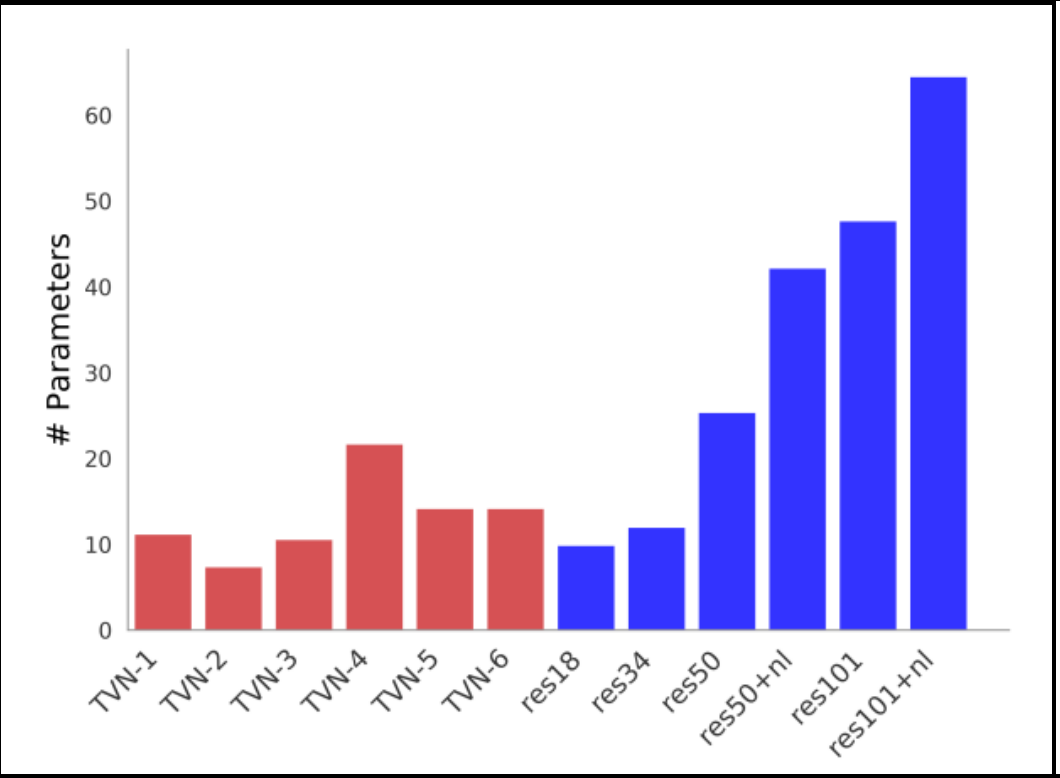
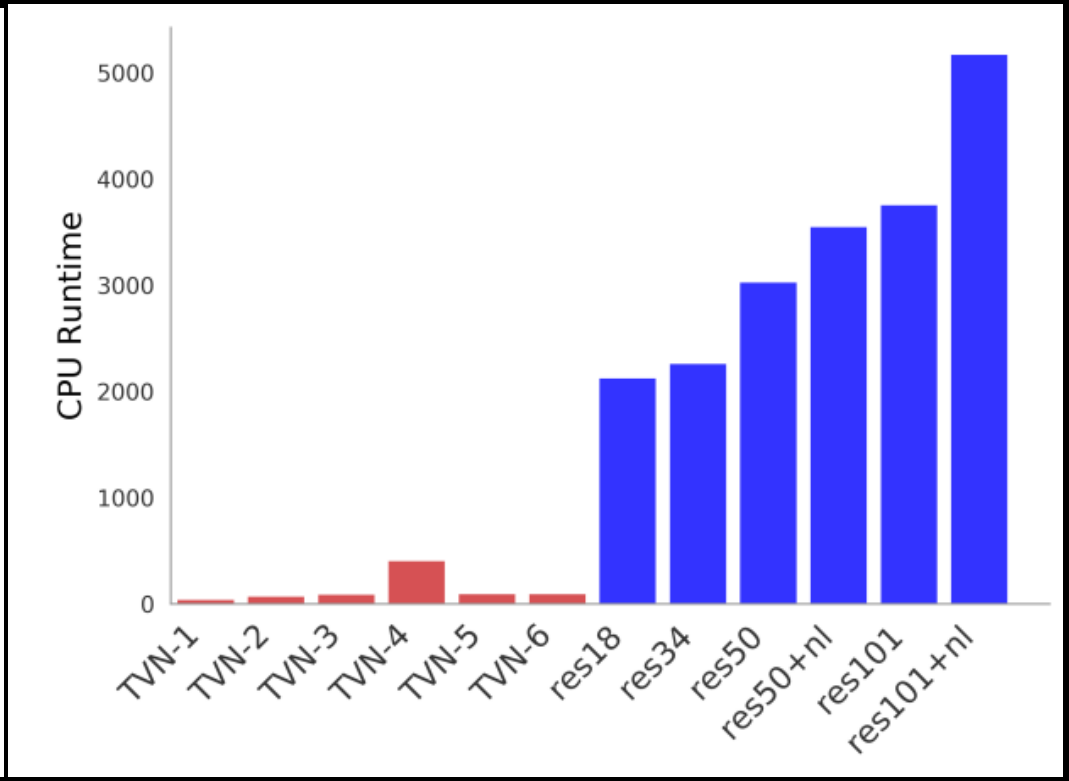
# Tiny Video Networks on UCF50 dataset



