자율주행을 위한 딥러닝 기반의 라이다 카메라 센서융합 기술

김재겸, 고준호, 김예철, 최준원 (한양대학교)

1. 서론

자율주행 차량에는 카메라, 라이다, 레이더, 초음파 센서 같은 다양한 센서가 장착되어 있다. 이러한 센서들을 사용하여 주변의 물체를 식별하고 동적인 환경에서 주변 상황을 이해하게 된다. 주변 상황에는 도로나 신호등과 같은 정적인 물체와 차량이나 보행자와 같은 동적인 물체들이 있다. 특히 동적인 물체들은 항상 자율주행차량의 안전을 위협하기 때문에 이러한 동적인 물체들을 검출하는 기술이 중요하다. 이러한 물체검출 기술은 최근에 딥러닝이라는 머신러닝 방법이 출현하면서 높은 성능개선을 이루었다. 딥러닝 기법에서는 많은 양의 데이터를 사용하여 깊은 층을 갖는 신경망 구조를 학습시키게 된다. 특히 카메라 영상을 이용하여 머신러닝 작업을 수행하는 컴퓨터 비젼 분야에서는 딥러닝 구조 중 하나인 Convolution Neural Network (CNN)를 활용하여 기존 방법 대비 매우 높은 성능을 이루었고 이러한 비전 기술은 자율주행에서 적극적으로 활용되고 있다. CNN구조는 AlexNet[1]부터 시작해서 VGGNet[2], ResidualNet[3], DenseNet에 이르기까지 빠른 속도로 발전하였고 이러한 CNN 구조를 활용하여 특징값을 뽑고 이를 물체검출 및 영역분할, 고행상도 영상 복원 등 다양한 분야에 활용되고 있다.

물체 검출 기술은 자율주행의 인지기능에 있어 매우 중요한 기술이다. 자율주행 차량 주변의 존재하는 차량, 보행자, 사이클리스 등을 검출하고 이들이 어디에 존재하는 지 알아야 한다. CNN 기반 물체 검출은 크게 두 가지 종류로 나뉘는 데 이단계 검출법과 일단계 검출법이 있다. 이단계 검출법에서는 먼저 물체가 존재할 수 있는 부분과 크기를 표시하기 위한 박스를 찾는 작업이 수행된다. 하나의 영상에서 물체가 나올만한 곳에 여러 가지 제안된 박스를 찾아내면 각각의 박스들에 대해 어떤 종류의 물체가 있는지를 판단하는 물체분류 작업이 수행된다. 이러한 이단계 물체검출 기법에는 R-CNN[4], Fast R-CNN[5], Faster R-CNN[6], 그리고 Mask R-CNN[7]이 있다. 이러한 이단계 검출 기법은 두 가지 작업을 순차적으로 수행해야 하기 때문에 상대적으로 긴 연산시간을 갖는 단점이 있다. 한편, 일단계 검출 방법은 CNN을 이용해 찾은 특징값을 이용하여 물체가 존재하는 영역을 포함하는 박스를 찾는 작업과 박스가 잡은 물체를 분류하는 물체분류 작업을 하나의 네트워크에서 동시에 수행하게 된다. 이러한 기법에는 YOLO[8], SSD[9], RetinaNet[10]이 존재한다. 이 기법들은 상대적으로 적은 연산 시간으로 자율주행에 많이 사용된다.

딥러닝 기법이 카메라 기반 인지 성능에 많은 개선을 가져왔지만 일반적으로 자율주행 환경에서는 카메라만으로 주변 환경을 인지하는 데 한계가 존재한다. 예를 들면 날씨나 그림자 또는 물체에 의한 가림, 급격한 조도 변화 등은 카메라 기반의 물체 인식 성능에 큰 영향을 미치게 된다. 또한 카메라 영상으로는 자율주행을 위한 중요한 정보인 차량의 깊이정보, 차량 간의 거리, 그리고 차량의 지향 등의 3차원 정보를 정확히 얻기가 어렵다. 최근의 자율주행 연구에서는 라이다 센서를 카메라 센서와 함께 사용해 자율주행의 인지 성능을 높이는 연구가 활발히 진행되고 있다. 라이다 센서는 주변으로 레이저를 발사하고 반사되어 도달하는 시간을 측정하여 주변의 3차원 정보를 포인트 형태로 취득하게 된다. 일반적으로 라이다 센서는 직진성이 높은 레이저를 이용하기 때문에 높은 정밀도의 3차원 정보를 얻을 수 있다.

