Study Note for Mask R-CNN

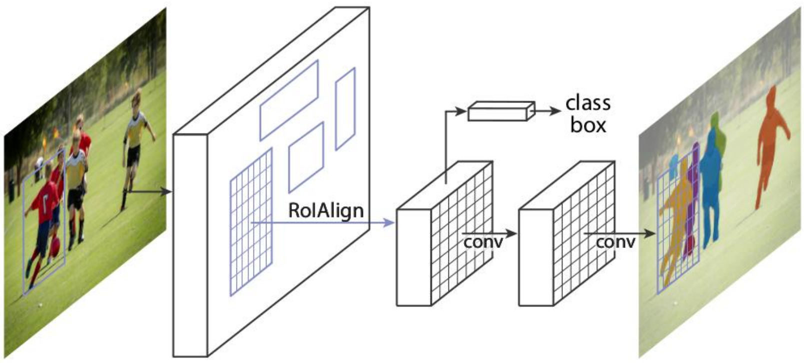
Qinghong Lin
2018/10

Abstract

• Mask R-CNN **add a branch for predicting an object mask in paralled** *(with the existing branch for bounding box recognition)* and is **simple to train and adds only a small overhead** *(to Faster R-CNN. running at 5 fps).*

1.Introduction

• Goal is to develop a comparably enabling framework for *instance segmentation.*
• Mask R-CNN extends Faster R-CNN by *adding a branch for predicting segmentation masks on each RoI*, ***in parallel** with* the existing branch for classification and bounding box regression.



• The mask branch is a small **FCN** applied to each RoI, predicting a segmentation mask in a pixel-to-pixel manner.

• *RoIPool* performs coarse spatial quantization for feature extraction, to fix the misalignment, ***RoIAlign*** preserves exact spatial locations.
• RoIAlign impact
-Improves mask accuracy by 10% to 50%, showing bigger gains under stricter localization metrics.
-It essential to **decouple mask and class prediction***:*
(*we predict a binary mask for each class independently without competition among classes, and rely on the network's RoI classification branch to predict the category*).

2.Related Work

R-CNN

• R-CNN approach **to bounding-box object detection** is to attend to a manageable number of candidate object regions and evaluate convolutional networks independently on each **RoI**.
• Fast R-CNN allow RoIs on feature maps using RoIPool, leading to fast speed and better accuracy.
• Faster R-CNN learning the **attention mechanism** with a **RPN**.

Instance Segmentation

- many instance segmentation approaches are based on ***segment proposals***.
- segmentation precedes recognition is slow and less accurate.

- ***FCIS*** predict a set of position-sensitive output channels FC, errors on overlapping instances and creates spurious edges.
- In contrast to *segmentation-first* strategy of these methods, Mask R-CNN is based on an *instance-first* strategy.

3.Mask R-CNN

- Faster R-CNN has **two outputs** for each candidate object, a class label and a bounding-box offset; to this we **add a third branch** that outputs the object mask.
- But the additional mask output is requiring extraction of much finer spatial layout of an object.

Faster R-CNN

- Faster R-CNN two stages

-The first stage is **RPN**.

-The second stage is extracts features using **RoIPool** from each candidate box and performs classification and bounding-box regression.

Mask R-CNN

- Adopts the same two-stage procedure

-The first stage is **RPN**.

-The second stage, *in parallel to* predicting the class and box offset, also outputs a **binary mask** for each RoI. (classification depends on mask predictions)

- multi-task loss on each sampled RoI as $L = Lcls + Lbox + Lmask$

-the mask branch has a Km^2-dim output for each RoI.(resolution $m \times m$ for K classes)

-Apply a per-pixel sigmoid and define *Lmask* as the average binary cross-entropy loss.

- Allow the network to generate masks for every class without competition among classes;

-FCNs uses a *per-pixel* softmax and a *multinomial* cross-entropy loss, masks across **classes compete**.

-In our case, with a *per-pixel sigmoid* and a *binary loss* do not.

Mask Representation

- A mask encodes an input object's *spatial* layout

-class labels or box offests are inevitably collapsed into short output vectors by *fc* layers

-**Extracting the spatial structure of masks** can be addressed naturally by the pixel-to-pixel correspondence provided by convolutions.

- Predict an $m \times m$ mask from each RoI using an FCN.

*(This allows each layer in the mask branch to **maintain the explicit m $\times$ m object spatial layout***

without collapsing it into a vector representation lacks spatial dimensions.)

- **To be well aligned** to faithfully preserve the explicit per-pixel spatial correspondence.

RoIAlign

- RoIPool extracting a small feature map from each RoI
  -first **quantizes** a floating-number RoI to the **discrete granularity of the feature map.**
  -then **subdivided** into spatial bins.
  -**finally** feature values covered by each bin are aggregated.

