Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

School of Industrial and Management Engineering, Korea University

Jin Hyeok Park

KOREA UNIVERSITY

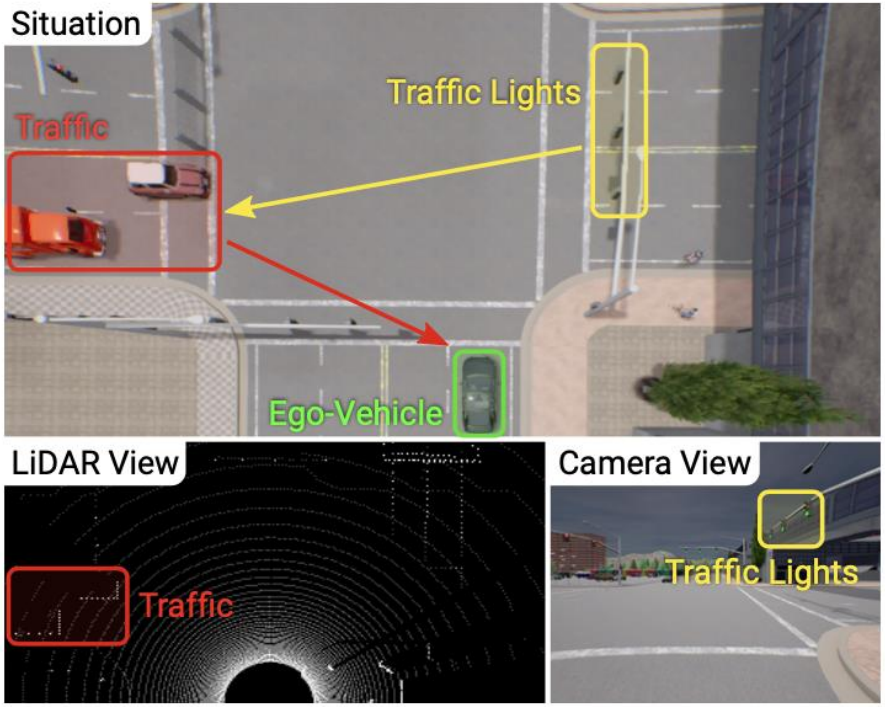
DMQA

Contents

Introduction

Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

- ❖ 기존에는 LiDAR 데이터 또는 Camera 이미지 만을 가지고 model을 설계함
- ❖ 단일 데이터 model의 경우 adversarial scenarios(운전시 생기는 변수)에 대응하기 어려움
- ❖ Adversarial scenarios를 해결하기 위해 LiDAR 데이터와 Camera 이미지를 가지고 Multi-Modal에 적용



Introduction

Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

- 2021년 9월 30일 기준 7회 인용
- Transformer를 Multi-Modal에 적용한 연구

Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

Aditya Prakash[*1] Kashyap Chitta[*1,2] Andreas Geiger[1,2]
[1]Max Planck Institute for Intelligent Systems, Tübingen [2]University of Tübingen
{firstname.lastname}@tue.mpg.de

Overview of paper

Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

- Task: Point-to-point navigation
- Point-to-point: 목표지점까지 waypoint를 따라 사고 없이 완주하는 것
- 학습 방식: Imitation learning
- Imitation learning을 적용한 이유: 고차원 데이터를 처리하고 연속된 action을 하는 경우에 적합함

$$\mathcal{D} = \{(\mathcal{X}^i, \mathcal{W}^i)\}_{i=1}^{Z}$$

- 가상환경에서 Expert가 주행하며 데이터 수집
- X는camera 이미지, LiDAR의 point cloud로 구성
- 이미지와 point cloud를 넣으면 T개의 Waypoint가 출력됨