최근에 딥러닝 기법을 이용하여 3차원 포인트 데이터로부터 물체 검출 수행하는 기법이 많이 연구되고 있다. 많은 기법들에서 3차원 포인트 정보를 2차원 평면에 투영하여 CNN기반의 물체검출 기법을 이용하여 물체 검출 수행한다. 라이다 데이터를 2차원 평면에 투영하는 방법에는 탑뷰와 프론트뷰로 투영하여 표현하는 두 가지 방법이 있다. 탑뷰 영상은 도로를 위에서 내려다보는 시점으로 포인트 데이터를 투영하여 2차원 영상을 만드는 반면 프론트뷰 라이다 영상은 카메라뷰 또는 정면 시점으로 투영하여 2차원 영상으로 표현하게 된다. 이렇게 만들어진 2차원 라이다 데이터는 물체검출 네트워크의 입력으로 사용된다. 특히 탑뷰 영상을 사용하게 되면 차량의 진행방향 및 운동 속도

등의 정보를 추가적으로 추출할 수 있게 된다. 앞에서 언급한 카메라와 라이다 데이터는 서로 장단점이 다르기 때문에 서로 상호보완적인 관계에 있다. 따라서 이처럼 각각의 센서가 가진 문제점을 극복하고 신뢰성과 강인성이 높은 환경인지를 수행하기 위해서는 다양한 센서정보를 동시에 융합하는 센서융합 기술이 필요하다. 아직은 딥러닝 기반의 센서융합 기술은 초기 단계이나 최근에 센서융합을 위한 딥러닝 기법이 많이 제안되고 있다. 특히 영상과 음성신호를 융합하여 음성인식을 수행하거나 영상과 문자 정보를 합쳐서 영상에 대해 해석을 하는 분야에도 연구가 되고 있다. 자율주행 분야에서도 카메라 및 라이더 신호가 제공하는 정보를 융합하여 보다 나은 물체 검출 및 인지 기능을 수행하는 방법들도 활발히 제안되고 있다.

본 논문에서는 최근에 제안된 카메라 및 라이다의 센서융합기법을 위한 기존의 딥러닝 기법을 소개하고 카메라 라이더 센서융합에 대한 최근 연구결과를 소개하고자 한다. 카메라 혹은 라이더 신호의 퀄리티가 저하되는 경우에 강인한 센서 융합을 수행하는 방법을 소개하고 이를 수행하기 위한 새로운 딥러닝 아키텍처를 설명한다. 마지막으로 센서융합에서의 앞으로의 연구 방향과 해결하여야 할 문제점들에 대해 간단히 논의하도록 한다.

2. 기존의 센서융합 기반 물체검출 기술

센서융합은 서로 다른 구조와 분포를 갖는 센서 신호로부터 효과적으로 정보를 추출하여 원하는 작업을 수행하는 기술을 말한다. 서로 다른 센서 신호들의 서로 이질적인 특성으로 인해 효과적으로 정보를 융합하는 것이 어렵다. 일반적으로 센서융합은 신호를 미리 합쳐서 하나의 모델로 처리하는 이른융합 (early fusion)방법과 각각의 신호를 다른 모델로 처리한 후 마지막에 결과를 합치는 늦은융합 (late fusion)방법으로 나눌 수 있다. 일반적으로 이른융합은 입력으로부터 상관도가 높은 특징을 찾을 수 있는 구조이지만 낮을레벨에서 정보 융합이 이루어지기 때문에 이질감이 높은 입력들에 대해서는 효과적인 정보융합이 어렵다. 반면 늦은융합은 이미 처리된 결과를 합치는 방법이기 때문에 각 센서처리 결과의 신뢰도에 맞게 융합이 가능하며 구현 역시 간단하다는 장점이 있다. 하지만 서로 다른 센서 신호의 공통적인 구조를 반영하는 특징값을 사용하지 못한다는 단점이 있다.

