Study Notes for R-CNN

QingHong Lin
2018/10

Abstract

best-performing methods before CNN

• complex ensemble models:combine multiple low-level image features with high-level context.

two key insights

• apply high-capacity **CNNs** to bottom-up region proposals to localize and segment objects.

• when labeled **trainging data is scarce**, supervised pre-training+domain-specific fine-tuning.

1.Introduction

history

SIFT&HOG

• before R-CNN,mainly based on SIFT and HOG,progress was **slow** by building ensemble systems and employing minor variants of successful methods.

• blockwise orientation histogram,but visual recognition might be hierarchical,multi-stage processes.

neocognitron

• proposed a biologically inspired hierarchical and shift-invariant model.

• lack supervised training algorithm.

• LeCun later provided **SGD**, backpropagation to train CNN.(extend the neocognitron)

AlexNet

• CNNs fell out of fashion with with the rise of **SVM**.

• Krizhevsky's **AlexNet** in 2012 proved CNN is most powerful above all in image classification

The significance of R-CNN

• first paper to show CNN beat other models on object detection.

focuse on two problems

• **localizing** objects with a deep network

• **training** a high-capacity model with only a small quantity of annotated detection data.

first problem:localization

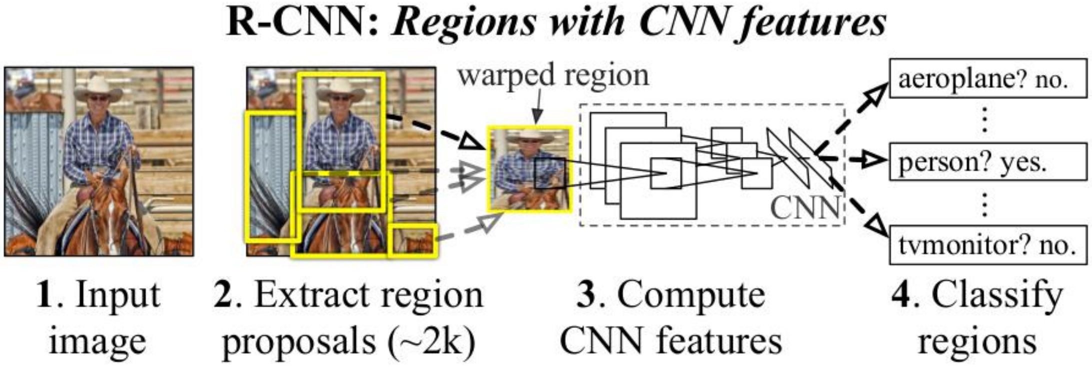
Early solutions

• **regression** is not fare well.

• **sliding window detector**:difficult to make precise localization

CNN's solutions

• **recognition using regions**



- test time, generate 2000 region proposals from image;
- **affine image warping** technique turn region proposal into fixed-size CNN input;
- use CNN to turn each proposal into fixed length feature vector;
- use linear SVMs to classify vectors into different classes.

second problem:labeled data is scarce for training a large CNN

- **conventional solution**:unsupervised pre-training+supervised fine-tuning
- **paper's contribution**:supervised pre-training on large auxiliary dataset+domain-specific fine-tuning on a small dataset.

else

efficient

- only computations:small matrix-vector product and greedy non-maximum suppression.

failure modes

- **bounding-box regression** method to reduces mislocalizations.

2.Object dection with R-CNN

R-CNN has three modules.

- category-independent region proposals
- CNN for extract features
- set of SVMs to classify objects

paper study about modules

- test-time usage
- how parameters are learned
- show detection results

2.1.Module degisn

Region proposals

- use **selective search** to enable a controlled comparison with prior detection work.

Feature extraction

- **AlexNet** network:Input 227x227 RGB images,output 4096D vector
- **tight bounding box** 227x227:**dilate** 16 pixels on each side as context before **warping**.

