패턴인식 보고서

120220210 고재현

2022년 12월 19일

1 개요

1.1 목적

- 최근 출시되는 자율주행 자동차에 포함된 pattern recognition/computer vision 관련 요소기술을 조사한다.
- 해당 요소기술과 수업에서 다룬 주제들 간의 연관성을 파악한다.

1.2 선정 모델 및 브랜드

선정한 모델은 **Tesla Model S** 이다. 테슬라가 오토파일럿 기능을 앞세워 자율주행 시장을 선도하고 있기 때문이다. 해당 모델에 적용된 pattern recognition/computer vision 관련 요소기술은 다음과 같다.

- **Autopilot**: 자율주행 기능
- **Autopark**: 주차 자동화 기능
- **Autosteer**: 자동 조향 기능

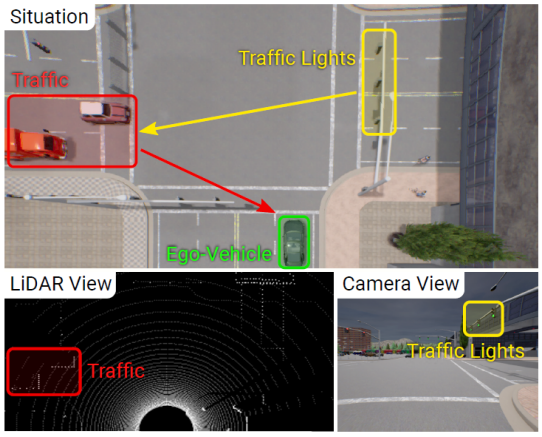
1.3 선정 기술 및 논문

자율주행 기능의 요소기술인 물체 검출과 교통신호 인식을 선정하여 관련 논문을 찾아보았다. [1, Multi-modal fusion transformer for end-to-end autonomous driving] 및 [2, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors] 논문을 선정하였다. 첫 번째 논문을 선정하게 된 이유는 자율주행 시스템의 구성요소 중 하나인 perception에 대한 연구이면서, 필자의 연구분야인 multi-modal fusion에 대한 연구이기 때문이다. 두 번째 논문을 선정하게 된 이유는 자율주행 시스템의 구성요소 중 하나인 perception에 대한 연구이면서, 영상처리의 주요한 분야 중 하나인 object detection에 대한 연구이기 때문이다.

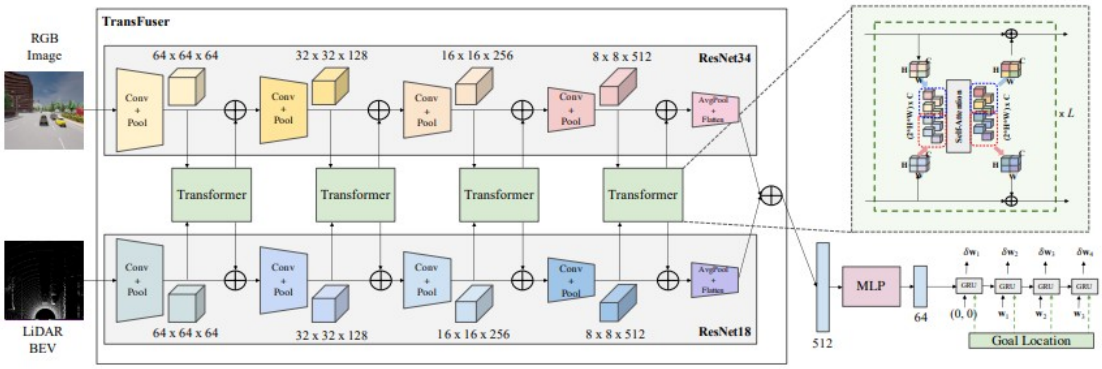
2 논문 요약

2.1 Multi-modal fusion transformer for end-to-end autonomous driving

이 논문은 그림 1 의 상황처럼 라이다(LiDAR : Light Detection And Ranging) 센서로 얻을 수 있는 주변의 차량의 위치에 따른 교통정보와 카메라로 얻을 수 있는 신호기에 따른 교통정보가 다른 경우,



〈그림 1〉 논문에서 해결하려는 문제 상황



〈그림 2〉 Transfuser 구조

두 센서로부터 얻을 수 있는 정보를 결합하여 차량의 주행을 제어하는 것을 목적으로 한다. 그림 2는 Transfuser의 구조를 보여준다. 두 센서의 출력으로부터 Resnet 구조[3]를 이용하여 정보를 추출하는 과정에서, 각 layer의 출력단으로부터 추출된 정보를 Transformer[4]를 이용하여 결합하는 것을 확인할 수 있다.

2.1.1 Method

목적 논문에서 제시한 목표는 시내 도로 주행에서의 point-to-point navigation이다. point-to-point navigation은 차량이 목표지점까지 waypoint를 따라 교통법규를 지키면서 다른 차량과의 상호작용을 하며 완주하는 것을 의미한다. 이를 달성하기 위한 방법으로 강화학습 기법 중 하나인 Imitation Learning을 채용하였다. Imitation Learning은 전문가가 직접 주행한 데이터를 따라하도록 agent의 policy를 학습하는 것을 의미한다.

