Single shot neural detectors

YOLO and SSD models

Adelina Bakieva
Higher School of Economics
Faculty of Computer Science
Moscow, Russia
aebakieva@edu.hse.ru

***Abstract*— You Only Look Once single neural network model for object detection was presented by J. Redmon et al in 2016. It refers to detection as a regression problem and solves it fast, comparing to existing state-of-art methods, e.g. Fast R-CNN, but making a lot of localization mistakes.**

Soon after YOLO, Single Shot MultiBox Detector paper was published by W. Liu et al. This model is based on VGG-16, pre-trained on ImageNet for image classification, and it also shows high speed in frames per second but has higher accuracy than YOLO.

Keywords—real-time object detection; convolutional neural network.

INTRODUCTION

Region-based Convolutional Neural Network was a dramatic improvement, which led to the prevalence of region proposal object detection methods over ones, based on a sliding window method.

However, in 2016 even state-of-arts models for object detection were too slow and couldn't reach the speed which would be considered as Real-Time or have low accuracy. For example, Faster R-CNN with mAP 73.2 had 7FPS, 30HzDPM and 100HzDPM while having 30FPS and 100FPS respectively had an accuracy of 26.1 mAP and 16.0 mAP, which definitely needed to be improved in order to maintain good performance in areas demanding fast and accurate object detection, such as systems for self-driving cars.

You Only Look Once and Single Shot MultiBox Detector models gave this improvement with 45FPS with 63.4 mAP and 59FPS with 74.3 mAP respectively. These methods skip the proposal step of R-CNN and predict bounding boxes for multiple classes directly, which fasten them.

YOU ONLY LOOK ONCE

Model.

YOLO [1] divides an input image into a grid ($S \times S$) and each grid cell predicts B bounding boxes and confidence for them (it reflect how accurate the box is and how confident the model is that box contains an object). So bounding box predictions contains five parameters: x, y – the center of the box, w, h - its width and height, and confidence score. Then YOLO [1] predicts C conditional class probabilities.

Because of YOLO detects only one object per cell it has difficulties with small objects of one class, which placed close to each other, such as flocks of birds.

Network design.

YOLO [1] network has 24 convolutional layers followed by 2 fully connected layers.

Training.

Convolution layers (first 20 of them, followed by an average-pooling layer and one fully connected layer) were pretrained on the ImageNet 1000-class competition dataset. Then J. Redmon et al added 4 convolutional layers and 2 fully connected layers with random weights and trained network on Pascal VOC 2007 and 2012 data sets.

Generalization.

YOLO [1] learns general representations of the objects and outperforms other detection methods when generalizing from natural images to artworks.

YOLO was compared to state-of-art detectors, such as R-CNN and DPM. At first, all models were trained on VOC 2007 dataset (APs on testing in this set: YOLO 59.2, R-CNN 54.2, DPM 43.2), then they were tested in finding faces on Picasso's works (APs: YOLO 53.3, R-CNN 10.4, DPM 37.8), where YOLO outperformed other models

SINGLE SHOT MULTIBOX DETECTOR

Model.

SSD [2] uses a convolutional network that creates bounding boxes of 6 different ratios and scores the probability of class presence in each box.

Network design.

Liu et al [2] use VGG16 model, which is high-quality image classification structure, which for SSD purposes was truncated before any classification layers, as a base model with added auxiliary structure, such as decreasing in size convolutional feature layers at the end of the base model, that allow predictions in multiple scales.

Training.

VGG16, used as a base model, were pretrained on ILSVRC CLS-LOC dataset. All dropout layers were removed, as well as a fc8 layer.

Instead of using all negatives examples, they are sorted for each default box (the highest confidence loss to the lowest). Only top ones are used, which allows getting a ratio of 3/1 between negative and positive examples.

Also, data augmentation and Atrous convolution are used for training.

COMPARISON

Model.

The first thing that differs among models is that SSD does not predict probabilities of objects like YOLO [1] but counts only probabilities of classes.

The second one is that SSD [2] uses the multi-scale feature maps for detection, but YOLO only uses one scale for detection.

