Detect to Track and Track to Detect

Анна Воронцова

ФКН НИУ ВШЭ

Москва, 2017

Задача

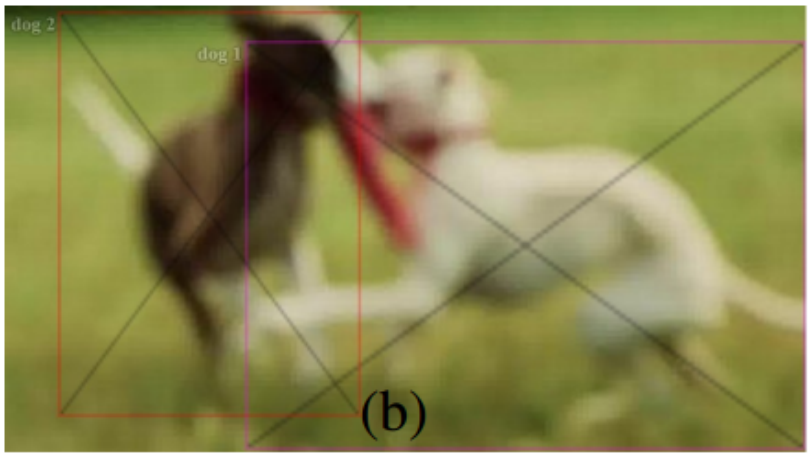
Детектировать объекты на видео (Detect) и отслеживать траектории их перемещения (Track).

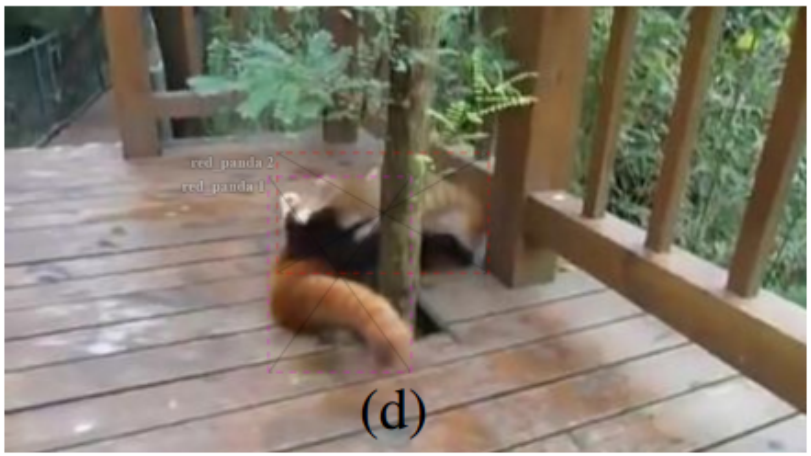
Задача

Детектировать объекты на видео (Detect) и отслеживать траектории их перемещения (Track).

- быстрое перемещение объекта в кадре => размытое изображение
- низкое качество видео (ср. с фото)
- в кадр попадает только часть объекта
- объект находится в нетипичном положении

Задача





(a) bicycle, bird, rabbit; (b) dog; (c) fox; (d) red panda

Задача

Детектировать объекты на видео (Detect) и отслеживать траектории их перемещения (Track).

Стандартный подход: Detect $\rightarrow$ Track (frame-level).

I. Detect

- Region Proposal (R-CNN[1] $\rightarrow$ Fast R-CNN[2] $\rightarrow$ Faster R-CNN[3], R-FCN[4])
- One-step (SSD[5], YOLO[6])

[1] R. Girshick. Rich feature hierarchies for accurate object detection and semantic segmentation.
[2] R. Girshick. Fast R-CNN.
[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks.
[4] J. Dai, Y. Li, K. He, J. Sun. R-FCN: Object Detection via Region-based Fully Convolutional Networks.
[5] W. Liu, D. Anguelov, D. Erhan, Ch. Szegedy, S. Reed, Ch.-Y. Fu, A. C. Berg. SSD: Single Shot MultiBox Detector.
[6] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection.

II. Track

- Regression-based (GOTURN[6], FCNT[7])
- Correlation-based ([8], SiamFC[9])

[6] D. Held, S. Thrun, and S. Savarese. Learning to track at 100 FPS with deep regression networks.

[7] L. Wang, W. Ouyang, X. Wang, and H. Lu. Visual tracking with fully convolutional networks.

