相対的な位置関係を考慮した一般物体検出の改良

Improving object detection performance using objects' relative positions

佐々木雄一[*1]

[*1] ファッションポケット株式会社
Fashion Pocket Inc.

We introduce a new post-processing method to improve general deep-learning object detectors. In this method, a simple inference model is inserted just before Non-Maximum suppression to update the predicted confidence levels using objects' relative positions. We show it improves the robustness of detections by reviving the items missed by the object detectors.

1. はじめに

一般物体検出では、SSD[Liu 2016]や YOLO[Redmon 2017]などといった手法が広く用いられている。これら手法では Neural Network 部によって候補領域を提案し、その中から妥当性の高い領域を Non-Maximum Suppression (NMS)、および、しきい値処理といった後処理によって抽出し、最終的な検出結果を bounding box として得る。NMS は、単一対象に対して複数の候補領域が提案されるケースに対して、ルールに基づいた機械的な重複除去を実施する。しきい値処理は、信頼度 (confidence level) がしきい値以下の候補領域を落とすという簡潔なルールである。これらを組み合わせた後処理は、一般的に非常に良い性能を達成することが知られているものの、特定の用途においては、さらなる工夫の余地が見受けられる。特に、人間が写った写真から、複数のファッションアイテム (shoes、skirt、tops、等) を検出させるようなタスクの場合には、各アイテム相互の位置関係を踏まえたアルゴリズムを構築することにより、より頑健な物体検出を行うことが可能である。

例えば、図 1 のようなケースを考える。tops や outer、shoes は十分な confidence level にて認識をされており、明確に bounding box が提案をされている。一方で、pants の confidence level は低くなっており、しきい値処理の後に bounding box として提示されていない。結果として、tops、outer、shoes のみが検出され、pants を履いていないという結果が認識結果として提示される。しかしながらこれは常識的に考えてありえない組合せであり、人間が写真を判断する場合、tops と shoes に間に pants もしくは skirt があると推測した上で、その付近を積極的に検索し、confidence level が低い候補領域であっても、最終的な bounding box として提示するのが普通である。

本提案では、neural network 部が提案する候補領域群の相対的な位置関係を学習させることで、常識的に考えづらい不検出に対する頑健性を持つ後処理手法を提案する。この手法は、neural network 部と NMS の間に、小規模な機械学習モデルを導入し、候補領域の confidence level を修正することによって実装されるため、SSD や YOLO などの従来手法に対して簡便に精度向上の手段を提供する。

尚、SSD や YOLO などの neural network には、本来、物体間の相対位置を考慮する仕組みが獲得されている。しかしながら、それによる頑健性の程度は限定的で、しばしば不十分であり、前述のような常識的に考えてありえない組み合わせの bounding box が出力されることは多い。ここでは、その弱点を明示的に解決する手法を提案する。

連絡先：y_sasaki@fashionpocket.jp

図 1　本提案で解決を目指す問題の事例。tops や outer、shoes が検出されているにも関わらず、パンツが検出されていないといった常識的には考えづらい組み合わせが出力されている検出結果

2. 関連研究

SSD と YOLO などから提案される候補領域に対し、続く後処理の改善によって、全体の検出率を改善しようとする試みは多い[Rothe 2014, Mrowca 2015, Bodla 2017, Hosang 2017]。特に、[Hosang 2017] で提案されている GossipNet は、候補領域対に対して機械学習を適用し、相互の重複量・位置関係・カテゴリ種別などに基づき、confidence level を更新する手法を提案している。本提案も同様の考え方に基づいているものの、候補領域対に替わり、3 つの候補領域間の位置情報等に基づき confidence level を更新する手法を採用しており、より相対的な位置関係に感度の高い手法となっている。

機械学習モデルについては、小規模な neural network を用いる。この network 構造については、3 つの候補領域の情報の情報を入力として用いつつも、更に写真全体に渡る特徴量を捉えた分析を行うため、PointNet[X]で用いられた考え方を援用している。PointNet では、点群の情報を小規模な network により個別に要約した後に、全点群からの要約情報を MaxPooling に入れることで、点群全体の特徴を抽出している。それを再度点群情報に足し合わせて再度 Neural Network による処理を行うことで、全体の特徴を取り入れた分析を可能にしている。本提案においては、MaxPooling に代わり AvgPooling を用い、候補領域群全体から抽出した全体特徴量を、再度候補領域の要約情報に足し合わせることで、性能をさらに向上させている。