2.2.Test-time detection

• **selective search** to extract around 2000 region proposals.

• **warp** each proposal and **forward propagate** it through the **CNN** to compute features.

• score each extracted feature vector using the **SVM**.

• apply a **greedy non-maximum suppression** rejects a region have high **IoU**.

Two properties make detection efficient

•all CNN parameters are **shared** across all categories,the only class-specific computations are **dot product** between features and SVM weights and non-maximum suppression.

• the feature vectors computed by the CNN are low-dimensional.

2.3.Training

Supervised pre-training

• **pre-train** CNN using datase(**ILSVRC2012**) no bounding box labels.

• using Caffe CNN library.

Domain-specific fine-tuning

• 1000-way classification layer to **(N+1)**-way classification layer.(N is the number of object classes,plus 1 for background)

• use only warped region proposals from VOC.

• positives:>=**0.5** IoU overlap,the rest as negatives.

• start SGD with **1/10th** of the initial pre-training rate while not clobbering the initialization.

•batch of size 128:**32** postive windows(over all classes) and **96** background windows.(background is general)

Object category classifiers

• **label** a region with an IoU overlap threshold(below as negatives).

• since the training data is too large to fit in memory,adopt **hard negative mining method.**

2.4.Results on PASCAL VOC 2010-12

• R-CNN is most directly comparable to **UVA** since all methods use seletive search region proposals,from 35.1% to 53.7% mAP while also **faster**.

2.5.Results on ILSVRC2013 detection

• R-CNN achieves a mAP of **31.4%** compared with **24.3%** from OverFeat.

• how CNNs can be applied to object detection,leading to greatly varying outcomes.

3.Visualization,ablation,and modes of error

3.1.Visualizing learned features

• **First-layer filters** are easy understand which capture oriented edges and opponent colors.

• **the subsequent layers** is difficult to understand.

• **non-parametricmethod**:single out a particular unit in the network and user it as if it were an object detector in its own right.

1.compute the unit's activations on a large set of held-out region proposals

2.sort the proposals from highest to lowest activation

3.perform non-maximum suppression

4.display the top-scoring regions

- **pool5**:6x6x256=9216-dim; top 16 activations; 6 of 256 features.
- the network learn a **representation** combines a small number of class-tuned features together with a distributed representation of shape,texture,color and material properties.
- **fc6** to model a large set of compositions of features.

3.2.Ablation studies

Performance layer-by-layer,without fine-tuning

- fc7 **worse** than features from fc6 mean that a large CNN's parameters can be removed without degrading mAP.
- only **pool5** has good results mean that much of the CNN's **representational power** comes from convolutional layers rather than fully connected layers.

Performance layer-by-layer,with fine-tuning

- improvement is striking.
- the **pool5** features learned from ImageNet are **general** and **the most of the improvement** is gained from learning domain-specific non-linear classifiers.

Comparison to recent feature learning methods

- DPM:uses only HOGfeatures. (R-CNN 54.2% vs DPM 33.7%)
- DPM ST:augments HOG features with histogram of "sketch token" probabilities.(2.5mAP)
- DPM HSC:replaces HOG with histograms of sparse codes(HSC).(4mAP improves over HOG)

3.3.Network architectures

- **the choice of architecture** has a large effect on R-CNN detection performance.
- VGG increasing mAP from 58.5% to 66.0%,but forward pass taking 7 times longer than AlexNet.

3.4.Detection error analysis

- 4 index

  Loc-poor localization

  Sim-confusion with a similar category

  Oth-confusion with a dissimilar object category

  BG-a FP that fired on background
- **more Loc** indicating CNN-features are more discriminative than HOG.
- **result**:Loose localization from our use of bottom-up region proposals and the positional invariance learned from pre-training the CNN for whole-image classification.

3.5.Bounding-box regression

•to reduce localization errors,train a **linear regression model** to predicct a new detection window given the pool5 features for a selective search region proposal.