# Выбор архитектуры

| Model | Evolved on | Runtime CPU/GPU | Image resol. | $f$ | $s$ |
|---|---|---|---|---|---|
| TVN-1 | MiT | 37 / 10 | 224x224 | 2 | 4 |
| TVN-2 | MLB | 65 / 13 | 256x256 | 2 | 7 |
| TVN-3 | Charades | 85 / 16 | 160x160 | 8 | 2 |
| TVN-4 | MiT | 402 / 19 | 128x128 | 8 | 4 |
| TVN-5 | MiT | 86 / 16 | 160x160 | 16 | 4 |
| TVN-6 | MiT | 142 / 18 | 160x160 | 32 | 4 |

Table 1. TVN models. Runtime is in milliseconds (ms).

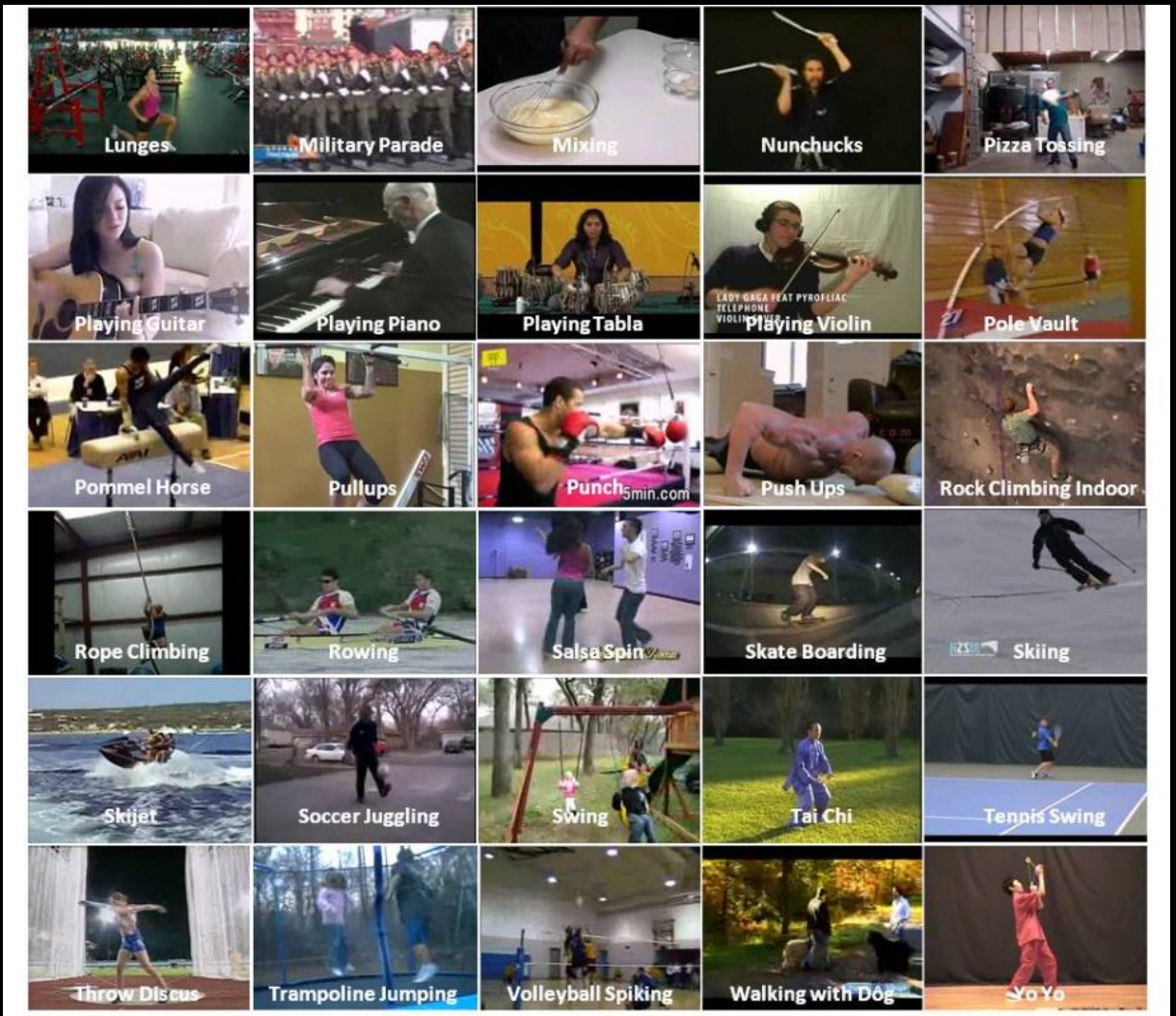
| Method | Runtime CPU(ms) | Runtime GPU(ms) | GFlops | Acc. (%) |
|---|---|---|---|---|
| ResNet-18 | 2120 | 105 | 38 | 21.1 |
| ResNet-34 | 2256 | 110 | 50 | 24.2 |
| ResNet-50 | 3022 | 125 | 124 | 28.1 |
| ResNet-101 | 3750 | 140 | 245 | 30.2 |
| TSN [47] | - | - | - | 24.1 |
| 2DResNet50 [26] | - | - | - | 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.