3. 提案手法

前述のように、本手法は 3 つの候補領域を組として入力する neural network モデルを核としている。以下では、3 つの候補領

域より入力バッチを構成する方法と、それを分析し新しい confidence level を出力する network 構造との 2 つに分けて説明を行う。

まず、入力バッチの構成について述べる。SSD や YOLO などにより提案された候補領域を、confidence level が高い順にソートし、そのうち上位 N_0 件を Box0 として選択する。各 Box0 に対し、それらの周囲に存在する候補領域を 2 つ選び、Box1, Box2 とする。この 2 つの候補領域は、confidence level が N_1 位までの間で、重複を許さないよう選定され、合計 N_1 $(N_1\text{-}1)/2$ 組が作られる。つまり、一枚の写真について、$N_0 \times N_1$ $(N_1\text{-}1)/2$ 個の組が準備される。最後に、それぞれの候補領域の組に対し、特徴量を計算する。この特徴量は、以下で挙げる数値を並べたものである:

- Box0, Box1, Box2 の confidence level
- Box0, Box1, Box2 の面積比
- Box0 に対する, Box1, Box2 の中心位置のユークリッド距離の比および、角度
- Box0, Box1, Box2 のカテゴリを示す one-hot vectors

上記で準備された入力ベクトルの次元数が k 次元、写真が N_p 枚ある場合、最終的に構築される要素の数は $N_p \times N_0 \times N_1$ $(N_1\text{-}1)/2 \times$ k 個となるが、これを、$[N_p \times N_0, N_1 (N_1\text{-}1)/2, k]$ へと形状変換し、$N_p \times N_0$ 件のバッチとして取り扱う。

他方、出力結果と比較される正解ラベルは、$[N_p \times N_0, 2]$の形状を持ったベクトルとする。2 次元目は、最終的な bounding box として残す/残さないに対応する one-hot vector であり、それを $N_p \times N_0$ 件の候補領域について準備する。

次いで、上記入力に対して学習・推論を行う network 構造について説明する。図 2 にあるように、network は 3 つの block を繰り返す構造としている。一つの block は、まず Box0, Box1, Box2 の情報から構成される k 次元の入力ベクトルを、kernel_size=1 の conv1d にて抽象化する。次いで、global average pooling を適用し、図中の他の Box1, Box2 の情報も含めて特徴量を抽出した後、full connect 層によって、入力と同じ k 次元に戻す。この k 次元情報の中には、Box0 を中心として、その周囲にある Box1, Box2 の情報も含めた全体的な特徴が抽出されている。最後に、その情報を、同じく k 次元の入力に加え合わせることによって、全体的な特徴を踏まえつつ、個別の Box0, Box1, Box2 に関する特徴量も保持した処理を可能としている。

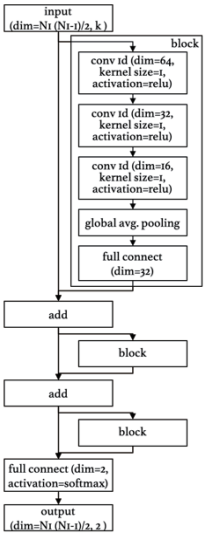

図 2　ネットワーク構造。"block"の構造は 3 回繰り返される。

上記を 1 つの block として、3block 分同様の処理を行った後、full connect 層および softmax にて、正解ラベルと比較される 2 次元の one-hot vector としている。推論の際には、このうちの片方を抽出し、更新された confidence level として扱う。

4. 実験

相対的な位置関係が重要となる問題設定として、今回は、写真の中からファッションアイテムを検出するタスクにおいて、本手法を評価する。

4.1 データ

本実験では、2 種類のデータを使用している。

一つは、独自に収集した画像に、ファッションアイテムの位置・カテゴリについてのアノテーションを行ったものである。基本的に日本人男女の全身写真となっており、日常の背景にて撮影されたものが大半を占める。カテゴリは、outer, tops, pants, skirt, onepiece, shoes, hat, bag の 8 種類とした。今回は、32k 件を母集団とし、8 割を学習用データ、1 割を学習の進捗監視とパラメータ最適化に利用し、1 割を定量評価用のデータとした。

二つ目は、DeepFashion データセット[Z. Liu 2016]であり、これは主に外国人の男女を対象とした全身写真である。本評価では、ここから抽出した数件の画像を用いて定性的な評価を行っている。

4.2 学習

Network は、SSD などが提示する候補領域を入力し、そのうちいずれを残す/残さないようにするかを推定させるよう学習させる。そのため、まず候補領域を準備する必要があり、別途ファッションアイテムの学習をさせた SSD において、NMS やしきい値処理を掛ける前の候補領域を抽出して用いる。

