Self-Attention を用いた 3 次元点群からの物体検出の高精度化

ER18068 三原一真　　指導教授：藤吉弘亘

1. はじめに

全方位 LiDAR は，レーザーを回転させながら全方位に照射し，3 次元点群データを獲得するセンサである．深層学習で 3 次元点群を扱うには，点群データを Voxel に置き換える手法，点群データを 2 次元画像に変換する手法，点群データを直接用いる手法がある．点群データを直接入力する手法は変換処理を必要としないため，計算コストが増加しないメリットに加え，他の手法と比較して正確な位置情報が保持される．一方で，点群数が少ない物体に対して検出できないことが多い．そこで，本研究では，点群数の少ない物体に対する精度の改善をするために，近傍の画素との関係を考慮することができる Self-Attention を導入し，高精度な物体検出手法を提案する．

2. PointPillars

PointPillars[1] は点群データを直接用いる物体検出手法である．この手法では，点群データを $x-y$ 平面に対して垂直な pillar(柱状) に含まれる点群データを切り出して入力特徴としている．PointPillars のネットワークである pillarFeatureNet は，入力特徴から点群を含む pillar の数，pillar 内部の点の数，点を構成する 9 次元の情報を，3 次元のテンソルへ変換し，特徴の抽出を行い CNN へ入力する．その後，CNN が出力した 3 次元のテンソルを pillar の元の位置に戻し，C, H, W の 3 次元のテンソルに変換して出力する．また，Backbone にて 2 次元畳み込み処理を使用することによって，処理の高速化を図っている．一方で，Cyclist や Pedestrian などの点群数が少ないクラスは電柱や樹木等の形状の近い物体に誤識別されることが多い．

3. 提案手法

従来手法出が検出が困難な点群数の少ない自転車や歩行者の検出精度を向上させるために，本研究では Self-Attention 機構を PointPillars に導入する．

3.1. ネットワーク構造

図 1 に，ネットワーク構造を示す．まず PillarFeatureNet の出力に対し，Self-Attention Block にて pillar 間の関係性を表現する．次に，Backbone にて点群数の少ない物体の特徴を捉え，小物体の認識精度の向上を目指す．最後に，Detection Head にて畳み込み処理を行い，クラス分類と bounding box の推定を行う．

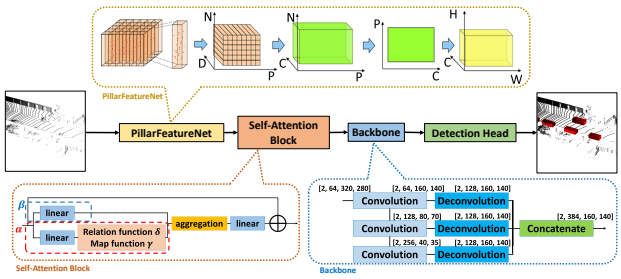

図 1：提案手法のネットワーク構造

3.2. Self-Attention Block

Self-Attention Block (SAB) では，PillarFeatureNet の出力に対して点群数の少ないクラスの特徴の抽出を行う．Self-Attention Block を式 (1) に示す．

$$y_i = \sum_{j \in R(i)} \alpha(x_{R(i)})_j \odot \beta(x_j) \tag{1}$$

ここで，α は加重和にかける重みの係数を算出する関数であり式 (2) より求める．

$$\alpha() = \gamma(\delta(x_{R(i)})) \tag{2}$$

ここで，δ は，線形変換後に注目点と周囲の点を用いてベクトルを出力する関係関数であり，ベクトルを整形するマッピング関数 γ に入力する．式 (1) の β は線形変換のみを行う．その後，2 つのベクトルのアダマール積を算出した後に入力ベクトルと結合させて出力する．

4. 評価実験

評価実験では，提案手法の有効性を検証する．

4.1. 実験概要

学習条件は，学習回数を 160 エポック，バッチサイズを 2 とする．また，最適化手法に Adam，初期学習率は 0.0002 とし，15 エポックごとに学習率を 0.8 倍とする．データセットには，自動車や歩行者，自転車等のクラスが定義されている KITTI 3D Object Detection Benchmark[2] を用いる．本実験では，7,481 枚の学習用データを 3,712 枚と 3,769 枚のデータに分割し，それぞれを学習用と評価用のデータとして使用する．評価指標として Precision を用い，クラスごとに Average Precision を算出する．

4.2. 評価結果

従来手法と提案手法による精度比較を表 1 に示す．表 1 より，従来手法と比較し，提案手法の mAP が KITTI test 3D Object Detection Benchmark において 2.25 pt 向上し，自転車と歩行者の識別精度が Moderate においてそれぞれ 3.72 pt，2.36 pt 向上した．また，KITTI test BEV Object Detection Benchmark において 1.96 pt 向上し，自転車と歩行者の識別精度が Moderate においてそれぞれ 4.10 pt，4.03 pt 向上した．図 2 に各手法の検出例を示す．

表 1：従来手法と提案手法の精度比較 [%]

	method	mAP	Car			Cyclist			Pedestrian		
			Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
3D	PointPillars	51.62	80.06	69.47	67.03	62.08	50.79	48.25	35.36	34.60	32.65
	proposed	**53.87**	**85.41**	**70.15**	**67.90**	**67.16**	**54.51**	**51.53**	**38.69**	**36.96**	**34.15**
BEV	PointPillars	60.41	89.83	**86.08**	**80.72**	64.42	54.16	51.04	43.06	40.98	39.76
	proposed	**62.37**	**89.89**	83.86	80.30	69.30	**58.26**	**55.38**	**47.24**	**45.01**	**41.59**

図 2 から，従来手法と比較して歩行者，自転車の誤識別が減少していることが分かる．以上から，小物体に対する提案手法の有効性を確認できる．

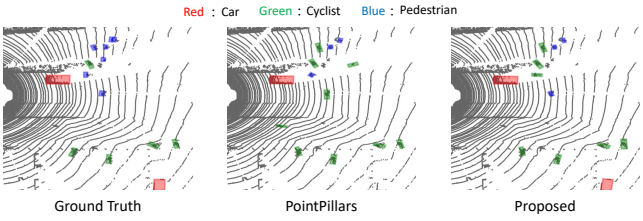

図 2：従来手法と提案手法の精度比較

5. おわりに

本研究では，Self-Attention 機構を導入した 3 次元物体検出手法を提案した．評価実験の結果，従来手法と比較して精度が向上した．今後は，さらに有効な Self-Attention Block を用いたネットワーク構造を検討する．

参考文献

[1] A. Lang, *et al.*, “PointPillars: Fast Encoders for Object Detection from Point Clouds”, CVPR, 2019.

[2] A. Geiger, *et al.*, “Are we ready for autonomous driving? the kitti vision benchmark suite”, CVPR, 2012.

[3] H. Zhao, *et al.*, “Exploring Self-attention for Image Recognition”, CVPR, 2020.