- RoIAlign layer removes the **harsh quantization** of RoIPool, ***aligning*** the extracted features with the input.
  -Avoid any quantization of the RoI boundaries or bins.
  -Use **bilinear interpolation** to compute the exact values of the input features at four regularly sampled locations in each RoI bin, and aggregate the result.

Network Architecture

- Instantiate Mask R-CNN with multiple architectures.

(i) **the convolutional *backbone* architecture** used for feature extraction over an entire image

(ii) **the network head** for bounding-box recognition and mask prediction.

- backbone:ResNet and ResNeXt of depth 50 or 101 layer, FPN
- network head,we add a FC mask prediction branch.

3.1.Implementation Details

Training

- *Lmask* is defined only on positive RoI
- Image-centric training
- RPN and Mask R-CNN have the same backbones and so they are shareable.

Inference

- predict K masks per RoI, but we only use the k-th mask.(k is the predicted class by the classification branch).

4.Experiments:Instance Segmentation

4.1.Main Results

- Mask R-CNN with ResNet-101-FPN backbone outperforms FCIS+++.

4.2.Ablation Experiments

net-depth-features	AP	AP_{50}	AP_{75}
ResNet-50-C4	30.3	51.2	31.5
ResNet-101-C4	32.7	54.2	34.3
ResNet-50-FPN	33.6	55.2	35.3
ResNet-101-FPN	35.4	57.3	37.5
ResNeXt-101-FPN	**36.7**	**59.5**	**38.9**

(a) **Backbone Architecture**: Better backbones bring expected gains: deeper networks do better, FPN outperforms C4 features, and ResNeXt improves on ResNet.

	AP	AP_{50}	AP_{75}
softmax	24.8	44.1	25.1
sigmoid	**30.3**	**51.2**	**31.5**
	+5.5	*+7.1*	*+6.4*

(b) **Multinomial *vs.* Independent Masks** (ResNet-50-C4): *Decoupling* via per-class binary masks (sigmoid) gives large gains over multinomial masks (softmax).

	align?	bilinear?	agg.	AP	AP_{50}	AP_{75}
RoIPool [12]			max	26.9	48.8	26.4
RoIWarp [10]		✓	max	27.2	49.2	27.1
		✓	ave	27.1	48.9	27.1
RoIAlign	✓	✓	max	**30.2**	**51.0**	**31.8**
	✓	✓	ave	**30.3**	**51.2**	**31.5**

(c) **RoIAlign** (ResNet-50-C4): Mask results with various RoI layers. Our RoIAlign layer improves AP by ~3 points and AP_{75} by ~5 points. Using proper alignment is the only factor that contributes to the large gap between RoI layers.

	AP	AP_{50}	AP_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
RoIPool	23.6	46.5	21.6	28.2	52.7	26.9
RoIAlign	**30.9**	**51.8**	**32.1**	**34.0**	**55.3**	**36.4**
	+7.3	*+ 5.3*	*+10.5*	*+5.8*	*+2.6*	*+9.5*

(d) **RoIAlign** (ResNet-50-**C5**, *stride 32*): Mask-level and box-level AP using *large-stride* features. Misalignments are more severe than with stride-16 features (Table 2c), resulting in massive accuracy gaps.

	mask branch	AP	AP_{50}	AP_{75}
MLP	fc: 1024→1024→80·28^2	31.5	53.7	32.8
MLP	fc: 1024→1024→1024→80·28^2	31.5	54.0	32.6
FCN	conv: 256→256→256→256→256→80	**33.6**	**55.2**	**35.3**

(e) **Mask Branch** (ResNet-50-FPN): Fully convolutional networks (FCN) *vs.* multi-layer perceptrons (MLP, fully-connected) for mask prediction. FCNs improve results as they take advantage of explicitly encoding spatial layout.

Architecture (a)

- benefits from deeper networks

Multinomial vs.Independent Masks (b)

- *sigmoid and a binary loss* better than *per-pixel softmax and a multinomial loss*.

Class-Specific vs. Class-Agnostic Masks

- Our approach decouples classification and segmentation.

RoIAlign (c)

- RoIAlign improves AP over RoIPool.
- RoIWarp quantizes the RoI, losing alignment with the input.
- ResNet-50-C5 back-bone imroves mask AP. (d)

Mask Branch (e)

- FCNs imroves mask AP over multi-layer perceptrons(MLP).

4.3.Bounding Box Detection Results

	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Faster R-CNN+++ [19]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [27]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [21]	Inception-ResNet-v2 [37]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [36]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6	40.3	19.8	40.2	48.8
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**	**43.4**	**22.1**	**43.2**	51.2

- Mask R-CNN object detection 39.8% while instance segmentation 37.1%

4.4.Timing

Inference

Training

5.Mask R-CNN for Human Pose Estimation

- easy to extend to human pose estimation.

Implementation Details

Main Results and Ablations