ОБНАРУЖЕНИЕ АКТИВНОСТЕЙ В ВИДЕОПОТОКЕ С ПОМОЩЬЮ СЕТЕЙ ВРЕМЕННЫХ СЕГМЕНТОВ

Бокиев Н.С., оператор 4 научной роты ФГАУ «Военный инновационный технополис «ЭРА».

Аннотация

Глубокое обучение достигли больших успехов в распознавании на статических изображениях. Однако для распознаваний действий в видеопотоке преимущество перед традиционными методами не столь существенны. В данной статье описывается метод обнаружения активностей в видеопотоке с помощью сетей временных сегментов, которые моделируют пространственно-временную структуру на уровне видео, что позволяет эффективно решать задачи

Ключевые слова: компьютерное зрение, понимание видео, распознавание действий, распознавание событий, классификация видео, машинное обучение, машинное зрение, глубокое обучение, нейронные сети, сверточные нейронные сети

Распознавание действий на основе видео привлекло значительное внимание со стороны академического сообщества [1,2,3]. Данные работы могут быть применены во многих областях, таких как безопасность и анализ поведения. В распознавании действий есть два важнейших и взаимодополняющих аспекта: статика и динамика. Эффективность системы распознавания в значительной степени зависит от того, способна ли она извлекать и использовать из нее соответствующую информацию. Однако извлечение такой информации нетривиально из-за ряда сложностей, таких как изменение масштаба, изменение угла обзора и движения камеры. Таким образом, становится крайне важным разработать эффективные представления, которые могут справиться с этими проблемами, сохраняя при этом категориальную информацию классов действий. В последнее время сверточные нейронные сети (СНС) добились больших успехов в классификации объектов на изображении, сцен и сложных событий [4,5]. СНС так же используются для решения проблемы распознавания действий на основе видео [6, 7]. Глубокие СНС обладают большой способностью к моделированию и способны обучаться распознавательному представлению на больших наборах данных. Однако, в отличие от классификации изображений, в задаче распознавания действий сквозные глубокие СНС по-прежнему не могут достичь значительного преимущества по сравнению с традиционными методами на основе признаков, сконструированных вручную.

Сложности заключаются в следующем. Во-первых, долговременная временная структура играет важную роль в понимании действия [8, 9]. Тем не менее, основные структуры СНС обычно фокусируются на статике и кратковременных движениях, таким образом, не имея возможности включить долговременную временную структуру. В последнее время предпринимаются несколько попыток [10, 11] решить эту проблему. Эти методы в основном полагаются на плотную временную выборку с заранее заданным интервалом выборки. Этот подход будет сопряжен с чрезмерными вычислительными затратами при применении к длинным видеопоследовательностям, что ограничивает его применение в реальной практике и создает риск пропуска важной информации для видео, длина которой превышает максимальную длину последовательности. Во-вторых, на практике обучение глубоких СНС требует большого объема обучающих образцов для достижения приемлемого качества. Однако из-за сложности сбора и аннотации данных общедоступные наборы данных по распознаванию действий (например, UCF101 [12], HMDB51 [13]) остаются ограниченными как по размеру, так и по разнообразию. Следовательно, очень глубокие СНС, которые достигли значительных успехов в классификации изображений, сталкиваются с высоким риском переобучения.

Таким образом, необходимо решить две проблемы: поиск эффективного преставления пространственно-временной структуры на большие временные промежутки и обучение глубоких СНС, имея при этой небольшой объем данных.

Решая первую проблему, можно обнаружить, что последовательные кадры избыточны, потому что они оказываются слишком похожими. Вместо этого можно брать кадры по разреженной схеме. Сначала видеопоследовательность равномерно разделяется на фрагменты, затем анализируется по

одному кадру из фрагмента. Это позволяет осуществлять сквозное обучение по длинным видеопоследовательностям при разумном бюджете времени и вычислительных ресурсов.

