다수 자율주행용 센서 기반 멀티모달 딥러닝을 통한 레이더 고스트 표적 탐지

이인균, 이현희, 신동준
한양대학교 융합전자공학과

rbstkfka@hanyang.ac.kr, fly4hyun@hanyang.ac.kr, djshin@hanyang.ac.kr

Radar Ghost Target Detection via Multimodal Deep Learning Based on Multiple Self-Driving Sensors

Ingyun Lee, Hyunhee Lee, Dong-Joon Shin
Dept. of Electronic Engineering, Hanyang University

요 약

본 논문은 레이더에서 고스트 표적(ghost target) 탐지를 위해 기존 레이더, 라이다를 이용하는 멀티모달 딥러닝(multimodal deep learning) transformer 모델을 발전시켜 레이더, 라이다, 카메라를 이용하는 새로운 멀티모달 딥러닝 모델을 제안한다. 제안된 딥러닝 모델에서 카메라, 레이더 및 라이다의 융합(fusion) 방법을 제안하고 제안된 딥러닝 모델이 기존 딥러닝 모델보다 향상된 성능을 보임을 확인한다.

Ⅰ. 서 론

자율주행 시장이 급속도로 커지면서 인공지능을 이용한 자율주행은 매우 중요한 기술이 되었다. 특히 자율주행에 카메라, 레이더, 라이다 센서와 같이 다양한 모달리티(modality)가 사용되면서 이를 융합하는 멀티모달 딥러닝(MMDL)의 중요성이 커지고 있다. 자율주행기술에 있어 전파의 잡음, 반사 등의 다양한 신호 왜곡으로 레이더에서 고스트 표적이 나타나며 이로 인해 차량 사고가 발생할 수 있어 레이더 고스트 표적 탐지가 매우 중요하다. 이를 위해 기존 연구에서는 레이더와 라이다를 이용한 멀티모달 transformer 모델을 제안하였지만[1] 자율주행 센서에 있어 가장 많이 사용되는 카메라를 이용하지 않았다.

본 논문에서는 레이더 고스트 표적 탐지를 위해 레이더, 라이다, 카메라를 이용하는 멀티모달 딥러닝 모델을 제안한다. 또한 제안한 모델이 기존 멀티모달 transformer 모델보다 성능이 좋음을 확인한다.

Ⅱ. 기존 멀티모달 transformer 딥러닝 모델

기존에 제시된 멀티모달 transformer 모델은 그림 1 처럼 레이더와 라이다의 인코딩 과정 및 MMA (multimodal attention)와 SA (self-attention)로 구성된다[1]. MMA와 SA를 수행하기 위해서는 Q(query), K(key), V(value) 행렬이 필요하다.

$$(Q, K, V) = (FW_Q, FW_K, FW_V). \quad (1)$$

F는 크기가 $N \times d_{in}$인 feature map 행렬이고, W_Q, W_K, W_V는 각각 Q, K, V 행렬을 만드는데 이용되는 학습 파라미터이다. 행렬의 크기는 $Q, K \in \mathbb{R}^{N \times d_{qk}}$, $V \in \mathbb{R}^{N \times d_{out}}$, $W_Q, W_K \in \mathbb{R}^{d_{in} \times d_{qk}}$, $W_V \in \mathbb{R}^{d_{in} \times d_{out}}$ 이며 여기서 N 은 센서가 탐지한 포인트의 수이고 d_{qk}, d_{in}, d_{out}은 값을 설정해 주어야 하는 변수이다. SA 는 소프트맥스 함수($softmax(\cdot)$)를 이용하여 다음과 같이 계산된다.

$$SA(F) = softmax\left(\frac{QK^T}{\sqrt{d_{qk}}}\right)V. \quad (2)$$

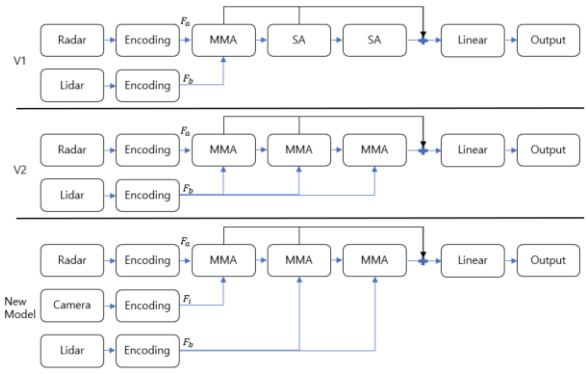

그림 1. 기존 MMDL 모델(V1, V2)과 제안한 모델 구조도.