| Method | Runtime CPU/GPU | GFlops | mAP |
|---|---|---|---|
| Asyn-TF, VGG16 [38] | - | - | 22.4 |
| I3D [3] | - | 216 | 32.9 |
| Nonlocal, R101[48] | - | 544 × 30 | 37.5 |
| SlowFast, two-str. [9] | 3594/135 | 213 × 30 | 42.1 |
| SlowFastNL, two-str. [9] | 4354/152 | 234 × 30 | 45.2 |
| X3D-XL (Kin400 pretr) [8] | - | 48.6 × 30 | 43.4 |
| X3D-XL (Kin600 pretr) [8] | - | 48.6 × 30 | ***47.1*** |
| TVN-1 (from scratch) | 37/10 | 13 | 40.4 |
| TVN-2 (from scratch) | 65/13 | 17 | **47.4** |
| TVN-3 (from scratch) | 85/16 | 69 | **52.0** |
| TVN-4 (from scratch) | 402/19 | 106 | **53.8** |
| TVN-5 (from scratch) | 86/16 | 52 | **52.4** |
| TVN-6 (from scratch) | 142/18 | 93 | **52.8** |
| TVN-1 (MiT pretr) | 37/10 | 13 | 42.1 |
| TVN-2 (MiT pretr) | 65/13 | 17 | **48.3** |
| TVN-3 (MiT pretr) | 85/16 | 69 | **53.2** |
| TVN-4 (MiT pretr) | 402/19 | 106 | **53.9** |
| TVN-5 (MiT pretr) | 86/16 | 52 | **54.2** |
| TVN-6 (MiT pretr) | 142/18 | 93 | **54.6** |

Table 2. Comparison to the state-of-the-art on Charades. We report the best TVN models in bold, and the best prior work models as bolded italics. Many TVNs, even without pretraining, outperform the SOTA which uses strong Kinetics pre-training. TVNs

# Датасет: UCF50 - Action Recognition Data Set




50 Action Categories consisting of realistic YouTube videos

25 Groups of Videos per Action Category

133 Average Videos per Action Category

199 Average Number of Frames per Video

320 Average Frames Width per Video

240 Average Frames Height per Video

# Ключевые особенности

- Выбрана базовая архитектура TVN-1. Возможен переход к более сложным и тяжёлым вариантам.
- В силу легковестности обучение должно занимать 2-3 часа на GPU, но стоит учесть возможные риски.
- Датасет UCF-50 содержит нарезанные видио с Youtube. Perfomance на них может сильно отличаться
- Код для статьи реализован на tensorflow и включает в себя сразу несколько моделей, найденных с помощью NAS. Предполагается переписать зафиксированную архитектуру на pytorch, и адаптировать её для нового датасета.