최근에는 딥러닝 기술이 발전하면서 계층적인 구조를 통해 센서신호로 부터 추상적인 특징값을 추출하고 이들을 융합하여 추가적으로 딥러닝 처리를 하는 중간융합 (intermediate fusion)이 가능해 졌다. 이러한 접근방법의 효과는 여러 논문에서 제시되었다. [11]에서는 오디오 신호와 비디오 신호의 특징값을 뉴럴네트워크의 중간 계층에서 융합하여 음성인식의 성능을 높여줄 수 있음을 보였다. [12]에서는 RGB-D 영상에서 RGB 영상과 깊이영상을 각각 CNN 계층을 적용하여 특징맵을 추출하고 이를 융합한 후 fully connected 계층을 적용하여 물체인식 성능을 개선하였다. 최근에서는 자율주행에서의 카메라 라이더 센서융합 기법이 제시되었는데 MV3D [13]에서는 라이다 포인트 데이터를 탑뷰와 프론트뷰로 투영하여 라이다 영상을 만들고 카메라 영상과 융합하여 물체 검출을 수행하는 기법이 제안되었다. AVOD [14]는 MV3D 이 후에 제안된 3D 물체 검출 알고리즘로 높은 성능 개선정도를 보여주었다. ContFuse 네트워크[15]는 continuous convolution을 사용하여 카메라 이미지와 탑뷰 이미지를 융합해서 3D 물체 검출을 해준다. continuous convolution [16]을 통해 카메라 이미지의 연속하고 dense한 정보를 추출해서 탑뷰 이미지에 더해줘서 3D 물체검출의 성능을 높였다.

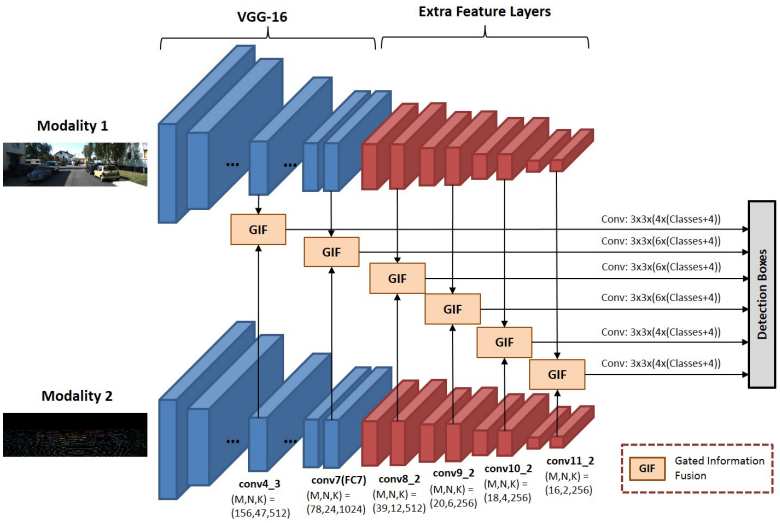
3. 제안하는 환경변화에 강인한 센서융합 기술

본 장에서는 센서융합 기법 중에서 환경변화에 강인한 구조를 갖는 물체검출 알고리즘을 제안하였다. 기존 융합기술들의 경우, 각 데이터를 concatenation 혹은 summation

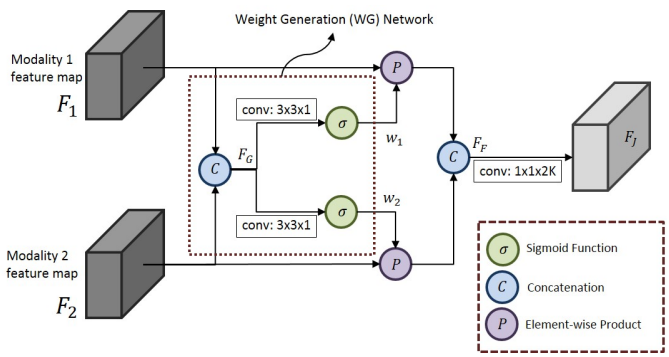
을 이용하여 단순히 융합하는 반면 제안하는 센서융합 기술에서는 gating 기법을 이용하여 각 특징맵의 중요도를 판단하여 융합한다. 이에 따라 입력 데이터에 퀄리티가 저하되는 등의 다양한 환경변화에도 강인한 융합을 이룰 수 있다. 제안하는 센서융합 기술은 물체검출 분야뿐 아니라, segmentation, recognition 분야 등 여러 데이터의 융합을 필요로 하는 분야에 모두 적용이 가능하다.