次いで、各候補領域に、最終的な bounding box として残すべきか否かを示す正解ラベルを準備する。これは、アノテーション結果の bounding box と最大の Intersection-over-Union (IoU)を示す候補領域を採ることによって行う。本来、confidence level に関わらず IoU 最大の候補領域を正解としてラベルを作成するべきであるが、本実験においては confidence level 順の上位 N_0 のみが学習に寄与するため、ある程度のしきい値を confidence level に対して要求し、それを満たす候補領域の中で IoU が最大になるものに限って正解ラベルを作成することとする。今回用いた学習データおよび SSD の場合、confidence level>0.05 の候補領域を対象とすることで、N_0=32 の中に、ほぼ 100%の正解ラベルを含めることができる。

Network は、一部候補領域の検出が上手くいかず confidence level が低くなっている場合にも頑健性を持つよう学習を行う。そのような状況を想定した下記のデータ水増しを適用し、データ量を 5 倍に増加させて学習を行う:

- 各画像において、confidence level 順で上位 1 位の候補領域の confidence level に対し、[0.3, 0.8]から乱択される係数を乗算する
- 上位 2 位の候補領域に対して同じ処理を行う
- 上位 3 位の候補領域に対して同じ処理を行う
- 上位 4 位の候補領域に対して同じ処理を行う
- 各画像において、confidence level 順で上位 1, 2, 3 位の候補領域の confidence level に対し、それぞれ [0.3, 0.8] から乱択される係数を乗算する

その他のパラメータについては、右の値を用いた: $N_0 = 32$, $N_1 = 5$, $k = 32$, NMS 後の confidence level しきい値 = 0.6。

Network の学習には Adam を使用した。学習の進捗監視データに対する loss の低減が十分に落ち着くまで約 700epochs 程度の学習を行い、以降の評価をしている。

4.3 評価

まず、mean Average Precision (mAP)による定量評価について述べる。図 3 に、各実験条件で得られた mAP の比較をまとめる。”SSD Output” (灰色) は、SSD から出力された候補領域に対して、通常の NMS およびしきい値処理を行った結果である。”SSD w/ Revive” (オレンジ) は、NMS の直前に本提案手法を挿入した場合を示す。mAP で 3.4%分の精度向上となっており、本提案手法にて狙う SSD の性能向上を実現している。それより下側に続く “Degraded 1st / 2nd / 3rd BB w/o Revive” (青) および、”Degraded 1st / 2nd / 3rd BB w/ Revive” (オレンジ) は、SSD が出力する候補領域の confidence level に 0.5 を乗算し、人工的に悪化させた場合と、それを本手法で回復させた場合の mAP を比較している。いずれの場合も悪化させた分の mAP 低下を回復させ、ほぼ”SSD w/ Revive”に匹敵する水準まで引き戻しており、SSD の不検出に対して頑健な性能を実現していることが分かる。

”SSD Original”と”SSD w/ Revive”を比較した際の、各カテゴリの AP 変化を図 4 にまとめた。特に、Shoes、Hat, Bag など、小さく隠れがちであり、SSD でも検出率が低くなりやすいアイテムに対して、大きく検出率が向上している。

図 5 に、DeepFashion の画像に対して本手法を適用した場合の定性的な結果を示す。DeepFashion は外国人のファッションが主な対象であり、SSD の学習を行った日本人を中心としたデータセットとはドメインを異にしているため、Pants などのアイテムについても SSD での検出が難しい場合がある (左列)。それらについても、本手法を適用することで回復されていることが分かる (右列)。

5. 結論

一般物体認識における後処理の一つとして、3 つの候補領域の相対的な位置関係を学習し、改善する手法を提案した。これは特に、ファッションアイテムの認識など、常識的な組み合わせがある程度固定されている条件で、mAP の向上に寄与することを示した。