Для максимизации качества сети, предлагается использовать продвинутые техники:

1. Очень глубокие архитектуры CHC
2. Кросс-модальная предварительная инициализация
3. Регуляризация
4. Аугментация данных

Проблема двухпотокотовых сетей в их неспособности моделирования пространственно-временной структуры на большой протяженности. Это заложено в их архитектуре: для изучения пространственной информации они “смотрят” на один фрейм, а для изучения временной структуры - на стек фреймов небольшого размера. Но есть действия, например, спортивные, которые включают в себя несколько этапов и для их понимания нужно “взглянуть” на более длительный участок видео. Для решения этой проблемы можно предложить схему временных сегментов (Рисунок 1):

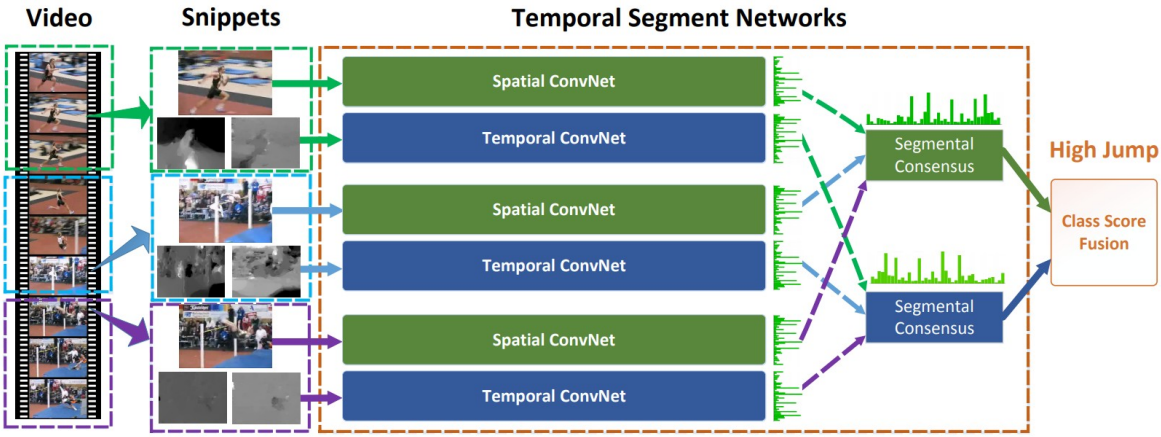

Рисунок 1 - Сверточная нейронная сеть временных сегментов

Видеопоследовательность на первом этапе равномерно делится на заданное количество фрагментов, из каждого фрагмента выбирается кадр, для которого предсказывается действие. Предсказания на всех фреймов агрегируются с помощью функции консенсуса. В процессе обучения функция ошибки будет учитывать каждый ошибочный прогноз.

Формально, модель принимает на вход видео V, которое делится на K сегментов {S1, S2, ⋯, SK} равной длительности. Затем сеть временных сегментов моделирует последовательность фрагментов следующим образом:

$$\mathrm{TSN}(T_1, T_2, \cdots, T_K) = \mathcal{H}(\mathcal{G}(\mathcal{F}(T_1; \mathbf{W}), \mathcal{F}(T_2; \mathbf{W}), \cdots, \mathcal{F}(T_K; \mathbf{W}))). \quad (1)$$

Где (T1, T2, · · ·, Tk) - последовательности кадров, каждый Tk случайным образом выбирается из соответствующего сегмента Sk. F(Tk; W) - функция, которая представляет CHC с параметрами W, которая работает на кадре Tk и возвращает распределение классов. G - сегментарная консенсусная функция, которая объединяет выходы F по всем кадрам. H – функция softmax, которая на основе консенсуса из G прогнозирует вероятность каждого класса.

В сочетании со стандартной категориальной кросс-энтропийной потерей итоговая функция потерь относительно сегментарного консенсуса G = G (F (T1; W), F (T2; W), · · ·, F (TK; W)) формируется как:

$$\mathcal{L}(y, \mathbf{G}) = -\sum_{i=1}^{C} y_i \left(G_i - \log \sum_{j=1}^{C} \exp G_j \right) \quad (2)$$

Где C - количество классов действий, yi - метки входных данных. Gi = g(Fi(T1), . . . , Fi(TK)), где Gi - оценка класса выведенный из оценок это же класса по всем фрагментам путем агрегирования функцией g, которая равномерно усредняет оценки.