제안하는 물체검출 네트워크의 전체 네트워크 구조는 [그림 1]에 설명되어 있다. 먼저, 카메라와 라이다 영상은 각 센서 데이터에 대한 특징맵을 생성하기 위해 두 개의 개별 CNN을 통과한다. 제안하는 네트워크의 CNN 구조는 SSD[9]에 사용된 구조와 유사하다. VGG[2] 네트워크가 처음 15층을 이루며, 8개의 추가 CNN이 형성되어 있다. 두 CNN에서 추출한 정보는 총 6개의 층에서 제안하는 센서융합 기술(GIF)을 통하여 융합한다. 최종적으로 융합된 특징맵을 이용하여 물체의 위치를 예측하고 분류하게 된다.

제안하는 센서융합 기술(GIF)은 카메라와 라이다의 특징맵을 입력으로 받으며, 각 입력의 중요도를 판단하여 융합하게 된다. 자세한 구조는 [그림 2]에 설명되어 있으며, 제안하는 융합기술의 전체 네트워크 구조는 중요도 판단



[그림 1] 제안하는 센서융합 물체검출 네트워크 구조



[그림 2] 제안하는 센서융합 기술 네트워크 구조

네트워크와 정보융합 네트워크 두 부분으로 이루어져 있다. 중요도 판단 네트워크는 각 입력을 기반으로 중요도를 출력하게 된다. 중요도를 판단하는 과정은 먼저 각 특징맵을 concatenation을 통하여 묶은 후, 각각의 CNN과 시그모이드 함수를 통하여 각 특징맵의 중요도를 픽셀별로 0부터 1 사이의 값으로 출력하게 이루어져 있다. 묶어진 특징맵을 이용하여 각 중요도를 판단하므로, 각 입력에 따른 상대적 중요도를 출력하게 된다. 정보융합 네트워크에서는 중요도 판단 네트워크에서 출력한 중요도를 각각 특징맵과 픽셀별로 곱하게 된다. 그 후, 각 특징맵을 concatenation을 통하여 다시 묶어 최종적으로 1×1 커널 사이즈의 CNN을 통과하여 융합하게 된다. 중요도 판단 네트워크에서 CNN을 통하여 높은 레벨의 특징을 추출하였으며, 이로부터 픽셀별로 중요도를 판단하였다.

4. 실험

이번 섹션에서는, KITTI 데이터 셋[17]을 이용하여 제안하는 물체검출 알고리즘의 성능을 평가하였다. 또한 제안하는 알고리즘을 학습시키기 위하여 새로운 data augmentation 기법을 사용하였다. 제안하는 센서융합 기술의 효율성을 검증하기 위하여 입력 데이터가 저하된 다양한 환경에서 성능을 평가하였다. 먼저 KITTI 데이터셋에서 3가지 물체(자동차, 사람, 자전거 탄 사람)에 대하여 학습하였으며, 3가지 레벨(easy, moderate, hard)에 대하여 평가하였다. 센서융합을 통한 물체검출을 위하여 KITTI 데이터의 RGB 이미지와 3D 라이다 데이터를 이용하였다. 제안하는 물체검출 네트워크에 맞게 3D 라이다 데이터를 카메라 도메인의 2D 이미지로 변환하였다. 3D 라이다 포인트는 3D 좌표인 (X, Y, Z)와 반사율 R 정보를 가지고 있다. 우리는 이를 KITTI에 주어진 calibration 행렬을 통하여 카메라 도메인 좌표로 변환하였으며, 깊이, 높이, 반사율 정보를 이용하여 총 3채널의 이미지를 만들었다. 각 채널의 픽셀값은 0부터 255까지로 normalization 해주었다. 초기 weight는 ImageNet에서 학습된 VGG-16 모델을 이용하였다. Stochastic gradient descent(SGD) 기법을 이용하여 학습하였으며, 배치 사이즈는 2로 하였다. 초기 학습률은 0.0005로 하였으며, weight decay는 0.0005, momentum은 0.9로 설정하였다. 제안하는 알고리즘은 총 130 epoch을 학습시켜주었다.

