S16　論文紹介

杉山明大

概要

選定基準

Alex Net + 物体検出に関する論文から。

論文

1. **ImageNet Classification with Deep Convolutional Neural Networks**
2. **You Only Look Once: Unified, Real-Time Object Detection**
3. **SSD: Single Shot MultiBox Detector**
4. **YOLO9000: Better, Faster, Stronger**
5. **YOLOv3: An Incremental Improvement**

物体検出論文の読む順番おすすめ

RCNN > Fast RCNN > (Faster RCNN) > YOLO > SSD > YOLOv2(YOLO9000) > YOLOv3 > ...

ImageNet Classification with Deep Convolutional Neural Networks

(2012) Alex Krizhevsky / Ilya Sutskever / Geoffrey E. Hinton

https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

どんなもの？

D-CNNによる画像認識の学習の高速化と精度向上をまとめた論文。通称AlexNet。
ImageNet ILSVRC-2010コンペの120万画素の画像の1000クラス分類問題で、畳込5層、全結合3層のD-CNNを使用し、top-1エラー率37.5%、top-5エラー率17.0%を達成した。
学習時間高速化のために非飽和ニューロン(=LeRUを使用)と効率的なGNN演算のGPUを実装し、過学習を減らすためにドロップアウトと呼ばれる正則化方法を採用。

どうやって有効だと検証した？

世界的な画像認識コンペILSVRC-2010, 2012のデータセットを使用し、従来一般的に使用されていたモデルと、本研究モデルの精度を比較し有効性を証明した。

技術の手法や肝は？

過学習抑制
- Data Augmentation: 学習サンプル増加。水平反転、切り出し、RGB強度変更。
- Dropout: 隠れニューロンの重みを確率50%で0にする。2つの全結合層で適用。
- Overlapping Pooling: Max Poolingをとるときに少し重複させる。

学習高速化
- ReLU: 一般的だったtanhと比較し5-6倍の学習速度でエラー率を下げた。
- マルチGPUで学習: 画像の上半分と下半分で分散処理。

議論はある？

- D-CNNを教師学習するだけで大幅な精度向上
- 中間層を1つ除くと2%精度悪化。Deepなことは本質的に精度に影響を与える。
  →学習データ量（連続画像=動画）増加、GPU性能向上で更なる精度向上を示唆。

先行研究と比べて何がすごい？

ILSVRCの2012年コンペでtop-5エラー率15.3%を記録。2番目のモデルは26.2%で、10%以上離す驚異的な結果を達成。以降、画像認識の研究トレンド=D-NN一色に変えた。
従来の画像認識は、職人技と思われていた特徴量設計(人手) が中心。手法：HaarLike, HOG, SIFT.
汎用的なD-NNで成果を出し、その後の研究を活発化させた。

次に読むべき論文は？

参照文献で読んだのをここにお書き

2020/9/18

You Only Look Once: Unified, Real-Time Object Detection

（2016）Joseph Redmon / Santosh Divvala / Ross Girshick / Ali Farhadi

https://arxiv.org/abs/1506.02640

どんなもの？

通称YOLO。高速化した物体検出のアルゴリズムでリアルタイムで処理が可能なCNN。既存のDPMやR-CNN系は領域推定と分類処理を分けて行うため処理時間が長く、YOLOはその問題を解決する手法。通常45fps, fast版150fps

どうやって有効だと検証した？

リアルタイム処理可能な手法と結果を比較。
その他）Fast R-CNN：70.0 mAP , 0.5fps

手法	mAP	FPS
100Hx DPM	16.0	100
30Hz DPM	26.1	30
Fast YOLO	52.7	155
YOLO	63.4	45

技術の手法や肝は？

objectnessの領域提案と分類を同時に実行。
画像をSxSのgridに分割する。各grid cellはB個のbounding boxes(物体の領域候補) と各bounding boxの信頼度のパラメータを保有する。また、各grid cellは下記のような条件確立を持つ。
Pr(Class_i) * IOU_truth pred (その物体が何であるか * そのgrid cellに物体がある確率)

議論はある？

各grid cellが2つのボックスしか予測せず、1つのクラスしか持つことができないため、空間的制約があり、予測できる近くの物体の数を制限する。鳥の群れのような集団で出現する小さな物体に苦戦する。
検出性能に近似した損失関数を用いて学習を行う。boundign boxesの大小に関わらず誤差を同じように扱うため、小さいboxは僅かな誤差でもIOUに大きな影響を与える。主なエラーの原因は、不正確なローカライズ。

