패턴인식 보고서

120220210 고재현

2023년 3월 8일

1 개요

1.1 목적

- 최근 대두되는 디지털 휴먼 관련 기반 기술들을 조사한다.
- 해당 기반 기술을 구현하기 위해 핵심적인 요소들에 대해 고려해 본다.

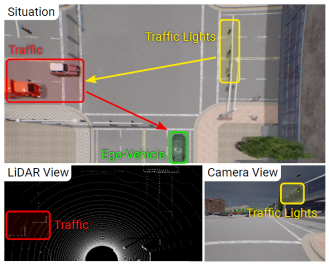
1.2 요소 기술

한유아와 같은 디지털 휴먼을 만들기 위해서는 얼굴 합성 기술, 모션 합성 기술, 음성 합성 기술이 필요하다. 사실적인 표현을 위한 캐릭터 모델링 및 렌더링 기술도 중요하지만, 인공지능 기반의 기술들만 서술하도록 한다. 얼굴 합성 및 모션 합성 기술은 주로 키포인트(랜드마크 등)를 생성한 뒤, 해당 키포인트를 기반으로 모델링을 덮어씌우는 방식으로 수행[1]되거나, 원본 영상을 변형하여 각 영상 프레임을 생성하는 방식으로 수행된다.

1.3 선정 기술 및 논문

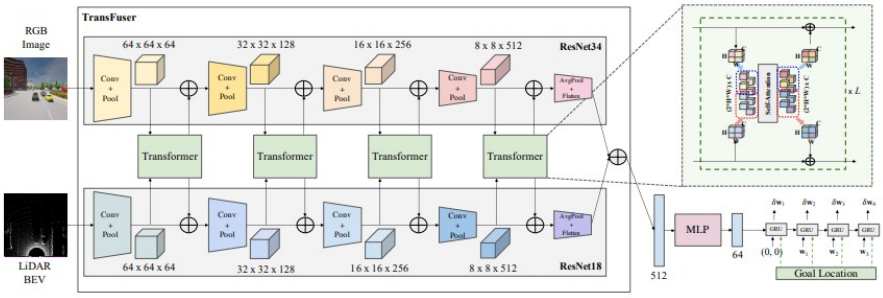
2 논문 요약

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



〈그림 1〉 논문에서 해결하려는 문제 상황

이 논문은 그림 1의 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, 두 센서로부터 얻을 수 있는 정보를 결합하여 차량의 주행을 제어하는 것을 목적으로 한다. 그림 2는



〈그림 2〉 Transfuser 구조

Transfuser의 구조를 보여준다. 두 센서의 출력으로부터 Resnet 구조[?]를 이용하여 정보를 추출하는 과정에서, 각 layer의 출력단으로부터 추출된 정보를 Transformer[?]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[?]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 $256 \times 256 \times 2$ 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.

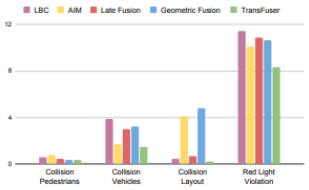
1. activation이 ReLU인 3-Layer Perceptron을 이용하여 $1 \times 1 \times 512$ 크기의 특징을 $1 \times 1 \times 64$ 크기의 특징으로 압축한다.
2. 압축된 특징을 GRU에 입력하여 4개의 waypoint를 예측한다.

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 3에서 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 2의 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

이 논문은 2015년부터 실시간 object Detection, segmentation 등의 분야에서 좋은 성적을 보이고 있는 YOLO(You Only Look Once)의 최신 버전인 YOLOv7을 제안한다. 이 논문의 목적은 영상으로부터의 실시간 물체 탐지이며 SOTA(State-of-the-art)를 달성하기 위해 다음과 같은 방법들을 제안한다.