Данная сеть дифференцируема, поэтому можно использовать все K фрагментов, объединив их для оптимизации параметров модели с помощью алгоритма обратного распространения ошибки:

$$\frac{\partial \mathcal{L}(y, \mathbf{G})}{\partial \mathbf{W}} = \frac{\partial \mathcal{L}}{\partial \mathbf{G}} \sum_{k=1}^{K} \frac{\partial \mathcal{G}}{\partial \mathcal{F}(T_k)} \frac{\partial \mathcal{F}(T_k)}{\partial \mathbf{W}} \quad (3)$$

Когда мы оптимизируем веса нейронной сети, например, стохастическим градиентным спуском, оптимизатор гарантирует, что обновления параметров происходит с учетом всего видео, а не по короткому фрагменту. Между тем, фиксируя K для всех видео, мы собираем разреженную временную стратегию выборки, где отобранные фрагменты содержат только небольшую часть кадров. Это резко снижает вычислительные затраты на оценку CHC на кадрах, по сравнению с предыдущими работами, использующими плотно отобранные кадры.

Архитектура сети является важным фактором в проектировании нейронных сетей. В ряде работ показано, что более глубокие структуры повышают эффективность распознавания объектов. Однако исходные двухпоточные CHC [1] использовали относительно неглубокую сетевую структуру (ClarifaiNet [14]). В этой работе в качестве строительного блока мы выбираем Inception с пакетной нормализацией (BN-Inception) [15], что обусловлено его хорошим балансом между точностью и производительностью. Пространственный поток CHC работает на одном изображении RGB, а временной поток CHC принимает стек последовательных полей оптического потока в качестве входных данных.

В качестве входа рассматриваются несколько модальностей (Рисунок 2):

- RGB изображение для пространственного потока
- Стек оптического потока для временного потока
- Разносить RGB
- Искаженные оптические поля

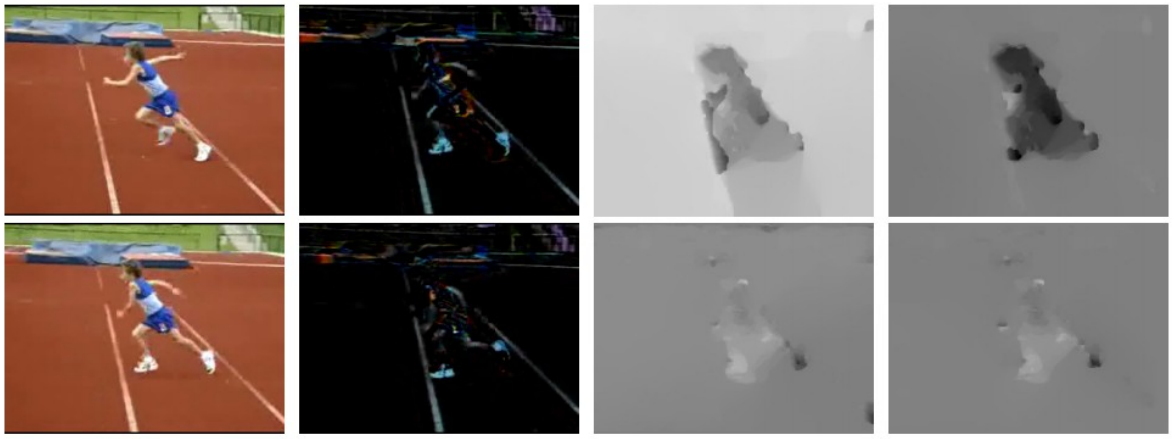

Рисунок 2 - RGB изображения, Разность RGB изображений, оптический поток (направления x, y), искаженных оптический поток (направления x, y)

Одно RGB изображение обычно кодирует статическую информацию в определенный момент времени и не содержит концептуальную информацию о предыдущем и следующем кадрах.

