マルチクラス化を導入した Grid-wise-attention による物体検出の判断根拠の可視化

EP18085 早川和希　　指導教授：山下隆義

1. はじめに

物体検出は，動画像内の物体の位置とクラスを特定するタスクであり，これまでに深層学習を用いた高精度かつ高速な手法が提案されている．これらの手法は複数のクラスを同時に検出することができる．しかしながら，誤検出や検出漏れがわずかながら生じる．従来手法はそれらに対する判断根拠の視覚的説明がなされていない．そこで，本研究ではマルチクラスの物体検出時に注視した領域をアテンションマップとして獲得可能な物体検出手法を提案する．

2. Grid-wise-attention

物体ごとのアテンションを獲得できる物体検出手法として Grid-wise-attention を用いた手法がある．この手法は画像をグリッドに分割し，各グリッドに対するアテンションマップを出力している．しかしながら，検出できる物体クラスは歩行者のみとなっている．

3. 提案手法

本研究では，Grid-wise-attention による物体検出をマルチクラス化することで検出した全てのクラスに対する判断根拠を獲得可能とする．

3.1. マルチクラス化した Grid-wise-attention

図 1 にマルチクラスに対応した Grid-wise-attention を用いた物体検出の構造を示す．本手法は，Grid-wise-attention による物体検出と同様に Feature Extractor，Classification branch，Localization branch から構成される．Classification branch は各クラスに対する確率をグリッドごとに出力できるようにする．そのため，出力する Class probability は $W \times H \times C$ の確率マップとなる．Localization branch は各グリッドに対して物体の位置と大きさを出力する．また，Grid-wise-attention 機構により，グリッドごとのアテンションマップを出力する．

3.2. アテンションマップの出力

Classification branch の中間特徴に対して，1×1 の畳み込み処理を行い，D 次元の特徴を得る．Localization branch では，この特徴を Query，feature extractor から獲得した特徴マップを Key として内積を行う．これにより，グリッドに対して，最もクラス確率の高いクラスに対するアテンションマップを獲得することが可能である．これにより，マルチクラス化した物体検出における判断根拠の可視化を実現する．

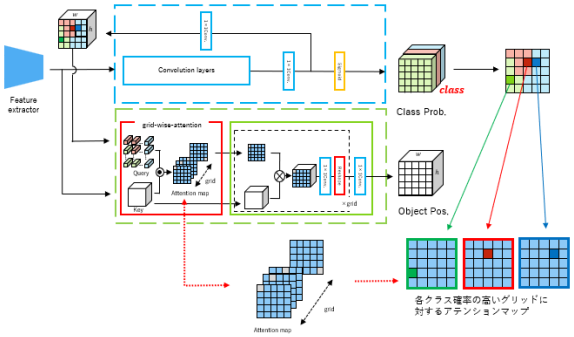

図 1：Grid-wise-attention

4. 評価実験

本実験では，従来手法として CSP[1] と比較を行う．また，Grid-wise-attention によるマルチクラスの検出に対する判断根拠の可視化を行う．

4.1. 実験条件

データセットは CityPersons を使用し，クラスは pedestrian，rider，sitting の 3 クラスである．学習用画像は 2,975 枚，評価用画像は 500 枚を用いる．ベースネットワークモデルは ResNet-50 を使用する．学習回数は 100 エポック，学習率は 5.0×10^{-6}，バッチサイズは 15 である．定量的な評価指標には log-average miss rate を用いる．また，誤検出数に対する miss rate を示す DET (Detection Error Trade-off) カーブを用いた評価を行う．

4.2. 実験結果

CSP と提案手法の比較結果を表 1 に示す．提案手法は CSP に比べて 23.96 ポイント低下しており，精度が向上していることが分かる．

表 1：log-average miss rate

	CSP	提案手法
miss rate [%]	33.7	9.74

また，DET カーブを図 2 に示す．これにより，提案手法は miss rate が低く，精度向上していることがわかる．

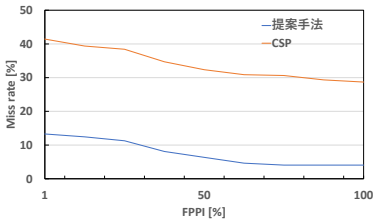

図 2：DET カーブ

4.3. アテンションマップの可視化

提案手法による出力画像とアテンションマップを図 3 に示す．このアテンションマップでは，それぞれのクラスに対するアテンションマップであり，黄色ほど強く，紫ほど弱く注視していることを示す．図 3(b) から，歩行者に対して上半身を強く注視していることから歩行者として正しく検出されている．また，図 3(c) から，歩行者と共通する頭部での注視が強いことから，rider を歩行者と誤検出されていると考える．

図 3：提案手法での検証画像による検出結果とアテンションマップ

5. おわりに

今研究では，物体検出をマルチクラス化することで全てのクラスに対する判断根拠の獲得を行った．結果として，本手法によるマルチクラス検出は各クラスに共通する頭部付近を強く注視していることがわかり，誤検出が発生した判断根拠を示すことができた．今後はさらなる高精度化と，検出対象のみの注視領域の獲得を目指す．

参考文献

[1] W. Liu, *et al.*, "Center and Scale Prediction:Anchor-free Approach for Pedestrian and Face Detection", CVPR，2019