1. 모델 구조 관점

a) Extend ELAN
b) compound scaling

2. bag-of-freebies: 학습 과정에서의 부담을 추론 과정에서는 추가 계산 비용이 들지 않으면서 네트워크의 성능을 향상시키는 기법 관점
   a) Planned re-parameterized convolution
   b) Coarse for auxiliary and fine for lead loss
   c) Batch normalization,

2.2.1 Method

Architecture 성능 개선을 위해 사용된 구조는 두 가지이다. 첫째는 Extended efficient layer aggregation networks(ELAN)이고, 둘째는 Model scaling for concatenation-based models(compound scaling)이다. gradient decent 관점에서 보면, 모델의 학습에서 batch 하나에 대한 학습 속도는 gradient path의 길이에 비례한다. ELAN은 이 관점에서 "가장 효율적인 네트워크를 구성하려면 어떻게 해야 할까?" 라는 질문을 통해 가장 긴 gradient path와 가장 짧은 gradient path를 조정함으로써 학습 속도를 높이는 방법을 제안했다. 그러나 이 방식조차 block을 매우 많이 쌓을 경우 stable state가 무너지는 문제가 생겼다. 이를 해결하기 위해 YoLoV7에서는 Expand, shuffle, merge cardinality를 이용한 Extended ELAN이라는 구조를 제시하였다. 이 구조를 통해 scaling에 따라 transition layer는 변경되지 않고 computational block의 크기만 조절할 수 있게 되었다.

ResNet[?] 등에서 사용된 Model scaling은 모델의 일부 특성을 조절하고, 추론 시간을 원하는 대로 조절하기 위해 다양한 크기의 모델을 학습시키고, 이를 조합하는 방법이다. 그런데 앞서 제시한 E-ELAN과 같은 concatenation 기반의 결합은 scaling에 따라 transition layer의 깊이뿐만 아니라 너비(입력 차원)도 함께 달라지기 때문에, 이를 함께 고려하는 방식으로 compound scaling을 제안하였다. Computation block에만 depth scaling을 수행하고 나머지 block은 이 변경에 따라 width scaling을 수행하는 방식이다.

Trainable bag-of-freebies

Planned re-parameterized convolution RepConv에서 제안된 re-parameterization을 ResNet 등에서도 활용하기 위해 제안된 방법이다. re-parameterization은 학습 시에는 병렬적으로 여러 convolution layer 및 Batch normalization layer를 학습한 뒤 추론시에는 이를 하나로 합치는 방식이다. 이를 위해 identity connection을 사용하게 되는데, ResNet에 이를 적용하게 되면 Residual connetion을 파괴하게 된다. 이를 해결하기 위해 YoloV7에서는 identity connection을 사용하지 않는 방식의 RepConvN을 제안하였다.

Coarse for auxiliary and fine for lead loss Deep supervision 기법을 사용하면 Lead head와 Auxiliary head의 label을 동시에 만들 수 없는데, lead head prediction으로 두 head를 동시에 학습하거나 lead head prediction을 guidance로 사용하는 방식으로 해결했다. 두번째 방법의 label은 coarse-to-fine hierarchical labels이라고 부른다. coarse label은 positive assignment constraint를 완화시킨 label이며, Fine label은 Lead head guided label과 같은 방식으로 lead head의 prediction으로 만든 label이다.

2.2.2 Result

COCO dataset으로 모델의 성능을 평가하였다. 지면상 한계로 실험 결과는 보고서에서는 생략하였다. 기존 모델들과의 성능을 비교하였더니 수행 시간 및 정확도의 trade-off 면에서 가장 좋은 성능을 보여주었다.

2.2.3 Discussion

수업 시간에 다룬 Neural Network의 학습 속도(Gradient Decent 관점에서)를 개선하는 방법으로 Gradient Path를 조절하는 방법이 제안되었다. 또한 학습 과정에서 labeling은 매우 번거롭고 어려운 일인데, 특히 auxiliary head의 label을 구하는 과정에서 coarse-to-fine hierarchical labels를 도입하여 해결하였다. 모델의 복잡도와 추론 시간은 비례하는데, 이를 개선하기 위해 모델의 구조를 개선하는 과정 또한 포함되어 있다.

참고 문헌

[1] A. Richard, M. Zollhöfer, Y. Wen, F. D. la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," *CoRR,* vol. abs/2104.08223, 2021.