Как показано на Рисунке 2, разница RGB между двумя последовательными кадрами описывает изменение внешнего вида, которое может соответствовать области видимости движения. Вдохновленные [16], мы экспериментируем с добавлением сложенной разности RGB в качестве другой входной модальности и исследуем ее эффективность в распознавании действий. СНС временного потока принимают оптическое поле потока в качестве входного сигнала и стремятся захватить информацию о движении. В реалистичных видео, однако, обычно существует движение камеры, и поля оптического потока могут не концентрироваться на действиях человека.

Как показано на Рисунке 2, благодаря движению камеры на заднем плане выделяется значительное количество горизонтальных перемещений. Вдохновленные работой над улучшенными плотными траекториями [2], мы предлагаем использовать искаженные поля оптического потока в качестве дополнительной входной модальности. Следуя [2], мы извлекаем искривленный оптический поток сначала оценивая матрицу гомографии, а затем компенсируя движение камеры. Как показано на Рисунке 2, искаженный оптический поток подавляет фоновое движение.

Поскольку наборы данных для распознавания действий относительно невелики, обучение глубоких сетей сопряжено с риском чрезмерного переобучения. Чтобы смягчить эту проблему, предлагается следовать трем стратегиям обучения.

Кроссмодальная предобучение. Техника предварительного обучении оказалась эффективным способом инициализации СНС, когда целевой набор данных небольшой. Пространственная сеть получает RGB изображение, логично инициализировать предобученной на ImageNet. Другие же модальности, по сути, захватывают информацию о движении, поэтому их распределение отличается от распределения RGB изображений. Сначала мы дискретизируем поля оптического потока в интервале от 0 до 255 путем линейного преобразования. Этот шаг делает диапазон полей оптического потока одинаковым с изображениями RGB. Затем мы модифицируем веса первого слоя свертки моделей RGB для обработки ввода полей оптического потока. В частности, мы усредняем веса по каналам RGB и копируем это среднее значение по номеру канала временного сетевого входа. Этот метод инициализации работает очень хорошо для временных сетей и уменьшает эффект переобучения в экспериментах.

Регуляризация. Пакетная нормализация [15] является важным компонентом для решения проблемы ковариантного сдвига. В процессе обучения пакетная нормализация будет оценивать среднее значение активации и дисперсию внутри каждого пакета и использовать их для преобразования этих значений активации в стандартное гауссово распределение. Эта операция не только ускоряет сходимость обучения, но и приводит к переобучению в процессе переноса из-за предвзятой оценки распределений активации из ограниченного числа обучающих выборок. Поэтому после инициализации с предварительно подготовленными моделями мы замораживаем средние и дисперсионные параметры всех слоев пакетной нормализации, кроме первого. Так как распределение оптического потока отличается от RGB-изображений, то значение активации первого слоя свертки будет иметь другое распределение, и нам необходимо заново оценить среднее значение и дисперсию соответственно. Мы называем эту стратегию частичной BN. Между тем, мы добавляем дополнительный слой dropout после глобального пула в архитектуре BN-Inception для дальнейшего уменьшения эффекта переобучения. Коэффициент dropout устанавливается равным 0.8 для пространственного потока СНС и 0.7 для временного потока СНС.

Аугментация. Еще одним из классических способов борьбы с переобучением является увеличение объема данных. В качестве преобразований использовались масштабирование, угловое обрезы, кропы, горизонтальной отражение. В технике угловой обрезки выделенные области выбираются только из углов или центра изображения, чтобы избежать неявной фокусировки на центральной области изображения. В технике многомасштабного кадрирования мы адаптируем технику джиттера шкалы [4], используемую в классификации ImageNet, к распознаванию действий. Мы представляем эффективную реализацию джиттера шкалы. Мы фиксируем размер полей входного изображения или оптического потока 256×340, а ширина и высота области кадрирования выбираются случайным образом из {256, 224, 192, 168}. Наконец, эти кадрированные области будут изменены до размеров 224 × 224 для обучения

работы в сети. Фактически, эта реализация не только содержит джиттер масштаба, но и включает в себя джиттер соотношения сторон.