[8] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking.

[9] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. Fully-convolutional siamese networks for object tracking.

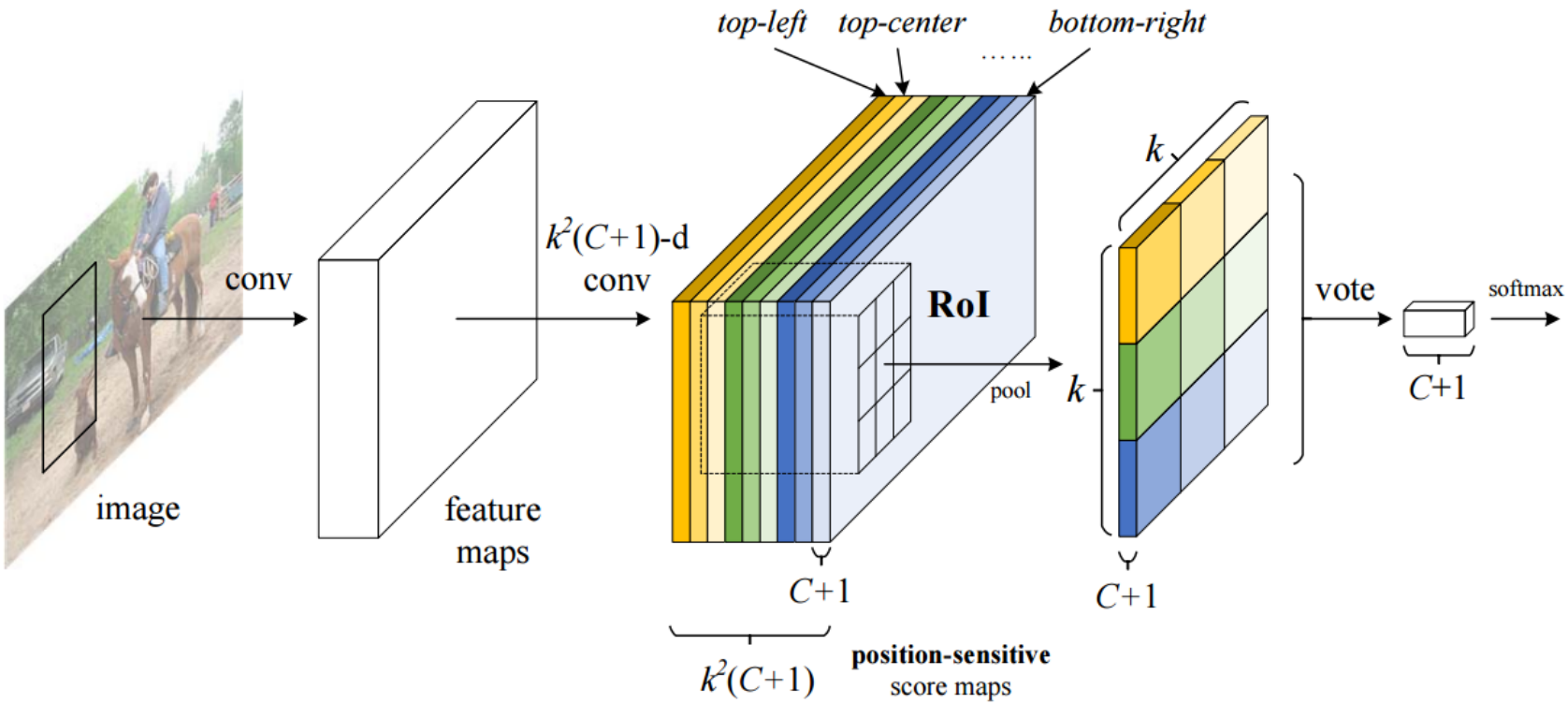
Задача

Детектировать объекты на видео (Detect) и отслеживать траектории их перемещения (Track).

Стандартный способ решения: Detect $\rightarrow$ Track (покадрово).