입출력 자율주행 오픈소스 시뮬레이터 CARLA[5]에 있는 urban 가상환경에서 수집한 데이터를 입출력으로 사용하였다. **입력**은 두 가지로, 카메라와 라이다 센서의 출력이다. 카메라로부터 얻은 영상의 왜곡을 줄이기 위해 이미지 입력의 중앙을 잘라내어 $256 \times 256 \times 3$ 크기로 사용했다. LiDAR 센서의

출력 또한 주변부분의 정보를 기반으로 256 × 256 × 2 사이즈로 잘라내어 사용하였다. 채널의 한쪽은 지면 위, 한쪽은 지면 아래를 의미한다. **출력**은 PID controller로 차량을 제어하기 위해 4개의 waypoint $\{w_t = (x_t, y_t)\}_{t=1}^{T}$ 로 설정했다.

모델 모델은 그림 2에서 두 가지 부분으로 나눌 수 있다. 첫째는 Resnet과 Transformer를 이용하여 구성한 Multi-Modal Fusion Transformer(Transfuser)이고, 둘째는 MLP와 GRU로 구성된 Waypoint Prediction Network이다. 먼저 Transfuser의 동작을 살펴보자. 전반적인 동작은 subsection 2.1 의 표제 문단에서 작성하였으므로 Transformer의 적용 방법만을 확인한다. Transformer로는 GPT 모델을 사용하였다.

1. 라이다 입력과 영상 입력에 대해 컨볼루션과 풀링을 진행하여 채널 수를 늘리면서 특징을 추출한다.
2. 특징의 크기를 average pooling을 통해 8x8 로 줄인다.
3. 각 특성맵을 concat하여 16x8 크기의 특성맵을 만든다.
4. velocity를 value로, 16x8 특징을 key와 query로 사용하여 self attention(dot product attention)을 적용한다.
5. bilinear interpolation을 통하여 원본 영상의 크기로 확대한다.
6. 이전 단의 특성맵과 attention을 통해 추출한 특성맵을 더하여 특성맵을 업데이트한다.

둘째로 Waypoint Prediction Network의 동작을 살펴보자.

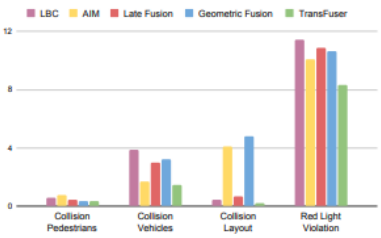
1. activation이 ReLU인 3-Layer Perceptron을 이용하여 1 × 1 × 512 크기의 특징을 1 × 1 × 64 크기의 특징으로 압축한다.
2. 압축된 특징을 GRU에 입력하여 4개의 waypoint를 예측한다.

학습 모델의 Loss는 전문가의 주행 데이터와의 L2 distance를 사용하였다.

2.1.2 result

Method	Town05 Short		Town05 Long	
	DS ↑	RC ↑	DS ↑	RC ↑
CILRS [16]	7.47 ± 2.51	13.40 ± 1.09	3.68 ± 2.16	7.19 ± 2.95
LBC [8]	30.97 ± 4.17	55.01 ± 5.14	7.05 ± 2.13	32.09 ± 7.40
AIM	49.00 ± 6.83	81.07 ± 15.59	26.50 ± 4.82	60.66 ± 7.66
Late Fusion	51.56 ± 5.24	83.66 ± 11.04	31.30 ± 5.53	68.05 ± 5.39
Geometric Fusion	54.32 ± 4.85	**86.91** ± 10.85	25.30 ± 4.08	**69.17** ± 11.07
TransFuser (Ours)	**54.52** ± 4.29	78.41 ± 3.75	**33.15** ± 4.04	56.36 ± 7.14
Expert	*84.67 ± 6.21*	*98.59 ± 2.17*	*38.60 ± 4.00*	*77.47 ± 1.86*

(a) **Driving Performance.** We report the mean and standard deviation over 9 runs of each method (3 training seeds, each seed evaluated 3 times) on 2 metrics: Route Completion (RC) and Driving Score (DS), in Town05 Short and Town05 Long settings comprising high densities of dynamic agents and scenarios.



(b) **Infractions.** We report the mean value of the total infractions incurred by each model over the 9 evaluation runs in the Town05 Short setting.

〈그림 3〉 Transfuser 실험 결과

그림 3 에서 Transfuser의 실험 결과를 확인할 수 있다. 왼쪽 장표는 주행 완료도와 안전운전 정도를 나타낸 것이고, 오른쪽 막대그래프는 사고율을 나타낸 것이다. 비교한 방법들에 비해 주행의 완료도도 좋아지고, 안전운전의 정도도 좋아졌고, 사고율도 줄어든 것을 확인할 수 있다.

2.1.3 Discussion

이 논문은 앞서 section 2 의 첫 문단에서 밝힌 바와 같이 교통신호 인식과 차량 및 보행자 인식을 동시에 처리하는 방식을 제안하였다. 이는 수업에서 다룬 패턴인식 과정과는 달리 raw data로부터 feature를 구하는 과정마저 network에게 맡긴 것으로 볼 수 있지만, 다차원의 입력의 차원을 줄이고, 그로부터 원하는 결과물을 얻는다는 점에서 수업의 내용과 맥락을 같이한다. 수업의 6장에서 다룬 neural network의 대부분의 내용을 포함하고 있다.

- criterion function으로 2차 minkowski distance를 사용
- multi-layer perceptron을 사용하여 정보 압축

2.2 YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors

asdf

참고 문헌

[1] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 7073–7083, 2021.

[2] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR,* vol. abs/1512.03385, 2015.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR,* vol. abs/1706.03762, 2017.

[5] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.