Так как CNN уровня фрагментов имеют общие параметры, можно выполнять фреймовую оценку моделей. Это позволяет сравнивать предложенную модель с другими архитектурами. Мы следуем схеме тестирования исходных двухпотоковых СНС [1], где мы выбираем 25 RGB-кадров или стеков оптического потока из боевиков. Затем мы обрезаем 4 угла и 1 центр, и их горизонтальное отражение от выбранных кадров для оценки СНС. Для объединения пространственных и временных сетей потоков мы берем их средневзвешенное значение.

Эксперименты проводились на двух больших датасетах: HMDB51 и UCF101. UCF101 содержит 101 классов, 13320 видеороликов. Схема оценки задачи соответствует THUMOS13 [17]. В наборе данных HMDB51 51 категорий, 6766 видеоклипов. Для оптимизации используется mini-batch SGD со следующими параметрами: batch size = 128; momentum = 0.9; learning rate (spatial networks) = 0.001, уменьшется до 1/10 каждые 1500 итераций (Всего 3 500); learning rate (spatial networks) = 0.005, уменьшается до 1/10 после 12000 и 18000 итераций (максимальное итерация 20000).

Для выделения оптического потока и искаженного оптического потока мы выбираем алгоритм оптического потока TV L1 [18], реализованный в OpenCV с CUDA. Результаты различных стратегий обучения показаны в таблице 1, а в таблице 2 показаны результат обучения временной сети на различных модальностях. По результатам можно сделать вывод, что оптический поток лучше улавливает информацию о движении, и иногда разница RGB может быть нестабильной для описания движений. С другой стороны, разница в RGB может служить низкокачественной, высокоскоростной альтернативой для описания движения.

Таблица 1 - Результаты обучения моделей

Training setting	Spatial ConvNets	Temporal ConvNets	Two-Stream
Baseline [1]	72.7%	81.0%	87.0%
From Scratch	48.7%	81.7%	82.9%
Pre-train Spatial(same as [1])	84.1%	81.7%	90.0%
+ Cross modality pre-training	84.1%	86.6%	91.5%
+ Partial BN with dropout	84.5%	87.2%	92.0%

Таблица 2 - Качество в зависимости от модальностей на датасете UCF101

Modality	Performance
RGB Image	84.5%
RGB Difference	83.8%
RGB Image + RGB Difference	87.3%
Optical Flow	87.2%
Warped Flow	86.9%
Optical Flow + Warped Flow	87.8%
Optical Flow + Warped Flow + RGB	**92.3%**
All Modalities	91.7%

В качестве функции консенсуса было выбрано усреднение, так как по результатам экспериментов, представленных в таблице 3, эта функцию дает наибольший прирост качества. Также были исследованы различные глубокие СНС, результаты отображены в таблице 4.

Таблица 3 - Исследование функций консенсуса на датасете UCF101

Consensus Function	Spatial ConvNets	Temporal ConvNets	Two-Stream
Max	85.0%	86.0%	91.6%
Average	85.7%	87.9%	**93.5%**
Weighted Average	86.2%	87.7%	92.4%

Таблица 4 - Результаты обучения в зависимости от типа глубоких сверточных нейронных сетей

Training setting	Spatial ConvNets	Temporal ConvNets	Two-Stream
Clarifai [1]	72.7%	81.0%	87.0%
GoogLeNet	77.1%	83.9%	89.0%
VGGNet-16	79.8%	85.7%	90.9%
BN-Inception	84.5%	87.2%	92.0%
BN-Inception+TSN	85.7%	87.9%	**93.5%**

Выбрав лучшие параметры, в таблице 5 показаны окончательные результаты сравнения с другими существующими решениями.