Also, YOLO [1] model uses the fully connected layer for detection, but SSD [2] use the convolution filter for detection. (Fig.1)

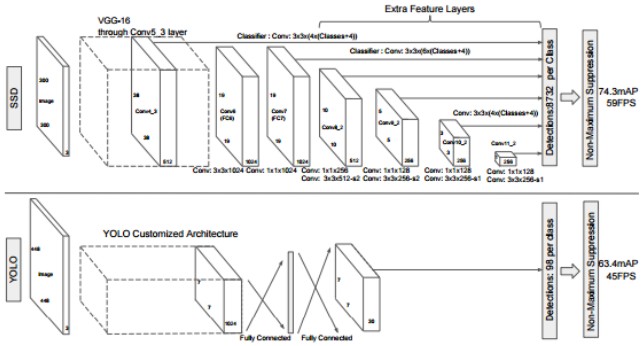

Fig.1:A comparison between SSD and YOLO architectures

Limitations.

As was mentioned in the description of the YOLO model it has difficulties in detecting small objects, especially when they are grouped. But SSD also has typically poor performance on small objects [3].

YOLO model compared with SSD [2], which has 6 different ratios for default boxes, struggles when working with unusual aspect ratios.

It also has frequent localization errors because of the described absence of multi-scale feature maps.

Testing comparison.

The first version of YOLO presented in 2016 has 45FPS and 63.4% mAP, while SSD [2] has 59FPS and 74.3% mAP on VOC 2007 test.

Liu et al. [2] testing of SSD300 (which works with 300×300 images), SSD500 (500×500) and YOLO (448×448) on VOC 2007, 2012 and COCO data sets are displayed on fig.2. It shows, that SSD outperforms the first version of YOLO in accuracy.

CONCLUSION

Single Shot MultiBox Detector and You Only Look Once both are big improvements in object detection. They are relatively fast and accurate because these models are single shot.

Although SSD outperforms YOLO both in speed and accuracy, new versions of YOLO [4, 5] outperform it in speed, which demonstrates that object detection solution are becoming better.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshink, A. Farhadi. You Only Look Once: Unified, Real-Time, Object Detection. arXiv perprint arXiv: 1506.02640, 2016.

[2] W. Liu, D. Anguelov, D. Erhan, Ch. Szegedy, S. Reed, Ch-Y. Fu, A. Berg. SSD: Single Shot MultiBox Detector. arXiv perprint arXiv: 1512.02325, 2016.

[3] J. Huang, V. Rathod, Ch. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer Z. Wojna, Y. Song, S. Guadarrama, K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. arXiv preprint arXiv: 1611.10012v3, 2017.

[4] J. Redmon, A. Farhadi. YOLO9000: Better, Faster, Stronger. arXiv preprint arXiv: 1612.08242, 2016.

Method PASCAL VOC 2012 test detection results

Method	PASCAL VOC 2012 test detection results																					
		mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast[6]	07++12	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0	55.0	87.5	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
Faster[2]	07++12	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
Faster[2]	07++12+COCO	75.9	87.4	83.6	76.8	62.9	59.6	81.9	82.0	91.3	54.9	82.6	59.0	89.0	85.5	84.7	84.1	52.2	78.9	65.5	85.4	70.2
YOLO[5]	07++12	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
SSD300	07++12	72.4	85.6	80.1	70.5	57.6	46.2	79.4	76.1	89.2	53.0	77.0	60.8	87.0	83.1	82.3	79.4	45.9	75.9	69.5	81.9	67.5
SSD300	07++12+COCO	77.5	90.2	83.3	76.3	63.0	53.6	83.8	82.8	92.0	59.7	82.7	63.5	89.3	87.6	85.9	84.3	52.6	82.5	**74.1**	**88.4**	74.2
SSD512	07++12	74.9	87.4	82.3	75.8	59.0	52.6	81.7	81.5	90.0	55.4	79.0	59.8	88.4	84.3	84.7	83.3	50.2	78.0	66.3	86.3	72.0
SSD512	07++12+COCO	**80.0**	**90.7**	**86.8**	**80.5**	**67.8**	**60.8**	**86.3**	**85.5**	**93.5**	**63.2**	**85.7**	**64.4**	**90.9**	**89.0**	**88.9**	**86.8**	**57.2**	**85.1**	72.8	**88.4**	**75.9**

[5] J. Redmon, A. Farhadi. YOLOv3: An Incremental Improvement. arXiv preprint arXiv: 1804.02767, 2018.

1066 words