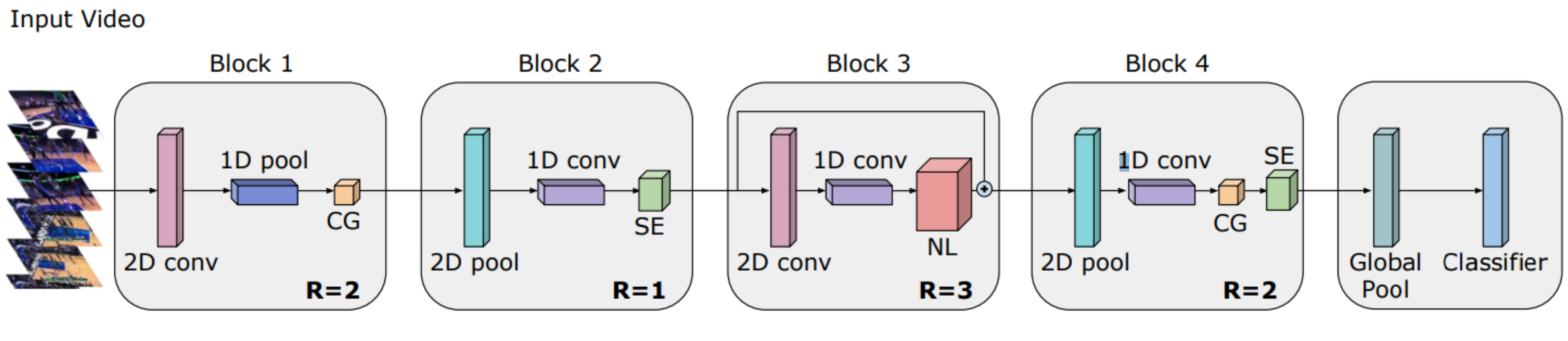


Figure 1: An example of a highly efficient 'Tiny Video Network', working on a video snippet. TVN-1 is shown. It takes 37 ms (CPU), 10ms (GPU).

- Архитектура TVN1
- Входной размер 200x200
- Нет аугментации при обучении
- Нет предобучения

- Соотношение train/test: 85/15
- Берём каждый 10ый кадр из видео
- Всего берём 8 кадров (из 80)
- Если кадров не хватает – повторяем последний

Обучение

localhost:8888/notebooks/task_project/tvn1_code/train_tvn1_framed.ipynb

jupyter train_tvn1_framed Last Checkpoint: несколько секунд назад (unsaved changes)