Таблица 5 - качество распознавания действий различных алгоритмов на датасетах HMDB51 и UCF101

HMDB51		UCF101	
DT+MVSV [37]	55.9%	DT+MVSV [37]	83.5%
iDT+FV [2]	57.2%	iDT+FV [38]	85.9%
iDT+HSV [25]	61.1%	iDT+HSV [25]	87.9%
MoFAP [39]	61.7%	MoFAP [39]	88.3%
Two Stream [1]	59.4%	Two Stream [1]	88.0%
VideoDarwin [18]	63.7%	C3D (3 nets) [13]	85.2%
MPR [40]	65.5%	Two stream +LSTM [4]	88.6%
FSTCN (SCI fusion) [28]	59.1%	FSTCN (SCI fusion) [28]	88.1%
TDD+FV [5]	63.2%	TDD+FV [5]	90.3%
LTC [19]	64.8%	LTC [19]	91.7%
KVMF [41]	63.3%	KVMF [41]	93.1%
TSN (RGB+Flow)	68.5%	TSN (2 modalities)	94.0%
TSN (RGB+Flow+Warped Flow)	69.4%	TSN (3 modalities)	94.2%

Визуализировав СНС с помощью DeepDraw [19], можно попробовать интерпретировать результаты и оценить работу моделей с различными подходами (Рисунок 3).

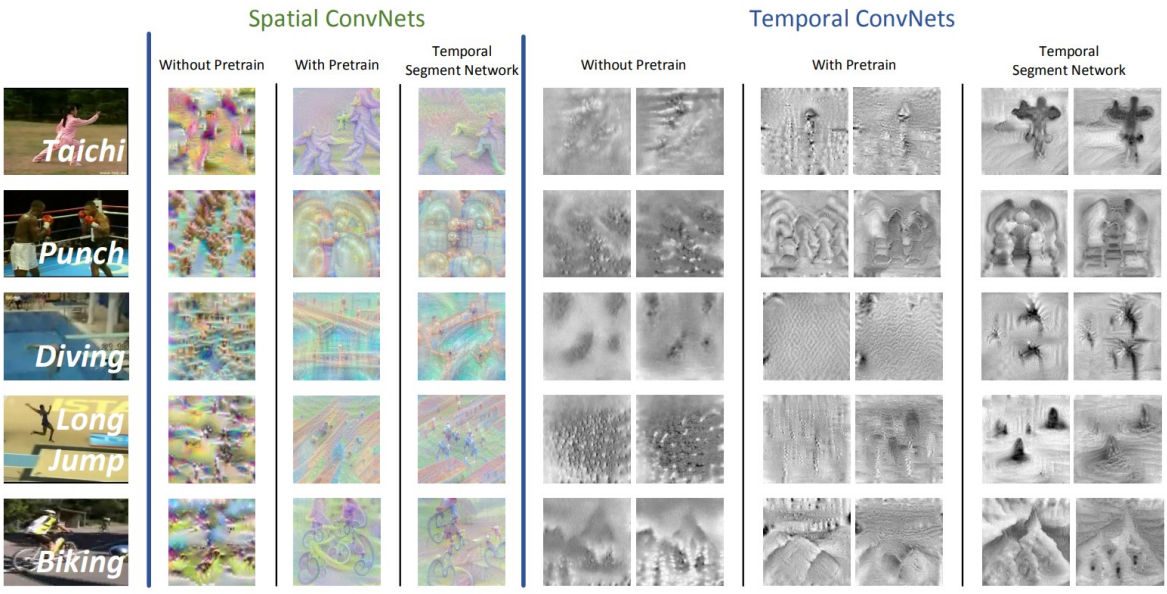

Рисунок 3 - Визуализация сверточных сетей с помощь DeepDraw [19]

Предварительно обученные модели более способны представлять визуальные концепции, чем модели без предварительной подготовки. Можно видеть, что как пространственные, так и временные модели без предварительной подготовки едва ли могут генерировать какую-либо осмысленную визуальную структуру. С помощью знаний, переданных из процесса предварительной подготовки, пространственные и временные модели способны улавливать структурированные визуальные паттерны. Также легко заметить, что модели, обученные только с кратковременной информацией, такой как одиночные кадры, склонны ошибочно принимать декорации и объекты в видео как важные свидетельства для распознавания действий.