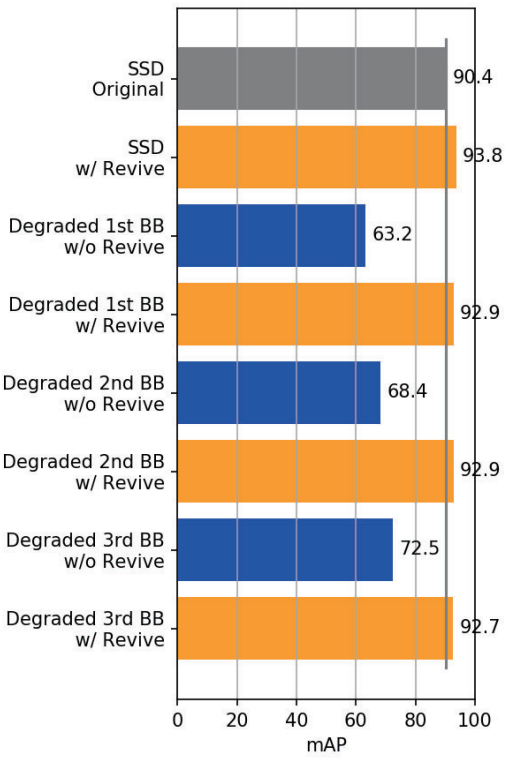

図 3　各実験条件で得られた mAP の比較。最上部 (SSD Original) が SSD の出力そのものでの mAP であり、その一つ下 (SSD w/ Revive) が提案手法を適用した場合。それより下側は、1, 2, 3 位の候補領域の confidence level を 0.5 倍することにより人工的に検出失敗を再現した場合の結果 (w/o Revive)と、提案手法を適用した場合の結果 (w/ Revive)。

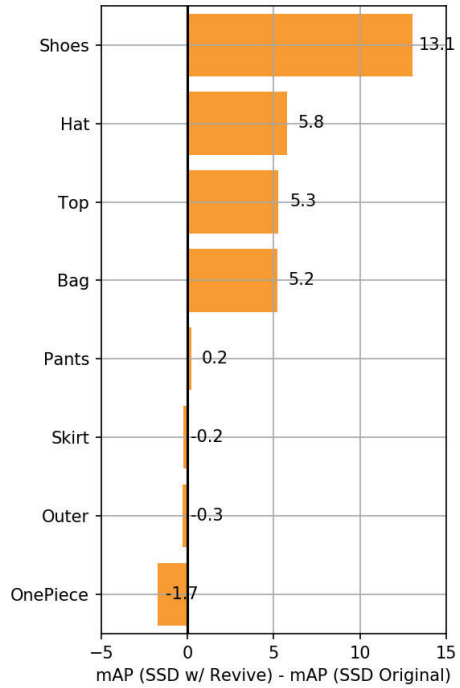

図 4　カテゴリ毎で見た時の、”SSD Original”と”SSD w/ Revive”における AP の差。

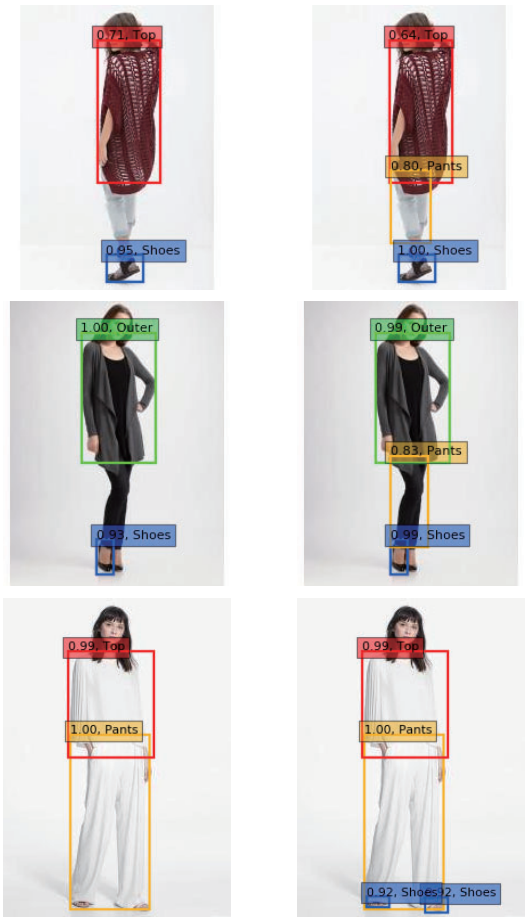

図 5　DeepFashion 画像に対して SSD を適用した結果 (左列)と、検出できなかった bounding box を、提案手法にて回復した例 (右列)

参考文献

[Liu 2016]　W. Liu, et.al., “SSD: Single Shot MultiBox Detector”, ECCV2016

[Redmon 2017]　J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger”, CVPR2017

[Rothe 2014] R. Rothe, et.al., “Non-Maximum Suppression for Object Detection by Passing Messages between Windows”, ACCV2014

[Mrowca 2015] D. Mrowca, et.al., “Spatial Semantic Regularisation for Large Scale Object Detection”, ICCV2015

[Bodla 2017] N. Bodla, et.al., “Soft-NMS: Improving Object Detection With One Line of Code”, ICCV2017

[Hosang 2017] J. Hosang, et.al., “Learning non-maximum suppression”, CVPR2017

[Z. Liu 2016] Z. Liu, et.al., “DeepFashion: Powering Robust Clothes, CVPR2016