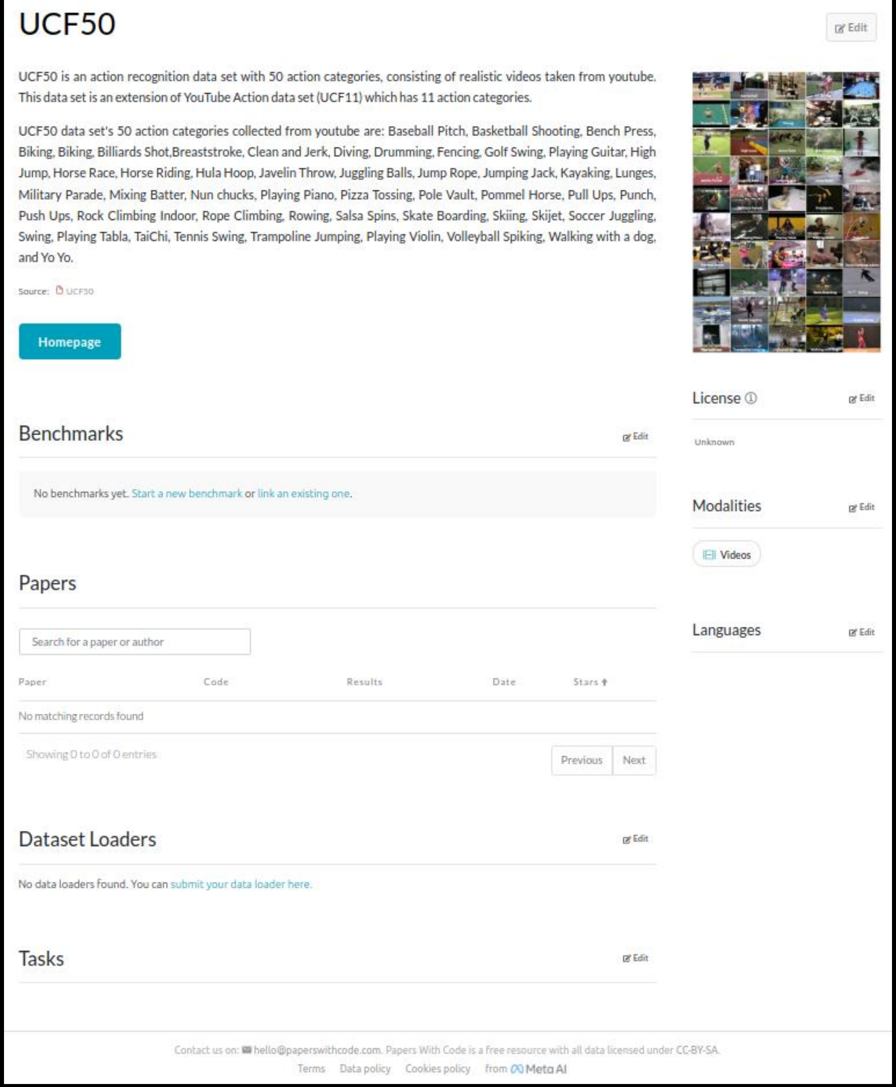
File Edit View Insert Cell Kernel Widgets Help — Trusted — Python 3 (ipykernel)

```
100%|██████████████████████████████| 1420/1420 [14:32<00:00,  1.63it/s]
epoch  6 :
mean loss =  tensor(3.9050)
100%|██████████████████████████████| 251/251 [01:44<00:00,  2.41it/s]
mean val loss =  tensor(3.9034)
accuracy =  0.0037387836490528413
100%|██████████████████████████████| 1420/1420 [14:29<00:00,  1.63it/s]
epoch  7 :
mean loss =  tensor(3.9050)
100%|██████████████████████████████| 1420/1420 [14:33<00:00,  1.63it/s]
epoch  8 :
mean loss =  tensor(3.9050)
100%|██████████████████████████████| 251/251 [01:46<00:00,  2.36it/s]
mean val loss =  tensor(3.9029)
accuracy =  0.0037387836490528413
100%|██████████████████████████████| 1420/1420 [14:37<00:00,  1.62it/s]
epoch  9 :
mean loss =  tensor(3.9045)
100%|██████████████████████████████| 1420/1420 [14:28<00:00,  1.63it/s]
epoch  10 :
mean loss =  tensor(3.9049)
100%|██████████████████████████████| 251/251 [01:44<00:00,  2.39it/s]
mean val loss =  tensor(3.9056)
accuracy =  0.0037387836490528413
100%|██████████████████████████████| 1420/1420 [14:31<00:00,  1.63it/s]
epoch  11 :
mean loss =  tensor(3.9048)
 35%|██████████                    | 501/1420 [05:10<09:45,  1.57it/s]
```

Сравнение с другими моделями

Performance	Experimental Setup	Paper
76.90%	Leave One Group Out Cross-validation (25 cross-validations)	Reddy and Shah. (MVAP), 2012
57.90%	5-fold group-wise cross-validation	Sadanand and Corso. (CVPR), 2012
76.40%*	Video Wise Cross-validation (*Since videos belonging to a group are obtained from a single long video, similar videos can end up in both training and testing in "video-wise cross-validation" leading to high performance)	Sadanand and Corso. (CVPR), 2012
81.03%*	2/3 training and 1/3 testing for each class (*From the details given in the paper, we are not sure if videos belonging to the same group are kept seperate in training and testing sets and the paper does not give details on number of cross-validations)	Todorovic. (ECCV), 2012
73.70%	Leave One Group Out Cross-validation (25 cross-validations)	Solmaz, et al. (MVAP), 2012
72.60%	Leave One Group Out Cross-validation (25 cross-validations)	Kliper-Gross, et al. (ECCV), 2012

Сравнение с другими моделями



UCF50

Edit

UCF50 is an action recognition data set with 50 action categories, consisting of realistic videos taken from youtube. This data set is an extension of YouTube Action data set (UCF11) which has 11 action categories.

UCF50 data set's 50 action categories collected from youtube are: Baseball Pitch, Basketball Shooting, Bench Press, Biking, Biking, Billiards Shot,Breaststroke, Clean and Jerk, Diving, Drumming, Fencing, Golf Swing, Playing Guitar, High Jump, Horse Race, Horse Riding, Hula Hoop, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Lunges, Military Parade, Mixing Batter, Nun chucks, Playing Piano, Pizza Tossing, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Skate Boarding, Skiing, Skijet, Soccer Juggling, Swing, Playing Tabla, TaiChi, Tennis Swing, Trampoline Jumping, Playing Violin, Volleyball Spiking, Walking with a dog, and Yo Yo.