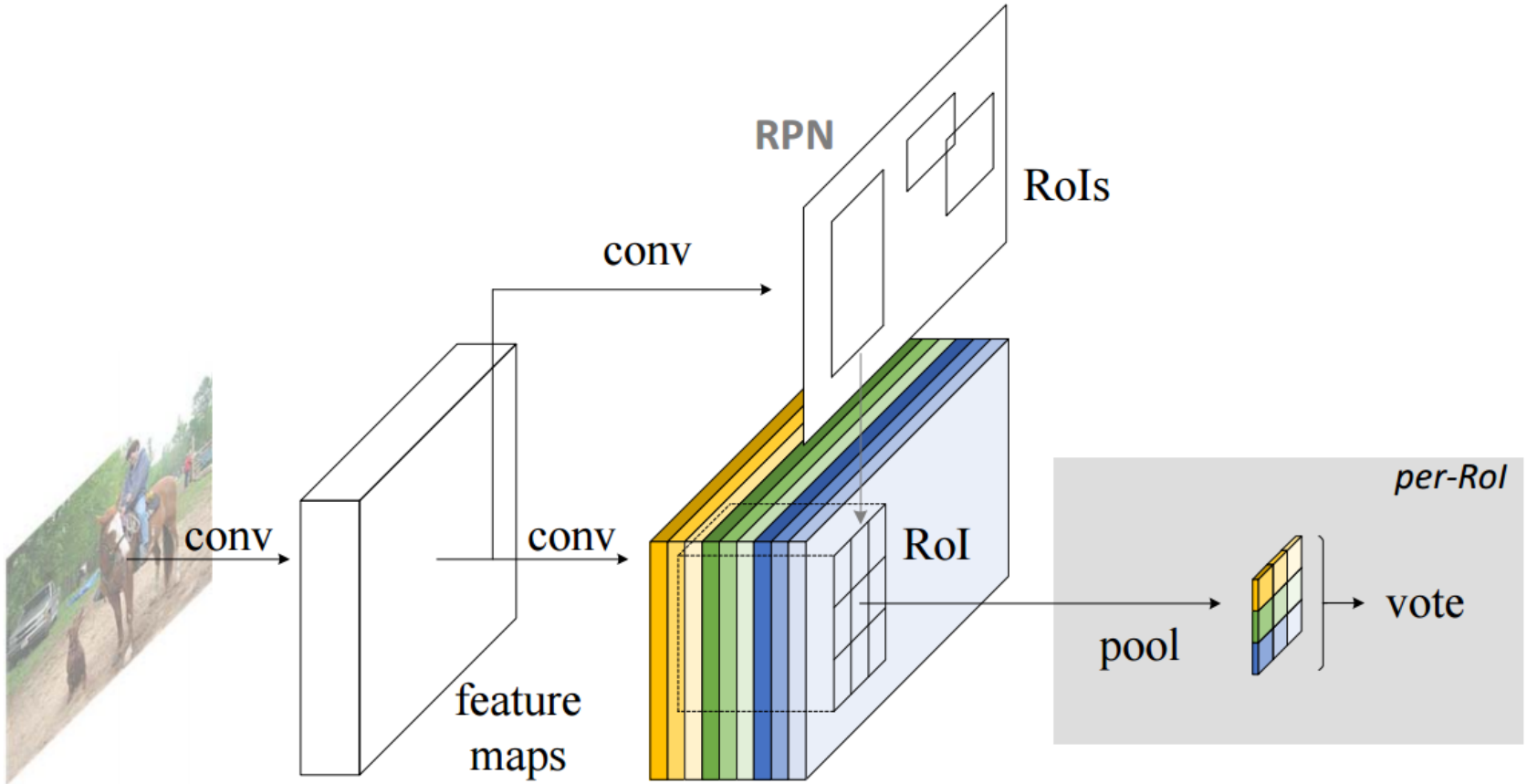
Предложение: Detect & Track.