Overview of paper

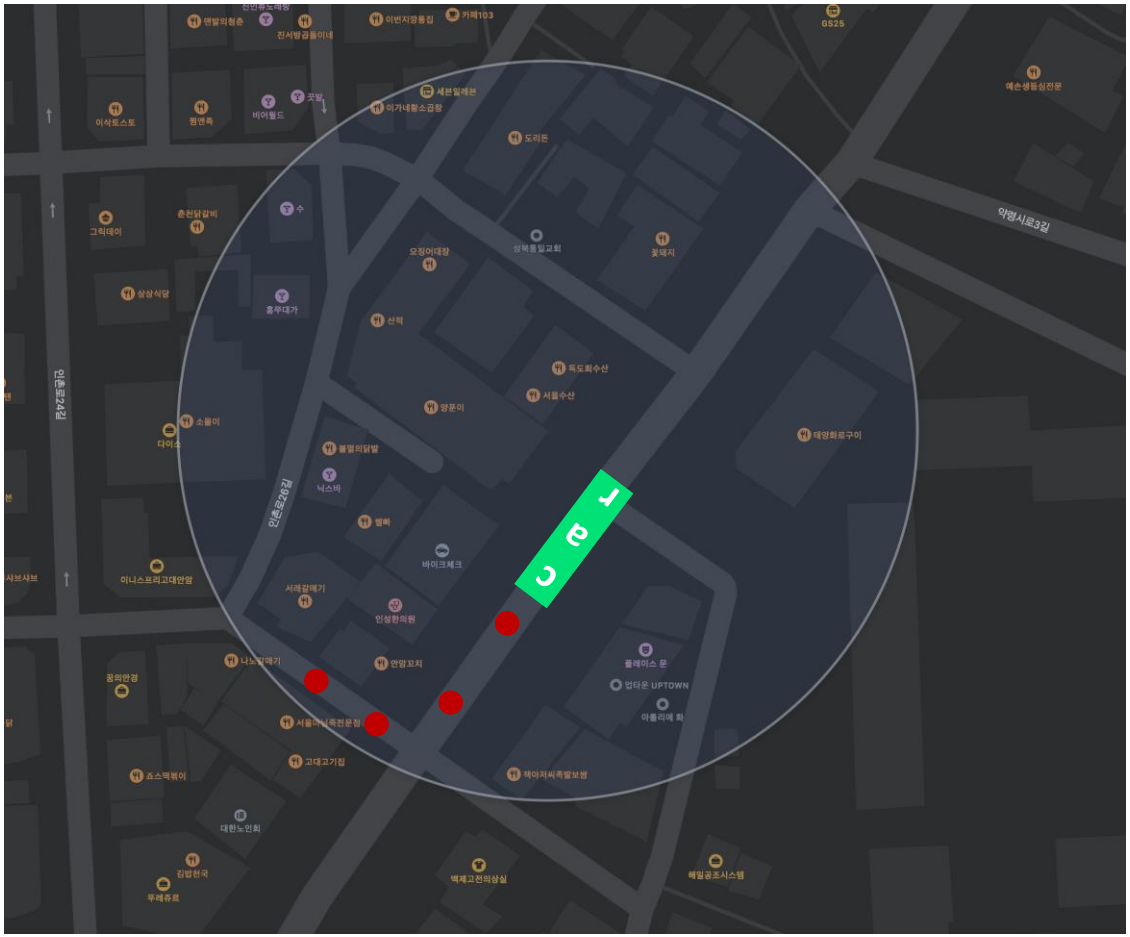
Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

- Camera 이미지
  - 256×256×3사이즈의 이미지 데이터
- LiDAR Point Cloud
  - LiDAR를 통해 얻은 Point cloud를 2D데이터로 변환한 256×256×2 사이즈의 데이터로 구성

