CNN と Transformer を活用したマルチタスクモデルの学習方法に関する研究

TP21007 張 陳雨　　指導教授：山下隆義

1. はじめに

自動運転システムを実現するためには，道路上の車両や歩行者の認識，交通信号や標識の理解，適切な車両制御など，複数のタスクを同時にかつ高精度に実行する必要がある．深層学習によるマルチタスク学習は，互いに関連した複数のタスクを単一モデルで実行することで，計算コストの削減及び精度向上を実現している．そのため，自動運転システムには，マルチタスク学習が適していると考えられる．

従来のマルチタスク学習は，CNN をバックボーンネットワークに用いてタスク間で共通した特徴を抽出する．CNN は局所特徴の抽出能力が高く，画像の細部の特徴表現を獲得できる．一方，大域特徴の抽出能力は低いため，画像全体のコンテキストを捉えることが難しい．逆に Transformer は，大域特徴の抽出能力は高いが，局所特徴の抽出能力が低い．自動運転におけるマルチタスク学習では，局所特徴が重要となる物体検出，大域特徴が重要となるセマンティックセグメンテーションを対象とする．そのため，バックボーンネットワークの特徴抽出能力が不十分であると，全てのタスクで精度低下を引き起こす可能性がある．また，画像から層ごとに特徴抽出する際，低層ではエッジ，色など詳細な特徴を捉える．一方で，高層ではクラス固有の複雑な特徴が多い．各タスクに必要な特徴が異なるため，精度向上にはタスクごとに適した層の特徴を利用することが求められる．

そこで本研究では，より強い特徴抽出能力を獲得するために，バックボーンネットワークに局所かつ大域特徴を獲得できる Next-ViT [1] を採用する．また，タスクごとに適切な特徴を利用するために，特徴融合手法である BiFPN [2] を導入する．

2. 自動運転におけるマルチタスク学習

MultiNet[3] は分類，検出，セマンティックセグメンテーションを行う Encoder-Decoder 型のマルチタスク学習手法である．MultiNet は，CNN をバックボーンに用いて特徴を抽出するため，大域な特徴抽出が不十分である．DLT-Net[4] は，FPN 構造を Encoder に採用し，CNN をバックボーンに用いて抽出した特徴を融合する．Decoder では Context Tensor モジュールにより，異なる層から抽出した特徴の情報を連結する．DLT-Net は，タスクごとに適切な特徴を利用することが各タスクの性能に大きく影響を及ぼすことを示した．

3. 提案手法

本研究では，従来の自動運転におけるマルチタスク学習の課題を解決するために，Encoder には局所かつ大域特徴を獲得できる Next-ViT [1] をバックボーンとする．また，タスクごとに適切な特徴を得るために特徴融合手法である BiFPN [2] を導入する．提案手法のモデル構造を図 1 に示す．

3.1. Encoder

Encoder には 4 層で構成する Next-ViT を用いる．Next-ViT は CNN と Transformer を活用したモデルである．Next-ViT で抽出した各層の特徴をマルチスケール特徴として扱う．マルチタスク学習には，タスクに合わせた特徴が必要であるため，特徴融合手法 BiFPN を導入する．BiFPN は，各層で抽出した特徴に重み付けし，双方向で特徴を融合する手法である．Next-ViT から抽出したマルチスケール特徴をタスクに合わせた特徴融合を行う．

3.2. Decoder

本研究では自動運転のマルチタスクモデルとして，物体検出と運転可能領域セマンティックセグメンテーション，車線検出を対象とする．そのため，Decoder には，検出ヘッドとセグメンテーションヘッドを構築する．検出ヘッドでは，バウンディングボックスのオフセット，各クラスの確率，および信頼度を予測する．

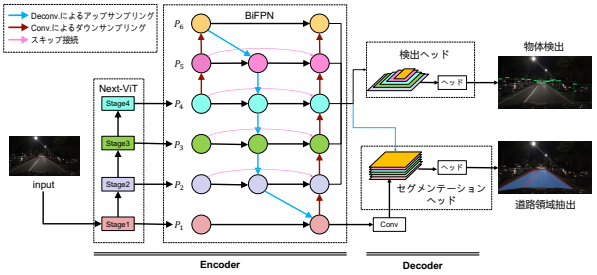


図 1: 提案手法のモデル構造：P_* は各層の特徴である．

セグメンテーションヘッドでは，各画素を運転可能領域，車線，および背景の 3 クラス分類する．また，セグメンテーションタスクでは，大域的な特徴だけでなく，運転可能領域，車線検出タスクで重要であるエッジ情報を捉えるために，低層の特徴も入力する．