MMA 를 구하기 위해 먼저 CMA (cross-modal attention)를 구해야 한다. CMA를 구하기 위해 레이더와 라이다를 각각 인코딩하여 추출한 F_a 와 F_b 가 사용된 경우 CMA는 식(3)과 같이 구해진다.

$$CMA(F_a, F_b) = softmax\left(\frac{Q_a {K_b}^T}{\sqrt{d_{qk}}}\right)V_b. \quad (3)$$

Q_a는 F_a를 이용하여 구한 query 행렬이며 K_b, V_b는 F_b를 이용하여 구한 key, value 행렬이다. MMA 는 SA 와 CMA 를 통해 구한 행렬을 서로 연접하여($concat(\cdot)$) 최종적으로 구하게 된다.

$$MMA(F_a, F_b) = concat\big(SA(F_a), CMA(F_a, F_b)\big). \quad (4)$$

MMA 를 통해 레이더 포인트와 라이다 포인트 간의 관계에 대한 정보를 담은 F를 추출하게 된다. 그림 1 과 같이 각각의 MMA와 SA 단계에서 추출한 F를 연접한 후 완전연결계층(linear)을 통해 얻은 output 으로 레이더 고스트 표적 여부를 판단한다.

Ⅲ. 카메라 센서를 추가한 MMDL 모델

(1) 카메라를 추가한 MMDL 모델 구조

본 논문이 제안하는 카메라를 추가한 MMDL 모델은 그림 1 과 같이 레이더 신호를 인코딩한 F_a와 카메라 이미지를 인코딩한 F_i로 첫 번째 MMA를 구하는 과정을

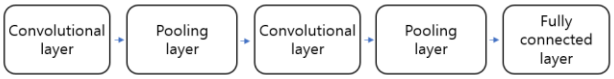

그림 2. 이미지의 픽셀 데이터 인코딩 과정 블록도.

거친다. 이 과정에서 하나의 이미지의 각 부분들과 레이더 포인트의 관계 정보를 담은 F를 MMA를 통해 추출한다. 라이다 신호를 인코딩한 F_b는 두 번째와 세 번째 MMA를 구하는 과정에서 이용하게 된다.

(2) 이미지 데이터 인코딩 과정

이미지 데이터는 픽셀 데이터와 카메라 번호, ΔT(시간 변화량) 데이터로 나뉘며 픽셀 데이터는 합성곱 신경망(convolutional neural network)을 통해 그리고 카메라 번호와 ΔT 데이터는 완전연결계층을 통해 인코딩하여 각각 F_{C_i}와 F_{F_i}를 추출한다. RGB 픽셀 데이터의 인코딩 과정은 그림 2 와 같이 합성곱 계층, 풀링 계층 및 완전연결계층으로 구성된다. 첫 번째 합성곱 계층에서는 $channel = 16$, $kernel\ size = 3$, $stride = 2$, $padding = 1$로 두 번째 합성곱 계층에서는 $channel = 1$로 한다. 풀링 계층에서는 최대 풀링(max pooling)을 사용하고 $kernel\ size = 2$, $stride = 1$로 한다. 완전연결계층은 한 층으로 구성되고 F_{C_i}의 크기는 $N_i \times a_1$이다. 카메라 번호, ΔT 데이터의 인코딩 과정은 한 개의 완전연결계층으로 구성되며 F_{F_i}의 크기는 $N_i \times a_2$이다. N_i는 F_{C_i}와 F_{F_i}의 행의 크기이며 a_1, a_2는 각 feature 의 length 이다. F_i는 식(5)와 같이 F_{C_i}와 F_{F_i}를 연접하여 추출한다.

$$F_i = concat(F_{C_i}, F_{F_i}),\ F_i \in \mathbb{R}^{N_i \times (a_1 + a_2)}. \quad (5)$$

(3) 레이더와 라이다 데이터 인코딩 과정

제안한 MMDL 모델에서 레이더와 라이다 인코딩 과정은 기존 MMDL 모델과 동일하다. 레이더와 라이다 데이터는 모두 좌표 값과 그 외의 값들로 나뉜다. 레이더와 라이다의 좌표 값 데이터는 각각 두 개의 완전연결계층을 통해 F_{C_a}와 F_{C_b}를 추출한다. 좌표 이외 값들의 데이터는 각각 두 개의 완전연결계층을 통해 F_{F_a}와 F_{F_b}를 추출한다. F_{C_a}, F_{C_b}, F_{F_a}, F_{F_b}의 크기는 각각 $N_a \times d_c$, $N_b \times d_c$, $N_a \times d_{F_a}$, $N_b \times d_{F_b}$ 이다. N_a, N_b 는 레이더와 라이다 포인트의 수이고 d_c, d_{F_a}, d_{F_b}는 각 feature 의 length 이다. 식(6)을 통해 레이더와 라이다 데이터를 각각 인코딩하여 F_a와 F_b를 추출한다.