R-FCN



C – кол-во классов; d – размерность признакового пр-ва

R-FCN

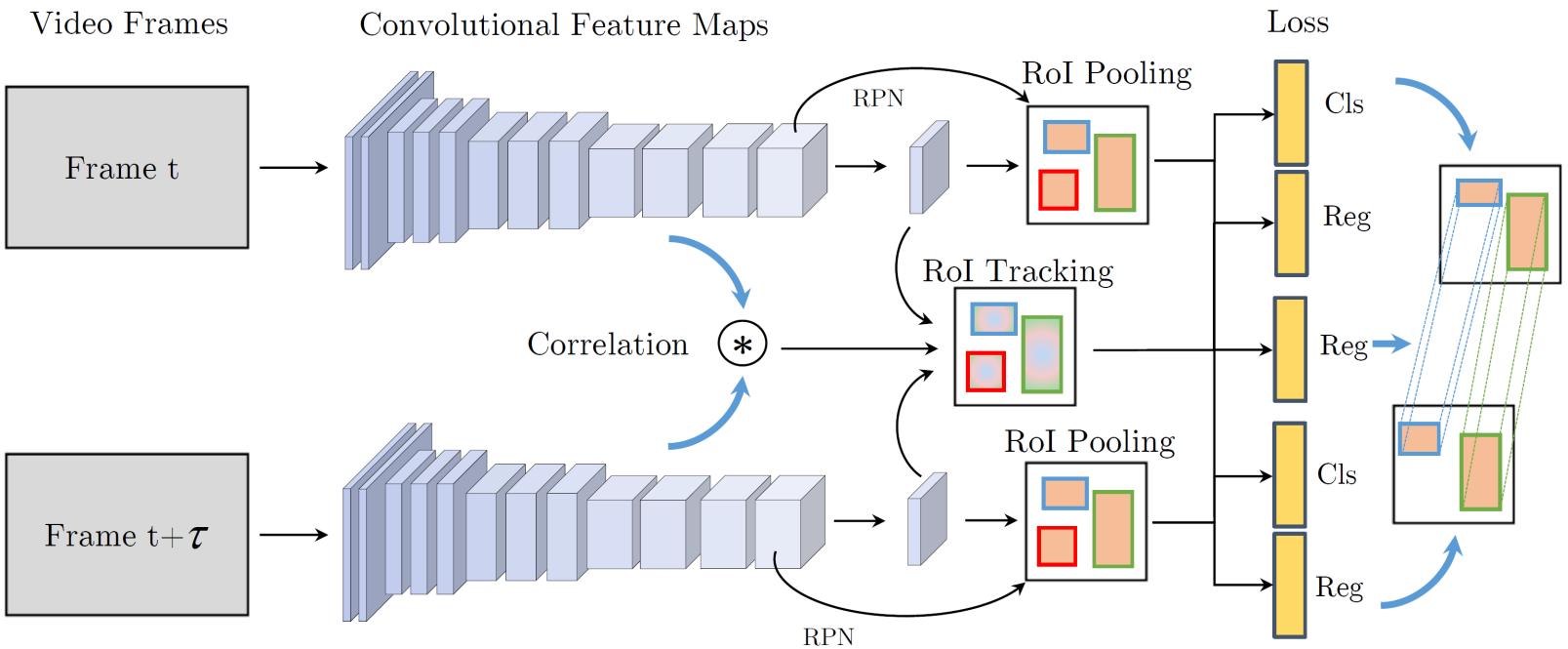


RPN[10] – полносверточная н/с на базе VGG16, сверточная н/с – ResNet101 [11]