3.3. 損失関数

マルチタスク学習は，タスク間における学習難易度の相違により，全タスクを効率的に学習できない問題がある．そこで，本研究における損失関数 L_{total} を以下に示す．

$$L_{total} = \alpha L_{det} + \beta L_{seg} \tag{1}$$

$$L_{det} = \alpha_{cls} L_{cls} + \alpha_{obj} L_{obj} + \alpha_{bbox} L_{bbox} \tag{2}$$

$$L_{seg} = L_{tversky} + \gamma L_{focal} \tag{3}$$

ここで，検出タスクの損失関数 L_{det} は，クラス分類 L_{cls}，信頼度 L_{obj}，バウンディングボックス位置に対する損失 L_{bbox} の和である．また，L_{cls} と L_{obj} には Focal loss[5]，L_{bbox} には smooth L1 loss を用いる．セグメンテーションタスクの損失関数 L_{seg} は，Focal loss と Tversky loss[6] を用いることで，クラス不均衡と分類困難なクラスの学習に対処する．α, β, γ は各損失に対する係数である．これらの係数を学習難易度に合わせてあらかじめ調整することで，効率的にマルチタスク学習できる．

4. 評価実験

本研究では，Berkeley Deep Drive Dataset (BDD100K) [7] を用いて学習及び評価を行う．学習用画像 70,000 枚，評価用画像 10,000 枚である．バックボーンである Next-ViT には ImageNet による事前学習済みモデルを用いる．各タスクの評価指標として，物体検出タスクは mAP50 と Recall，運転可能領域タスクは mIoU，車線検出タスクは Accuracy と IoU を用いる．Next-ViT と BiFPN の有効性を評価するため，提案手法 (Ours) と，バックボーンを単純な CNN としたモデル (w/ CNN) 及び BiFPN 特徴融合を利用しないモデル（w/o BiFPN）を比較した．

4.1. 各タスクに対する定量的評価

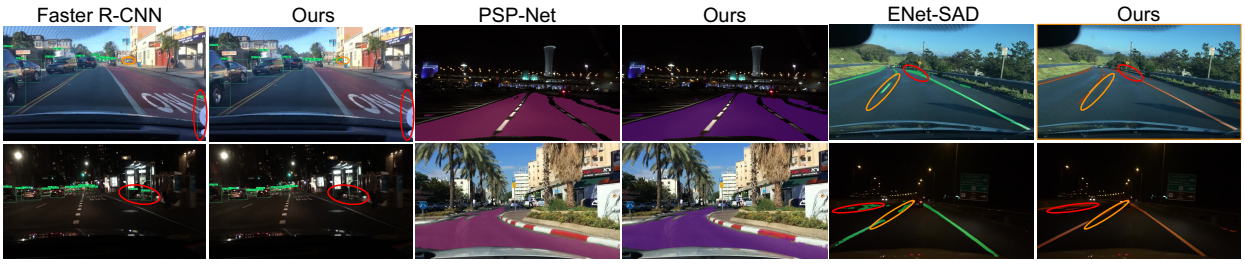
各タスクに対する定量的評価を表 1 に示す．物体検出タスクでは，提案手法は Recall が 95.0%， AP50 が 76.5% であり，Faster R-CNN，MultiNet，DLT-Net に対して精度が向上している．運転可能領域タスクでは，提案手法は mIoU が 87.5% であり， MultiNet と DLT-Net より 16pt 程度向上している．PSP-Net と比較して 2.1pt 低下しているが，推論速度は 3 倍近く高速である．車線検出タスクでは，提案手法は Accuracy が 82.1%，IoU が 24.0% であり，この精度は従来のシングルタスク手法を大きく上回っている．これらの評価結果により，提案手法が有効であることが言える．

4.2. 各タスクに対する定性的評価

各タスクに対する定性的評価を図 2 に示す．図 2(a) より，物体検出タスクでは，Faster R-CNN は車内のオブジェクトや歩行者を車と誤認識した．また，遠方の車も検出していない．一方，提案手法は未検出や誤認識がなく，正確に