3.6.Qualitative results

- precision greater than 0.5 are shown (not curated&been curated)

4.The ILSVRC2013 detection dataset

4.1.Dataset overview

compoent

- three sets:train(395918),val(20121),test(40152)
- val and test

same image distribution; similar to PASCAL VOC; exhaustively annotated(labeled with bounding boxes)

- train

more variable complexity with a skew towards images of a single centered object; not exhaustively annotated

- negative images: do not contain any instances of their associated class.

problem

- train images cannot be used for **hard negative mining**.


光明
cv phd
21 人赞同了该回答
研究了一下，希望对你有帮助。首先是negative，即负样本，其次是hard，说明是困难样本，也就是说在对负样本分类时候，loss比较大（label与prediction相差较大）的那些样本，也可以说是容易将负样本看成正样本的那些样本，例如roi里没有物体，全是背景，这时候分类器很容易正确分类成背景，这个就叫easy negative；如果roi里有二分之一个物体，标签仍是负样本，这时候分类器就容易把他看成正样本，这时候就是had negative。
hard negative mining就是多找一些hard negative加入负样本集，进行训练，这样会比easy negative组成的负样本集效果更好。主要体现在虚警率更低一些（也就是false positive少）。
编辑于 2017-12-27
▲ 已赞同 21 ▼ 5 条评论 分享 收藏 感谢 ···

• Where should negative examples come from.
• Should the train images be used or not,and if so,to what extent?

strategy

• rely heavily on the **val** set and use some of the **train** images as an auxiliary source of **positive** examples.
• use val both training and validation,split it into equally ”val1”&”val2”.
• to produce an approximately class-balanced partition, the smallest maximum relative imbalance(|a-b|/(a+b)) was selected.

4.2.Region proposals

• selective search is **not scale invariant**, the number of regions produced depend on the image resolution,so resized each image to a fixed width.

4.3.Training data

• val1+trainN:
• training data three procedures:

(1)CNN fine-tuning:run 50k SGD on val1+trainN
(2)detector SVM training:all ground-truth boxes from val1+trainN
Hard negative mining: randomly selected subset of 5000 images from val1.
(3)bounding-box regressor training:trained on val1.

4.4.Validation and evaluation

• validated data usage choices and the effect of fine-tuning and bounding-box regression on val.
• goal is to produce a preliminary R-CNN result on ILSVRC without extensive dataset tuning.

4.5.Ablation study

• argument:training data, fine-tuning and bounding-box regression

4.6.Relationship to OverFeat

• OverFeat as a **special case** of R-CNN

seletive search -> a multi-scale pyramid of regular square regions
per-class bounding-box regressors -> a single bounding-box regressor

- OverFeat is 9x **faster**(2s/image) than R-CNN
- OverFeat's **sliding windows** not warped at the image level and computation can be easily shared between overlapping windows.

5.Semantic segmentation

- **O2P** for **"second-order pooling"**:current leading semantic segmentation system

CNN features for segmentation

- **full**:ignores the region's shape and computes CNN features directly on the warped window.
- **fg**:computes CNN features only on a region's foreground mask.
- **full+fg**:simply concatenates the full and fg features.

Results on VOC 2011

- layer fc6 always outperforms fc7
- the fg strategy slightly **outperforms** full,indicating the masked region shape provides a stronger signal.
- full+fg indicating that full features is **highly informative** even given the fg features.

6.Conclusion

- the old way:**complex ensembles** combining multiple low-level image features with high-level context.
- R-CNN gives a 30% improvement over the best previous results on PASCAL VOC 2012.
- two insights

1.apply high-capacity **CNNs** to bottom-up region proposals to localize and segment objects.
2.**training** a large CNNs(supervised pre-training/domain-specific fine-tuning) when labeled training data is scarce.

- combination of **computer vision**(bottom-up region) and **deep learning**(CNN).