$$F_a = concat(F_{C_a}, F_{F_a}),\ F_a \in \mathbb{R}^{N_a \times (d_c + d_{F_a})},$$
$$F_b = concat(F_{C_b}, F_{F_b}),\ F_b \in \mathbb{R}^{N_b \times (d_c + d_{F_b})}. \quad (6)$$

전체 인코딩 과정에서 합성곱 계층은 ReLU 함수가 사용되고 완전연결계층에서는 ReLU 함수와 배치 정규화(batch normalization)가 수반된다.

Ⅳ. 실험 결과

nuScenes[2] 데이터셋 중 장면(scenes) 10 개를 트레이닝 데이터로 장면 2 개를 테스트 데이터로 이용한다. 레이더 데이터는 레이더 포인트 1,000 개의 좌표, 속도, RCS, $pdh0$, $dynProp$, $invalid\ state$, ΔT로 구성되고 라이다 데이터는 라이다 포인트 16,000 개의 좌표, 반사 강도 I, ΔT로 구성된다. 이미지 데이터는 시간 간격이 1 초인 것을 사용하며 픽셀, 카메라 번호, ΔT로 구성된다. 픽셀의 텐서 크기는 $6 \times 45 \times 80 \times 3$이고 카메라 번호는 원-핫 인코딩한 것을 사용한다. Ground truth 생성 방법은 [3]의 방법을 사용한다. 학습률은 0.001, 배치 크기는 4 로, $d_c = 32$, $d_{F_a} = 64$, $d_{F_b} = 64$, $N_i = 1140$, $a_1 = 32$, $a_2 = 32$, $d_{qk} = 64$ 로 설정한다.

V1	V2	New Model
MMA(128)	MMA(128)	MMA(128)
SA(128)	MMA(256)	MMA(256)
SA(128)	MMA(512)	MMA(512)
FC(128,0.5)	FC(512,0.5)	FC(512,0.5)
FC(2,0.5)	FC(128,0.5)	FC(128,0.5)
	FC(2,0.5)	FC(2,0.5)

그림 3. 딥러닝 모델 구조 및 변수 크기.

표 1. 제안한 모델과 기존 모델의 성능 비교

Model	AUC-ROC
New Model	77.668%
V1	71.962%
V2	71.699%

그림 3 은 $MMA(d_{out})$와 $SA(d_{out})$를 나타내고, d_{out}은 결괏값의 차원이다. $FC(l, dp)$는 완전연결계층에서 l개의 노드와 dropout 비율을 의미한다.

표 1 과 같이 제안한 MMDL 모델의 AUC-ROC 값은 77.668%로 기존 V1, V2 보다 각각 5.706%, 5.969% 높다. 기존 모델에서는 노드 수와 텐서 크기, 학습률을 조절하며 가장 높은 AUC-ROC 값을 얻었고, 제안한 모델에서는 성능 비교를 위해 이미지 인코딩 부분과 첫 번째 MMA를 구하는 부분을 제외한 다른 것은 기존 모델 V2 와 동일하게 설정하여 실험했다. 기존 모델 V2 와 제안한 모델의 학습 파라미터 수는 각각 742,658 개, 739,603 개로 거의 동일하게 맞추면서 모델의 크기는 작아졌지만 성능은 좋아진 것을 확인하였다.

Ⅴ. 결론

본 논문에서는 기존 레이더와 라이다를 이용한 멀티모달 transformer 딥러닝 모델에 카메라를 추가하여 레이더 고스트 표적을 찾는 새로운 멀티모달 transformer 딥러닝 모델을 제안하였으며 제안한 딥러닝 모델이 기존 딥러닝 모델보다 성능이 향상된 것을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1A2C2014057)

참 고 문 헌

[1] L. Wang, S. Giebenhain, C. Anklam, and B. Goldluecke, “Radar ghost target detection via multimodal transformers,” *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7758-7765, Oct. 2021.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 11621-11631, 2020.

[3] F. Nobis, “Autonomous driving: Radar sensor noise filtering and multimodal sensor fusion for object detection with artificial neural net-works,” MS Thesis, *Technical University of Munich*, 2019.