表 1: 各タスクに対する定量的評価：* はシングルタスク手法である．

Object detection				Drivable area seg.			Lane line detection			
method	Recall [%]	AP@.5 [%]	Speed [fps]	method	mIoU [%]	Speed [fps]	method	Acc [%]	IoU [%]	Speed [fps]
Multi-Net	81.3	60.2	8.6	Multi-Net	71.6	8.6	Enet*	34.1	14.6	**100.0**
DLT-Net	89.4	68.4	9.3	DLT-Net	71.3	9.3	SCNN *	35.8	15.8	19.8
Faster R-CNN*	77.2	55.6	8.8	PSP-Net*	**89.6**	11.1	Enet-SAD*	36.6	16.0	50.6
YOLOv5s*	86.8	**77.2**	**82.0**	-	-	-	-	-	-	-
Ours	**95.0**	76.5	31.0	Ours	87.5	**31.0**	Ours	**82.1**	**24.0**	31.0



(a) 物体検出タスクにおける比較例　(b) 運転可能領域タスクにおける比較例　(c) 車線検出タスクにおける比較例

図 2: 各タスクにおける可視化例

(a) w/ CNN　(b) Ours　(c) w/o BiFPN

図 3: Ablation study における可視化例

表 2: Ablation study

				Object detection.		Drivable area seg.	Lane line detection	
Method	CNN	Next-ViT	BiFPN	Recall [%]	AP@.5 [%]	mIoU [%]	Acc [%]	IoU [%]
w/ CNN	✓		✓	94.8	75.5	85.8	79.1	21.3
w/o BiFPN		✓		89.6	45.8	79.9	71.6	18.7
Ours		✓	✓	**95.0**	**76.5**	**87.5**	**82.1**	**24.0**

検出している．これは，提案手法で用いる Next-ViT により，画像の局所特徴と大域特徴の両方を捉えた結果であると考えられる．図 2(b) より，運転可能領域タスクでは，提案手法と PSP-Net のセグメンテーションは同等であることがわかる．これは，提案手法は他タスクで獲得したエッジ情報がセグメンテーションに貢献してたためであると考えられる．図 2(c) より，車線検出タスクでは，提案手法は ENet-SAD よりも明らかに車線領域が正確かつ連続的であることが確認できる．運転可能領域タスクと特徴を共有することで，提案手法では運転可能領域と車線間での誤認識が大幅に抑制されていると考えられる．

4.3. Ablation study

提案手法における Encoder の有効性に対する定量的評価を表 2 に示す．定量的結果から，w/ CNN の場合，全タスクに対してわずかに精度が低下している．これは，バックボーンの特徴抽出能力の相違により，各タスクに性能を影響与えたと考えられる．また，w/o BiFPN の場合，各タスクに対する精度が大幅に低下していることがわかる．

提案手法の有効性に対する定性的評価を図 3 に示す．定性的結果から, CNN と Transformer を活用した Next-ViT を利用することで，局所かつ大域特徴を獲得でき，未検出と誤検出が改善していることがわかる．また，w/o BiFPN の場合，多くの未検出と誤検出が発生している．つまりタスクごとに適切な特徴の獲得ができなかったと言える．

5. おわりに

本研究では，自動運転タスクにおいて CNN と Transformer を活用した Next-ViT と 特徴融合手法 BiFPN を導入したマルチタスクモデルを提案した．BDD100K による評価実験から，自動運転においてマルチタスクでの精度向上を確認した．今後はモデルの高速化を目指す予定である．

参考文献

[1] J. Li, *et al.*, “Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios”, arXiv, 2022.

[2] M. Tan, *et al.*, “EfficientDet: Scalable and Efficient Object Detection”, CVPR, 2020.

[3] M. Teichmann,*et al.*, “Multinet: Real-time joint semantic reasoning for autonomous driving.” IEEE IV, 2018.

[4] Y. Qian,*et al.*, “DLT-Net: Joint detection of drivable areas, lane lines, and traffic objects.” IEEE IV, 2019.

[5] T. Lin, *et al.*, “Focal Loss for Dense Object Detection”, ICCV 2017.

[6] S. Salehi, *et al.*, “Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks”, MLMI, 2017.

[7] F. Yu,*et al*, ,“BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, ” CVPR, 2020.

研究業績

[1] 張 陳雨 等,“Next-ViT と BiFPN による車載カメラ映像からのマルチタスクの高精度化”, ビジョン技術の実利用ワークショップ, 2023.