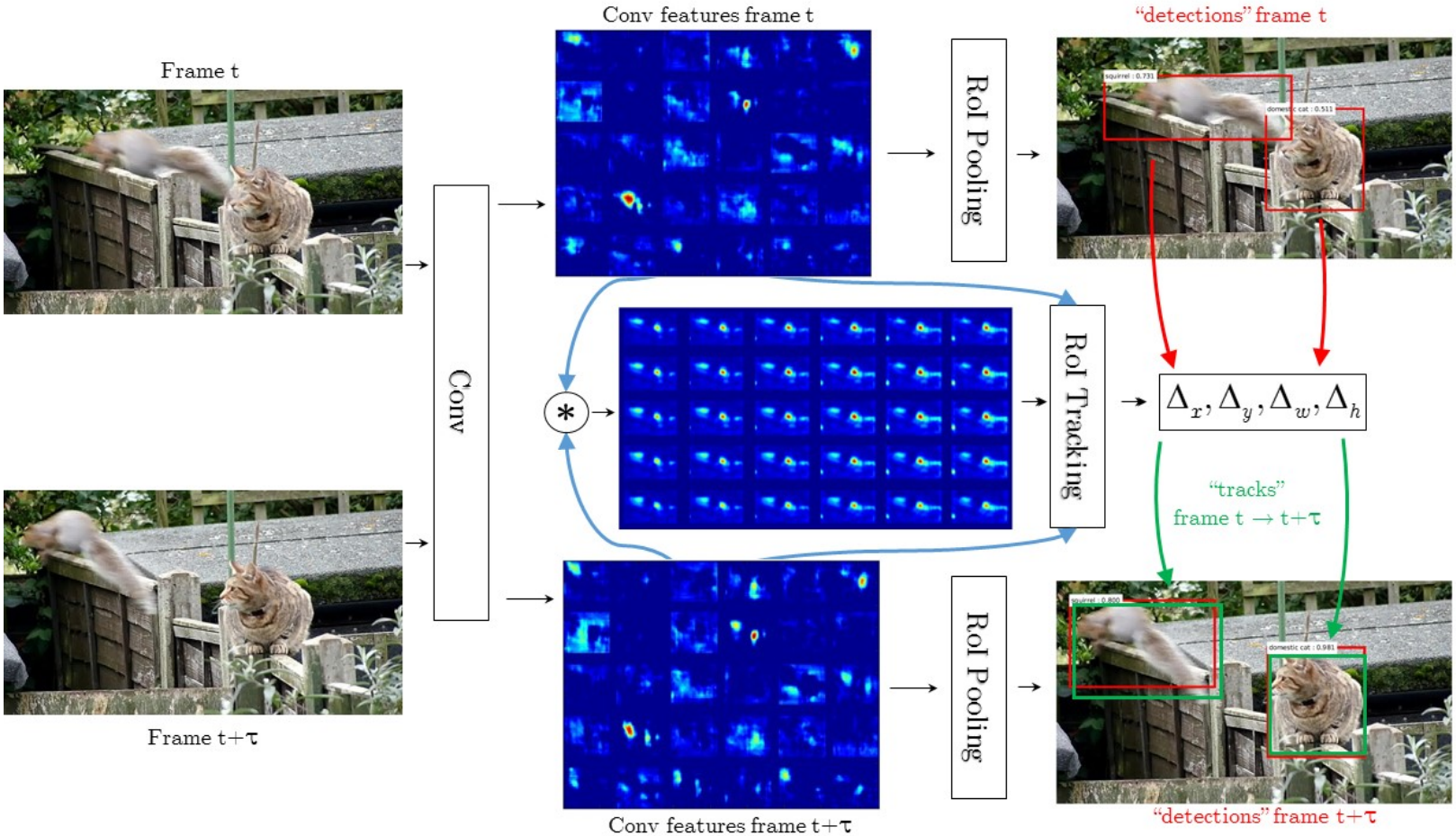
[10] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition.

D&T: Архитектура



D&T: принцип работы



RoI transformation: $\Delta^{t+\tau} = (\Delta_x^{t+\tau}, \Delta_y^{t+\tau}, \Delta_h^{t+\tau}, \Delta_w^{t+\tau})$

Correlation Layer

«Локальная» корреляция:

$$x_{corr}^{t,t+\tau}(i,j,p,q) = \left\langle x_l^t(i,j), x_l^{t+\tau}(i+p, j+q) \right\rangle$$

$-d \leqslant p, q \leqslant d$, d - максимальное смещение.

Расположение объекта – $\max p, q$ по карте корреляций.

Multitask Objective

$\{p_i\}_{i=1}^N$ – предсказания после softmax

$\{b_i\}_{i=1}^N$ – bounding boxes

$\{\Delta_i\}_{i=1}^N$ – траектории RoI

$*$ – истинные значения

$[c^* = 0]$ – фон

N_{fg} – #RoI, отн. к какому-либо классу $= \sum_{i=1}^{N} [c_i^* > 0]$

N_{tra} – #RoI, соотносящихся на последовательных кадрах

Multitask Objective

$\{p_i\}_{i=1}^N$ – предсказания после softmax | $*$ – истинные значения

$\{b_i\}_{i=1}^N$ – bounding boxes | $[c^* = 0]$ – фон

$\{\Delta_i\}_{i=1}^N$ – траектории RoI

N_{fg} – #RoI, отн. к какому-либо классу $= \sum_{i=1}^{N} [c_i^* > 0]$

N_{tra} – #RoI, соотносящихся на последовательных кадрах

$$L(\{p_i\}, \{b_i\}, \{\Delta_i\}) = \frac{1}{N} \sum_{i=1}^{N} L_{cls}(p_{i,c^*}) +$$

$$+ \frac{1}{N_{fg}} \sum_{i=1}^{N} [c_i^* > 0] L_{reg}(b_i, b_i^*) +$$

$$+ \frac{1}{N_{tra}} \sum_{i=1}^{N_{tra}} L_{tra}(\Delta_i^{t+\tau}, \Delta_i^{*,t+\tau})$$