Source: UCF50

Homepage

License

Unknown

Modalities

Videos

Languages

Benchmarks

No benchmarks yet. Start a new benchmark or link an existing one.

Papers

Search for a paper or author

Paper	Code	Results	Date	Stars
No matching records found				

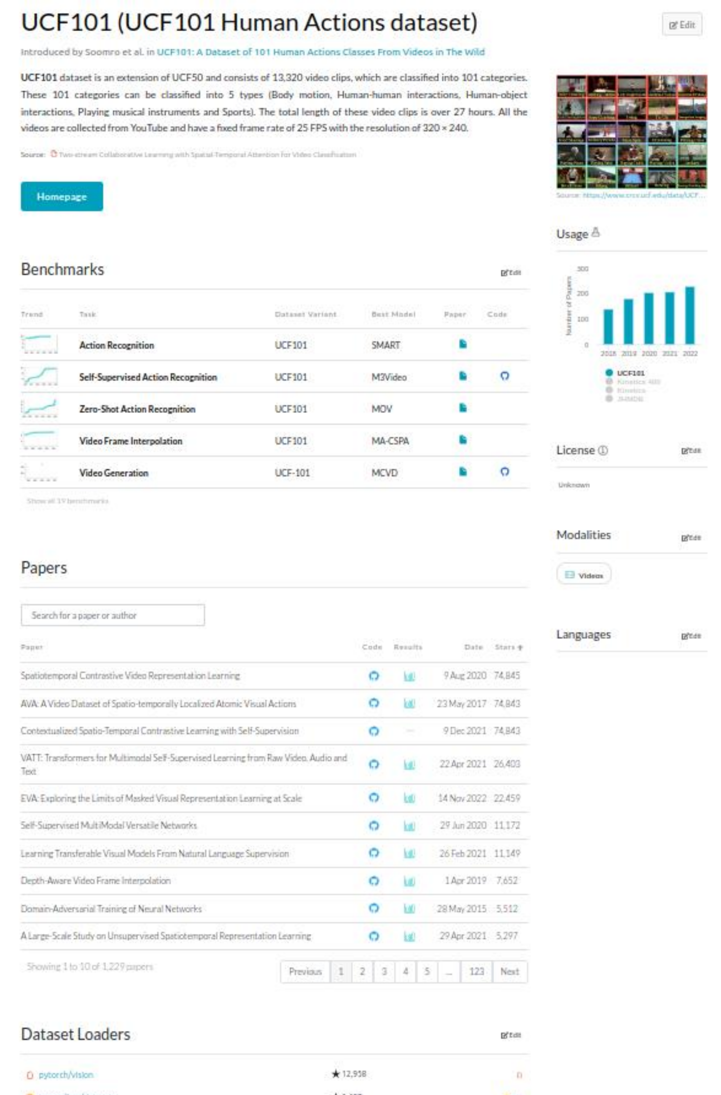
Showing 0 to 0 of 0 entries

Previous Next

Dataset Loaders

No data loaders found. You can submit your data loader here.

Tasks



UCF101 (UCF101 Human Actions dataset)

Edit

Introduced by Soomro et al. in UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild

UCF101 dataset is an extension of UCF50 and consists of 13,320 video clips, which are classified into 101 categories. These 101 categories can be classified into 5 types (Body motion, Human-human interactions, Human-object interactions, Playing musical instruments and Sports). The total length of these video clips is over 27 hours. All the videos are collected from YouTube and have a fixed frame rate of 25 FPS with the resolution of 320 × 240.

Source: Two-stream Collaborative Learning with Spatial-Temporal Attention for Video Classification

Homepage

Usage

UCF101

License

Unknown

Modalities

Videos

Languages

Benchmarks

Trend	Task	Dataset Variant	Best Model	Paper	Code
	Action Recognition	UCF101	SMART		
	Self-Supervised Action Recognition	UCF101	M3Video		
	Zero-Shot Action Recognition	UCF101	MOV		
	Video Frame Interpolation	UCF101	MA-CSPA		
	Video Generation	UCF-101	MCVD		