제안하는 알고리즘은 gating 기법을 통해 융합하여, 새로운 data augmentation 기법으로 다양한 저하된 환경의 데이터를 입력으로 학습하였다. 제안하는 data augmentation 기법은 i)Blank, ii)Occlusion, iii)Noise, iv)Illumination change 등 총 4가지 방법을 이용하였다. Blank는 픽셀값을 모두 0으로 주는 것이며, Occlusion은 이미지의 일부분을 랜덤하게 검은 박스로 가린 것이다. 또한, Noise는 가우시안 노이즈를 이미지의 모든 부분에 추가하였으며, Illumination change는 영상 일부분의 밝기를 변화시켜주었다. 우리는 각 augmentation 기법을 동일한 확률로 적용하여 학습하였다.

3. 실험 결과

먼저, 우리는 제안하는 알고리즘의 강인성을 증명하기 위하여 저하된 테스트 입력에 대하여 성능을 평가하였다. 비교 알고리즘은 베이스라인, SSD 기반 빠른 융합, SSD 기반 늦은 융합 총 3가지를 비교하였다. 베이스라인은 제안하는 알고리즘에서 중요도를 모두 1로 고정시킨 경우이며, SSD 기반 빠른 융합은 라이다와 카메라 입력을 concatenation을 통하여 묶은 후 한 개의 SSD를 이용한 것이다. SSD 기반 늦은 융합은 라이다와 카메라 입력 데이터를 각각 SSD를 통하여 검출한 후 결과를 합친 것이다. 우리는 모든 알고리즘을 똑같은 data augmentation 기법을 이용하여 학습하였으며, 같은 저하된 테스트 데이터에 대하여 성능을 평가하였다. 테스트 입력은 두 입력 모두 온전한 경우와 한 가지 입력이 blank, occlusion, noise, 혹은 illumination change 등의 저하를 발생하는 게 공평

〈표 1〉 저하된 테스트 데이터에서의 물체검출 성능

테스트 입력	제안하는 알고리즘			베이스라인			SSD 기반 빠른 융합			SSD 기반 늦은 융합		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Total	93.95	86.70	78.05	89.86	825.21	72.21	91.10	85.65	75.83	89.69	82.03	72.96
RGB + Lidar	98.69	90.31	82.16	93.61	87.01	77.52	95.84	89.94	79.67	91.72	87.93	78.46
RGB (blank) + Lidar	88.86	78.12	69.68	86.56	74.30	64.71	89.94	78.99	69.56	87.92	77.83	69.11
RGB + Lidar (blank)	97.39	90.29	81.84	91.88	88.10	78.69	90.48	88.56	77.92	93.31	89.27	80.03
RGB (occl.) + Lidar	89.88	88.12	79.03	88.12	78.52	68.85	90.22	84.15	73.93	91.78	88.22	78.80
RGB + Lidar (occl.)	97.72	90.23	81.94	92.75	87.10	77.67	90.53	88.91	79.07	84.80	74.88	66.33
RGB (noise) + Lidar	89.33	80.15	71.12	86.75	75.13	65.71	90.18	81.29	72.04	88.67	76.12	67.18
RGB (illum.) + Lidar	95.82	89.71	80.58	89.37	85.31	75.87	90.48	88.42	78.60	89.69	79.96	70.82