Multitask Objective

$\{p_i\}_{i=1}^N$ – предсказания после softmax | $*$ – истинные значения

$\{b_i\}_{i=1}^N$ – bounding boxes | $[c^* = 0]$ – фон

$\{\Delta_i\}_{i=1}^N$ – траектории RoI

N_{fg} – #RoI, отн. к какому-либо классу $= \sum_{i=1}^{N} [c_i^* > 0]$

N_{tra} – #RoI, соотносящихся на последовательных кадрах

$$L(\{p_i\}, \{b_i\}, \{\Delta_i\}) = \frac{1}{N} \sum_{i=1}^{N} L_{cls}(p_{i,c^*}) + \quad \text{классификация}$$

$$+ \frac{1}{N_{fg}} \sum_{i=1}^{N} [c_i^* > 0] L_{reg}(b_i, b_i^*) + \quad \text{регрессия (RoI)}$$

$$+ \frac{1}{N_{tra}} \sum_{i=1}^{N_{tra}} L_{tra}(\Delta_i^{t+\tau}, \Delta_i^{*,t+\tau}) \quad \text{траектории}$$

Multitask Objective

$\{p_i\}_{i=1}^N$ – предсказания после softmax | $*$ – истинные значения
$\{b_i\}_{i=1}^N$ – bounding boxes | $[c^* = 0]$ – фон
$\{\Delta_i\}_{i=1}^N$ – траектории RoI

N_{fg} – #RoI, отн. к какому-либо классу $= \sum_{i=1}^{N} [c_i^* > 0]$

N_{tra} – #RoI, соотносящихся на последовательных кадрах

$$L(\{p_i\}, \{b_i\}, \{\Delta_i\}) = \frac{1}{N} \sum_{i=1}^{N} L_{cls}(p_{i,c^*}) + \quad \Big| \quad L_{cls}(\cdot) = -\log(\cdot)$$

$$+ \frac{1}{N_{fg}} \sum_{i=1}^{N} [c_i^* > 0] L_{reg}(b_i, b_i^*) + \quad \Big| \quad L_{reg}(\cdot) = smooth_{L1}$$

$$+ \frac{1}{N_{tra}} \sum_{i=1}^{N_{tra}} L_{tra}(\Delta_i^{t+\tau}, \Delta_i^{*,t+\tau}) \quad \Big| \quad L_{tra}(\cdot) = smooth_{L1}$$

Multitask Objective

Smooth L1

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5 & otherwise \end{cases}$$

Multitask Objective

$\{p_i\}_{i=1}^N$ – предсказания после softmax | $*$ – истинные значения
$\{b_i\}_{i=1}^N$ – bounding boxes | $[c^* = 0]$ – фон
$\{\Delta_i\}_{i=1}^N$ – траектории RoI

N_{fg} – #RoI, отн. к какому-либо классу $= \sum_{i=1}^{N} [c_i^* > 0]$

N_{tra} – #RoI, соотносящихся на последовательных кадрах

$$L(\{p_i\}, \{b_i\}, \{\Delta_i\}) = \frac{1}{N} \sum_{i=1}^{N} -\log(p_{i,c^*}) +$$

$$+ \frac{1}{N_{fg}} \sum_{i=1}^{N} [c_i^* > 0] smooth_{L1}(b_i, b_i^*) +$$

$$+ \frac{1}{N_{tra}} \sum_{i=1}^{N_{tra}} smooth_{L1}(\Delta_i^{t+\tau}, \Delta_i^{*,t+\tau})$$

Multitask Objective

$\{p_i\}_{i=1}^N$ – предсказания после softmax
$\{b_i\}_{i=1}^N$ – bounding boxes
$\{\Delta_i\}_{i=1}^N$ – траектории RoI

$*$ – истинные значения
$[c^* = 0]$ – фон

N_{fg} – #RoI, отн. к какому-либо классу $= \sum_{i=1}^{N} [c_i^* > 0]$

N_{tra} – #RoI, соотносящихся на последовательных кадрах

RoI совпадает с истинным, если:

- c^*, b^*– если $IoU \geqslant 0.5$
- $\Delta^{*,t+\tau}$ – если соотнесены объекты, присутствующему на обоих кадрах.