В этой статье мы представили сеть временных сегментов, структуру видео-уровня, которая призвана моделировать долгосрочную временную структуру. Как было продемонстрировано на двух сложных наборах данных, эта работа вывела современное состояние на новый уровень при сохранении разумных вычислительных затрат. Это в значительной степени объясняется сегментарной архитектурой с редкой выборкой, а также рядом передовых методов, которые мы исследовали в этой работе. Первый обеспечивает эффективный и действенный способ захвата долговременной временной структуры, в то время как второй позволяет обучать очень глубокие сети на ограниченном наборе обучения без серьезного переобучения.

Литература

1. Karen Simonyan, Two-Stream Convolutional Networks for Action Recognition in Videos. [Электронный ресурс] URL: https://arxiv.org/pdf/1406.2199.pdf
2. Heng Wang, Action recognition with improved trajectories. [Электронный ресурс] URL: https://arxiv.org/pdf/1406.2199.pdf
3. Limin Wang, Motionlets: Mid-level 3D parts for human motion recognition. [Электронный ресурс] URL: https://wanglimin.github.io/papers/WangQT_CVPR13.pdf
4. Karen Simonyan, Very deep convolutional networks for large-scale image recognition. [Электронный ресурс] URL: https://arxiv.org/abs/1409.1556
5. Christian Szegedy, Going deeper with convolutions. [Электронный ресурс] URL: https://arxiv.org/abs/1409.4842
6. Limin Wang, Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns. [Электронный ресурс] URL: http://guoshengcv.github.io/papers/scene.pdf
7. Yuanjun Xiong, Recognize complex events from static images by fusing deep channels. [Электронный ресурс] URL: https://ieeexplore.ieee.org/document/7298768
8. Andrej Karpathy, Large-scale Video Classification with Convolutional Neural Networks. [Электронный ресурс] URL: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42455.pdf
9. Du Tran, Learning spatiotemporal features with 3d convolutional networks. [Электронный ресурс] URL: https://arxiv.org/abs/1412.0767
10. Juan Carlos Niebles, Modeling temporal structure of decomposable motion segments for activity classification. [Электронный ресурс] URL: https://link.springer.com/chapter/10.1007/978-3-642-15552-9_29
11. Adrien Gaidon, Temporal localization of actions with actoms. [Электронный ресурс] URL: https://ieeexplore.ieee.org/abstract/document/6487513
12. Khurram Soomro, UCF101: A dataset of 101 human actions classes from videos in the wild. [Электронный ресурс] URL: https://arxiv.org/pdf/1212.0402.pdf
13. Hildegard Kuehne, HMDB: A large video database for human motion recognition. [Электронный ресурс] URL: https://ieeexplore.ieee.org/document/6126543
14. Matthew D Zeiler, Visualizing and understanding convolutional networks. [Электронный ресурс] URL: https://arxiv.org/abs/1311.2901
15. Sergey Ioffe, Batch normalization: Accelerating deep network training by reducing internal covariate shift. [Электронный ресурс] URL: https://arxiv.org/abs/1502.03167
16. Lin Sun, Human action recognition using factorized spatio-temporal convolutional networks. [Электронный ресурс] URL: https://arxiv.org/abs/1510.00562

17. Haroon Idrees, THUMOS challenge: The THUMOS Challenge on Action Recognition for Videos "in the Wild". [Электронный ресурс] URL: https://arxiv.org/abs/1604.06182
18. Horst Bischof, A duality based approach for realtime tv-L 1 optical flow. In: 29th DAGM Symposium on Pattern Recognition. [Электронный ресурс] URL: https://link.springer.com/chapter/10.1007/978-3-540-74936-3_22
19. Audun Mathias, Deep draw. [Электронный ресурс] URL: https://github.com/auduno/deepdraw