先行研究と比べて何がすごい？

- 処理が早い：シンプルな回帰問題に落とし込むことにより、複雑なパイプラインを考慮する必要がなくなった。
- 高精度：sliding window, Region Proposalではなく、画像全体の情報から学習や検証を実施。背景を物体と誤検出する件数がFast R-CNNの半分以下。
- 汎化性能が高い：どのような領域の画像に対しても学習可能

次に読むべき論文は？

(M. A. Sadeghi and D. Forsyth. 30hz object detection with dpm v5. In Computer Vision–ECCV 2014, pages 65–79.Springer, 2014. 5, 6)

2020/9/23

SSD: Single Shot MultiBox Detector

（2016）Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed,
Cheng-Yang Fu, Alexander C. Berg

https://arxiv.org/abs/1512.02325

どんなもの？

画像中の物体を単一のNNを使用して検出する手法。通称SSD。
特徴量地図のアスペクト比とスケールに関係なくバウンディングボックスの出力空間をデフォルトボックスのセットへと離散化する．画像をグリッドで分割して、それぞれのグリッドに対して固定されたいくつかのバウンディングボックスの当てはまり具合を見る手法

議論はある？

SSD512とFaster[25]のAvg.PrecisionとAvg.Recallを比べると、大きいサイズでは5%程度改善されるのに対して、小さいサイズでは1~2%程度しか改善していない。
SSDの計算時間の80%はベースネットワーク(今回はVGG16)に費やされているので、ベースネットワークの速度が改善されればより速いモデルになる。

どうやって有効だと検証した？

PASCAL VOC2012のテストデータとCOCOののデータセットでも検証されている。

先行研究と比べて何がすごい？

SD300とSSD512は、速度と精度の両面でFaster R-CNNを上回る。

技術の手法や肝は？

SSDではExtra Feature Layersという畳み込み層を挿入している。あとの特徴マップの分割領域数をスケールダウンさせている。この分割領域数を減らすと、デフォルトのボックはより大きくなり、大きい物体に当てはまりやすくなる。

次に読むべき論文は？

YOLO9000: Better, Faster, Stronger

2020/9/23

YOLO9000: Better, Faster, Stronger

（2016）Joseph Redmon, Ali Farhadi

https://arxiv.org/abs/1612.08242

どんなもの？

通称YOLOv2。YOLOより物体検出速度が向上。また検出できるクラス数が9000種類以上まで増えた。

議論はある？

どうやって有効だと検証した？

PASCAL VOCやCOCOのデータセットで検証。
VOC 2007: 76.8 mAP in 67 FPS, 78.6 mAP in 40 FPS
これはResNetやSSDを使用したRCNNの高速化など、最先端の方法よりも優れている。

先行研究と比べて何がすごい？

早く精度も高い。
YOLOv2は競合する方法よりもはるかに高速に動作しながら73.4 mAPを達成します。また、COCOを訓練し、表5の他の方法と比較します。VOCメトリック（IOU = .5）では、YOLOv2はSSDとFaster R-CNNに匹敵する44.0 mAPになります。

技術の手法や肝は？

バッチ標準化
ImageNetで自選額種された高解像度分類器
アンカーボックスを使用して、境界ボックスを予測。

次に読むべき論文は？

YOLOv3: An Incremental Improvement

2020/9/23

YOLOv3: An Incremental Improvement

（2018）Joseph Redmon / Ali Farhadi

https://arxiv.org/abs/1804.02767

どんなもの？

YOLOv2からのアップデート。ネットワークは少し大きくなったが、精度がより向上。
320 × 320 pixelの画像の物体検出の結果。
SSDの3倍早く検出が可能。

手法	mAP	inference time
SSD	28.0	61ms
YOLOv3	28.2	22ms

どうやって有効だと検証した？

COCOのデータセットを検証データとして、SSDやRetinaNetなどの他のモデルとmAPと検出時間を比較検証した。実行時間はM40またはTitanXで，基本的に同程度の性能を持つGPUで計算。

技術の手法や肝は？

- Class Prediction: softmaxからindependent logistic classifierに変更
  softmaxを使用するのは、各boxが1つのクラスをもつケースばかりではないという仮定の元。
- Feature Extractor: 特徴摘出のための新しいネットワークを採用。53個のconvolutional layersを持つのでDarknet-53と呼ぶ。

議論はある？

先行研究と比べて何がすごい？

mAPはRetinaNetに若干及ばない程度の精度であるが、検出時間は1/9に抑えられているため非常に高速に物体検出が可能。

次に読むべき論文は？

参考文献ではないが。。。

- YouTube - YOLOv3 (Joseph Redmon)
- TED Talks - コンピューターはいかに物体を即座に認識できるようになったのか (Joseph Redmon)

2020/9/23