Class-wise linking score

Object tube – последовательность детекций.

В момент t : $D_i^{t,c} = \left\{x_i^t, y_i^t, h_i^t, w_i^t, p_{i,c}^*\right\}$

Tracklet – изменение bounding boxes во времени.

$t \to t+\tau$: $T_i^{t,t+\tau} = \{x_i^t, y_i^t, h_i^t, w_i^t,$
$x_i^t + \Delta_x^{t+\tau}, y_i^t + \Delta_y^{t+\tau}, h_i^t + \Delta_h^{t+\tau}, w_i^{t+\tau} + \Delta_w^{t+\tau}\}$

Задача: связать object tubes и tracklets.

Class-wise linking score

Object tube – последовательность детекций.

В момент t : $D_i^{t,c} = \left\{x_i^t, y_i^t, h_i^t, w_i^t, p_{i,c}^*\right\}$

Tracklet – изменение bounding boxes во времени.

$t \to t+\tau$: $T_i^{t,t+\tau} = \{x_i^t, y_i^t, h_i^t, w_i^t,$
$x_i^t + \Delta_x^{t+\tau}, y_i^t + \Delta_y^{t+\tau}, h_i^t + \Delta_h^{t+\tau}, w_i^{t+\tau} + \Delta_w^{t+\tau}\}$

Class-wise linking score:

$s_c(D_{i,c}^t, D_{j,c}^{t+\tau}, T^{t,t+\tau}) = p_{i,c}^t + p_{j,c}^{t+\tau} + \left[D_i^t, D_j^{t+\tau} \in T^{t,t+\tau}\right]$

Оптимальный путь

Class-wise linking score:

$$s_c(D_{i,c}^t, D_{j,c}^{t+\tau}, T^{t,t+\tau}) = p_{i,c}^t + p_{j,c}^{t+\tau} + \left[D_i^t, D_j^{t+\tau} \in T^{t,t+\tau}\right]$$

Оптимальный путь:

$$\overline{D}_c^* = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}-\tau} s_c(D^t, D^{t+\tau}, T^{t,t+\tau})$$

[Алгоритм Витерби – алгоритм поиска наиболее подходящего списка состояний (т.н. *путь Витерби*), который в контексте цепей Маркова получает наиболее вероятную последовательность произошедших событий]

Эксперименты

ImageNet object detection from video (VID) dataset:

- 30 классов
- 3862(обуч.) / 555(вал.) видео
- Bounding Boxes с аннотациями и track ID.

Метрика – mAP.

Эксперименты

Methods	airplane	antelope	bear	bicycle	bird	bus	car	cattle	dog	d. cat	elephant	fox	g. panda	hamster	horse	lion
TCN [18]	72.7	75.5	42.2	39.5	25.0	64.1	36.3	51.1	24.4	48.6	65.6	73.9	61.7	82.4	30.8	34.4
TPN+LSTM [16]	84.6	78.1	72.0	67.2	68.0	80.1	54.7	61.2	61.6	78.9	71.6	83.2	78.1	91.5	66.8	21.6
Winner ILSVRC'15 [17]	83.7	85.7	84.4	74.5	73.8	75.7	57.1	58.7	72.3	69.2	80.2	83.4	80.5	93.1	84.2	67.8
D (R-FCN)	87.4	79.4	84.5	67.0	72.1	84.6	54.6	72.9	70.9	77.3	76.7	89.7	77.6	88.5	74.8	57.9
D (& T loss)	89.4	80.4	83.8	70.0	71.8	82.6	56.8	71.0	71.8	76.6	79.3	89.9	83.3	91.9	76.8	57.3
D&T ($\tau = 1$)	90.2	82.3	87.9	70.1	73.2	87.7	57.0	80.6	77.3	82.6	83.0	97.8	85.8	96.6	82.1	66.7
D&T ($\tau = 10$)	89.1	79.8	87.5	68.8	72.9	86.1	55.7	78.6	76.4	83.4	82.9	97.0	85.0	96.0	82.2	66.0