Papers

Search for a paper or author

Paper	Code	Results	Date	Stars
Spatiotemporal Contrastive Video Representation Learning			9 Aug 2020	74,845
AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions			23 May 2017	74,843
Contextualized Spatio-Temporal Contrastive Learning with Self-Supervision		—	9 Dec 2021	74,843
VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text			22 Apr 2021	26,403
EVA: Exploring the Limits of Masked Visual Representation Learning at Scale			14 Nov 2022	22,459
Self-Supervised MultiModal Versatile Networks			29 Jun 2020	11,172
Learning Transferable Visual Models From Natural Language Supervision			26 Feb 2021	11,149
Depth-Aware Video Frame Interpolation			1 Apr 2019	7,652
Domain-Adversarial Training of Neural Networks			28 May 2015	5,512
A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning			29 Apr 2021	5,297

Showing 1 to 10 of 1,229 papers

Previous 1 2 3 4 5 ... 123 Next

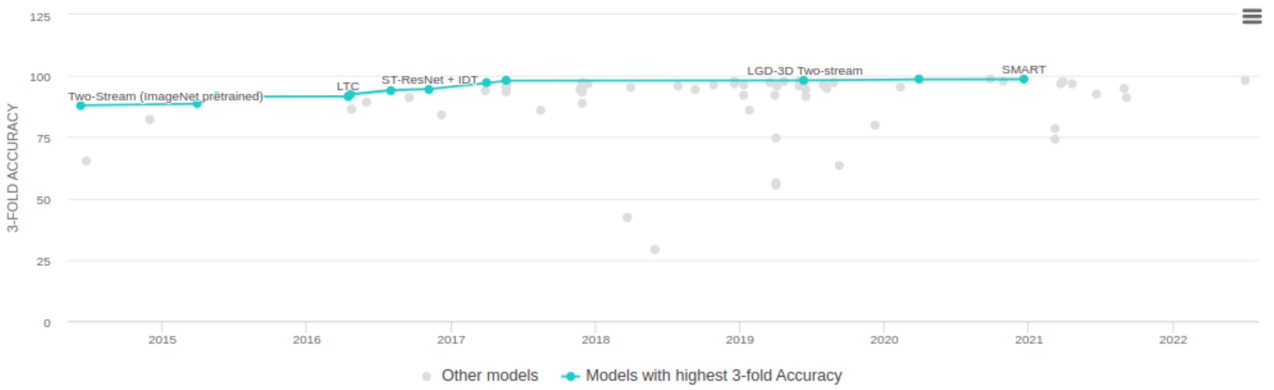
Dataset Loaders

pytorch/vision ★ 12,958

Action Recognition on UCF101

Leaderboard   Dataset

View 3-fold Accuracy by Date for All models



Filter: LSTM ResNet GCN untagged

Edit Leaderboard

Rank	Model	3-fold Accuracy	Accuracy	Extra Training Data	Paper	Code	Result	Year	Tags
1	**SMART**	98.64		×	SMART Frame Selection for Action Recognition			2020	
2	**OmniSource** (SlowOnly-8x8-R101-RGB + I3D-Flow)	98.6		✓	Omni-sourced Webly-supervised Learning for Video Recognition			2020	
3	**PERF-Net** (multi-distilled S3D)	98.6		×	PERF-Net: Pose Empowered RGB-Flow Net			2020	
4	**LGD-3D Two-stream**	98.2		×	Learning Spatio-Temporal Representation with Local and Global Diffusion			2019	
5	**Text4Vis**	98.2		×	Revisiting Classifier: Transferring Vision-Language Models for Video Recognition			2022	

SMART на данный момент является SOTA для аналогичной задачи на UCF101

Method	Backbone	UCF101	HMDB51
Two-stream	VGG	92.5	62.4
I3D	Inc v3	98.0	80.7
DynaMotion + I3D	Inc v3	98.4	84.2
TSN	BN-Inc	94.2	69.9
KI-Net	Res-152	97.8	78.2
AAS	TSN	94.6	71.2
SMART	TSN	**95.8**	**74.6**
AAS	TSN+Kinetics	96.8	77.3
SMART	TSN+Kinetics	**98.6**	**84.3**

Table 5: Extending SMART to other approaches

Method	UCF101	HMDB51
ISTPAN	95.5	70.7
ISTPAN + SMART	**96.4**	**72.1**
I3D	98.0	80.0
I3D + Smart	**98.2**	**81.1**
STM-Resnet	94.2	68.9
STM-Resnet + SMART	**94.9**	**69.7**

Проблемы

- Сравнительно малеником датасет UCF50, который входит в датасет побольше.
- Для работы с видео требуется значительно больше параметров
- Из-за отсутствия аугментаций модели не хватает информации, чтобы что-либо выучить