Overview of paper

Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

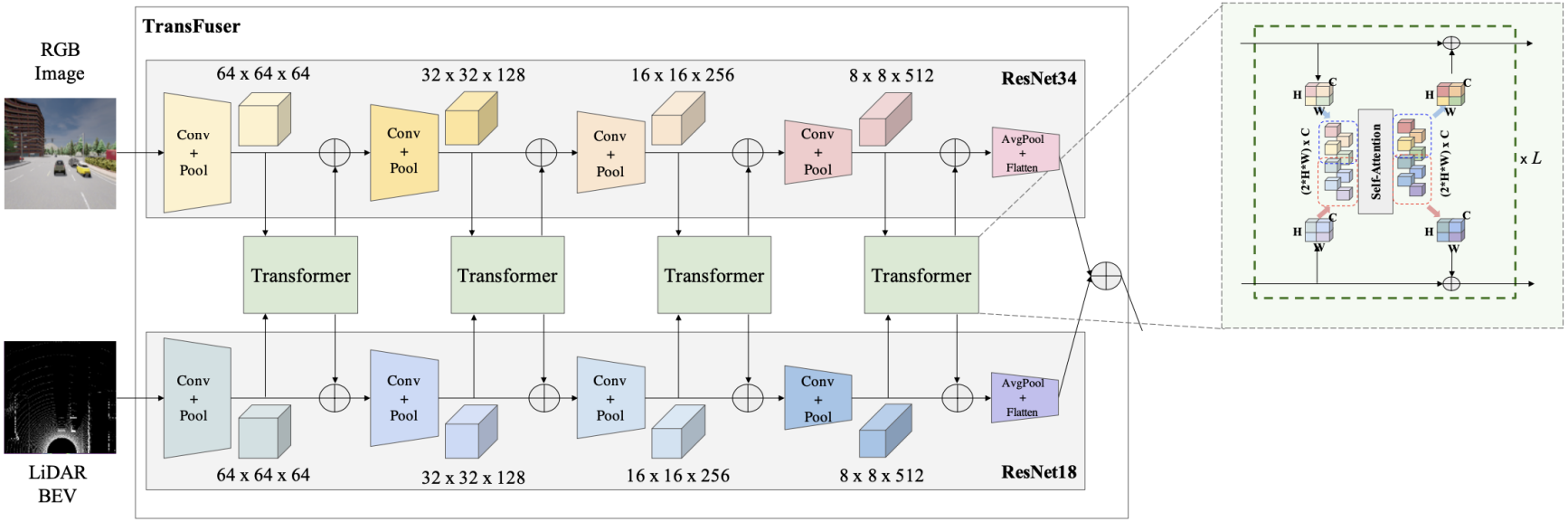
- ❖ Output Representation
  - ➢ BEV space상에서 (x, y)의 좌표 형식을 가짐



Architecture

Multi-Modal Fusion Transformer

- ❖ Conv + Pool을 통한 feature map extraction
- ❖ 추출한 feature map 사이즈를 8×8로 압축 후 각 데이터의 feature map을 16×8 사이즈의 feature map으로 합친 후 transformer의 입력 값으로 사용
- ❖ 16×8 feature map을 position embedding 후 linear layer를 통해 projection
- ❖ Transformer에 입력하여 self-attention 연산 후 attention이 반영된 임베딩 벡터를 데이터 별로 나눔
- ❖ 8×8 사이즈의 벡터를 압축하기 전의 크기로 scale up한 뒤 초기의 feature map과 원소끼리 더함

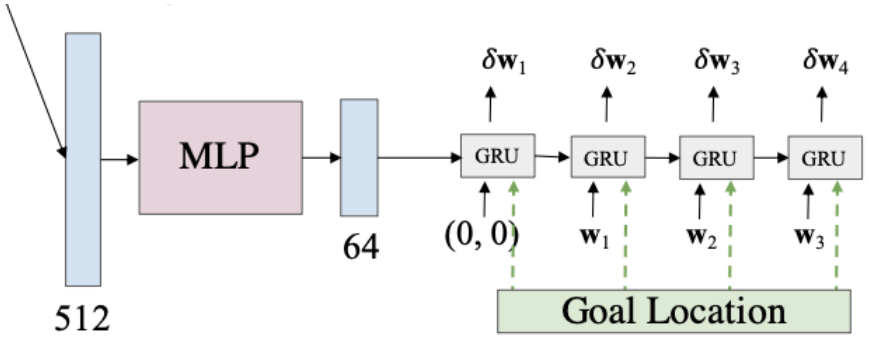


DMQA

Architecture

Waypoint Prediction Network

- Multi-Modal Fusion Transformer를 통해 나온 1×1×512 벡터를 1×1×64 벡터로 압축함
- 초기 입력 데이터는 x = (0,0)으로 (0,0)을 기준으로 미래의 4시점의 waypoint를 예측함
- 1×1×64 벡터를 가지고 hidden state를 초기화 함