Methods	lizard	monkey	motorcycle	rabbit	red panda	sheep	snake	squirrel	tiger	train	turtle	watercraft	whale	zebra	mAP (%)
TCN [18]	54.2	1.6	61.0	36.6	19.7	55.0	38.9	2.6	42.8	54.6	66.1	69.2	26.5	68.6	47.5
TPN+LSTM [16]	74.4	36.6	76.3	51.4	70.6	64.2	61.2	42.3	84.8	78.1	77.2	61.5	66.9	88.5	68.4
Winner ILSVRC'15 [17]	80.3	54.8	80.6	63.7	85.7	60.5	72.9	52.7	89.7	81.3	73.7	69.5	33.5	90.2	73.8
Winner ILSVRC'16 [39]	(single model performance)														76.2
D (R-FCN)	76.8	50.1	80.2	61.3	79.5	51.9	69.0	57.4	90.2	83.3	81.4	68.7	68.4	90.9	74.2
D (& T loss)	79.0	54.1	80.3	65.3	85.3	56.9	74.1	59.9	91.3	84.9	81.9	68.3	68.9	90.9	75.8
D&T ($\tau = 1$)	83.4	57.6	86.7	74.2	91.6	59.7	76.4	68.4	92.6	86.1	84.3	69.7	66.3	95.2	**79.8**
D&T ($\tau = 10$)	83.1	57.9	79.8	72.7	90.0	59.4	75.6	65.4	90.5	85.6	83.3	68.3	66.5	93.2	78.6

D&T($\tau = 1$) – mAP 79.8%, winner ILSVRC'16 – mAP 76.2%

[16] K. Kang et al. Object detection in videos with tubelet proposal networks.
[17] K. Kang et al. T-CNN: tubelets with convolutional neural networks for object detection from videos.
[18] K. Kang et al. Object detection from video tubelets with convolutional neural networks.
[39] J. Yang et al. ILSVRC2016 object detection from video.

Заключение

- D&T – фреймворк на основе сверточных нейросетей для детекции и отслеживания объектов на видео
- извлечение признаков с помощью корреляции – по ним можно определить соотношение между объктами на разных кадрах
- детекция на уровне кадра, основанная на tracklet $\rightarrow$ высокое качество на уровне видео.

Заключение

D&T:

- простота – за счет одновременного решения двух подзадач: детекции и отслеживания
- эффективность: достигает результатов, сравнимых с результатами победителя конкурса ImageNet 2016 года

Эксперименты

https://www.robots.ox.ac.uk/ vgg/research/detect-track/

https://www.robots.ox.ac.uk/ vgg/research/detect-track/videos/DT_detections.mp4

Список литературы

- Ch. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to Track and Track to Detect. In *ICCV*, 2017. [1710.03958]
- J. Dai, Y. Li, K. He, J. Sun. R-FCN: Object Detection via Region-based Fully Convolutional Networks [1605.06409]
- J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. [1411.4038]

Список литературы (дополнительно)

- S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. [1506.01497]
- R. Girshick. Fast R-CNN. In *ICCV*, 2015. [1504.08083]
- R. Girshick. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. [1311.2524]
- W. Liu, D. Anguelov, D. Erhan, Ch. Szegedy, S. Reed, Ch.-Y. Fu, A. C. Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016. [1512.02325]
- J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection [1506.02640]

Список литературы (дополнительно)

- L. Wang, W. Ouyang, X. Wang, and H. Lu. Visual tracking with fully convolutional networks. In *ICCV*, 2015.
- D. Held, S. Thrun, and S. Savarese. Learning to track at 100 FPS with deep regression networks. In *ECCV*, 2016.
- C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015.
- L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016. [1606.09549]

Список литературы (дополнительно)

- K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang. Object detection in videos with tubelet proposal networks. In *CVPR*, 2017. [1702.06355]
- K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, and W. Ouyang. T-CNN: tubelets with convolutional neural networks for object detection from videos, 2016. [1604.02532]
- K. Kang, W. Ouyang, H. Li, and X. Wang. Object detection from video tubelets with convolutional neural networks. In *CVPR*, 2016. [1604.04053]
- J. Yang, H. Shuai, Z. Yu, R. Fan, Q. Ma, Q. Liu, and J. Deng. ILSVRC2016 object detection from video.