치 않기 때문에 이를 빼고 진행하였다. 라이다 데이터는 noise와 illumination change를 적용시킬 수 없어 이에 대해서는 실험을 진행하지 않았다. 실험 결과에 대한 자세한 성능은 [표 1]에 제시하였다. 제안하는 알고리즘이 대부분의 경우에서 더 좋은 성능을 나타냈으며, blank의 경우에서는 베이스라인에 비하여 5% 이상의 성능향상을 이루었다. 또한 제안하는 알고리즘이 다른 융합 기법에 비하여 강인함을 나타낼 뿐만 아니라 두 입력이 모두 온전한 경우에도 성능이 향상되는 것을 확인하였다. 이는 gating 기법을 통한 센서융합 기술이 입력이 저하된 경우의 강인함 뿐 아니라 두 입력이 모두 온전한 경우에도 각 입력의 중요도를 판단하여 더 좋은 융합을 이루어내는 것을 확인할 수 있다.

5. 결론

본 논문에서는 라이다 기반의 물체검출 방법들을 소개하고 환경변화에 강인한 센서융합 기술에 기반한 물체검출 알고리즘을 제안하였다. 두 개의 CNN이 RGB 이미지와 라이다 이미지를 입력으로 받았으며, 추출된 특징맵을 제안하는 센서융합 기술을 통하여 융합하였다. 제안하는 센서융합 기술은 gating 기법을 이용하여 각 입력의 중요도를 판단하였으며, 이를 기반으로 환경변화에 강인한 융합을 하였다. KITTI 데이터 셋을 이용하여 실험을 진행한 결과 제안하는 알고리즘이 다른 융합기술에 비하여 강인함을 띠는 것을 확인하였으며, 기존의 state-of-the-art 알고리즘보다도 비슷하거나 더 좋은 성능을 나타내었다.

참고문헌

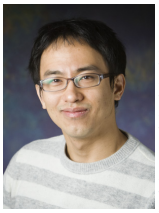
1. A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS 2012
2. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR 2015.
3. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR 2016.
4. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014
5. R. Girshick. Fast R-CNN. In ICCV 2015
6. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS 2015

7. K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In ICCV 2017
8. J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection. In: CVPR 2016
9. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single shot multibox detector. In ECCV 2016
10. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar. Focal loss for dense object detection. In ICCV 2017
11. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A.Y. Ng. Multimodal deep learning. In ICML 2011
12. A. Eitel, J.T. Springenberg, L. Spinello, M.A. Riedmiller, and W. Burgard. Multi-modal deep learning for robust rgb-d object recognition. In IROS 2015
13. X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In CVPR 2017
14. J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander. Joint 3d proposal generation and object detection from view aggregation. In IROS 2018
15. M. Liang, B. Yang, S. Wang, and R. Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In ECCV 2018
16. S. Wang, S. Suo, W.C. Ma, and R. Urtasun. Deep parameteric convolutional neural networks. In CVPR 2018
17. A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR 2012
18. X. Chen, K. Kundu, Y. Zhu, A.G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In: Adv. in NIPS 2015
19. X. Chen, K. Kundu, Z. Zhang, H. Fidler, and R. Urtasun. Monocular 3d object detection for autonomous driving. In: CVPR 2016
20. F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. Chateau. Deep manta: a coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular

김재겸
2016 한양대학교 전기공학과 졸업(학사)
관심분야 : 머신러닝, 딥러닝
Email: jkkim@spa.hanyang.ac.kr

고준호
2018 한양대학교 전기공학과 졸업(학사)
2018~현재 한양대학교 전기공학과 (석사)
관심분야 : 머신러닝, 딥러닝
Email: jhkoh@spa.hanyang.ac.kr
2016~현재 한양대학교 전기공학과(박사)

김예철
2018 한양대학교 전기공학과 졸업(학사)
2018~현재 한양대학교 전기공학과 (석사)
관심분야 : 머신러닝, 딥러닝
Email: yckim@spa.hanyang.ac.kr



최준원
2000 서울대학교 전기공학과 졸업(학사)
2002 서울대학교 전기공학과 졸업(석사)
2010 미국 Univ. of Illinois at Urbana-Champaign 졸업 (박사)
2010~2013 퀄컴, 샌디에고 미국
2013~현재 한양대학교 전기생체공학부 부교수
관심분야 : 신호처리, 머신러닝, 무선통신, 자율주행
Email: junwchoi@hanyang.ac.kr