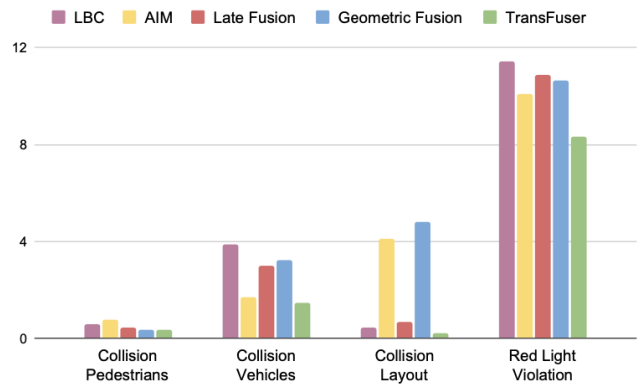
Result

Experiments

- DS(Driving Score): RC에 infraction multiplier(충돌, 이탈, 차선 침입, 신호 위반)를 곱한 값
- RC(Route Completion): 주행 경로를 몇 %나 주행했는지 알려주는 수치
- 한가지 데이터만 입력 받는 CILRS, LBC, AIM에 비해 두가지 데이터를 입력으로 받는 모델 성능이 뛰어남
- TransFuser는 Late Fusion과 Geometric Fusion에 안전운전을 더 잘하는 결과를 보임
- 전반적으로 다른 모델에 비해 사고율이 매우 낮음

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.

Result

Experiments

- ❖ 카메라 이미지와 LiDAR의 상호보완성
  - ➢ 이미지 토큰의 62.75%가 attention을 가장 많이 한 5개의 token이 LiDAR에서 나온 토큰
  - ➢ LiDAR 토큰의 78.45%가 attention을 가장 많이 한 5개의 token이 이미지에서 나온 토큰

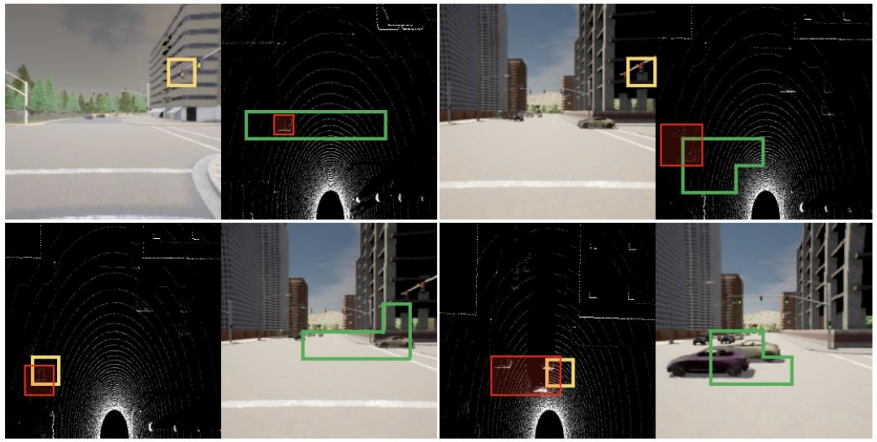


Figure 3: **Attention Maps.** For the yellow query token, we show the top-5 attended tokens in green and highlight the presence of vehicles in the LiDAR point cloud in red. TransFuser attends to the vehicles and traffic lights at intersections, albeit at a slightly different location.

Conclusion

❖ Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

- 카메라 이미지와 LiDAR point cloud에 해당하는 두 종류의 데이터를 사용한 모델
- 두 종류의 데이터를 사용하기 위한 Multi-Modal을 적용한 Transformer
- 해당 모델에서 다른 센서 데이터를 추가하거나 다른 AI task에 적용 가능한 연구


Thank You