딥러닝을 이용한 시각 장애인 보조 시스템 개발

허규진, 오진숙, 김한샘, 이민학, 강우철
인천대학교 임베디드시스템공학과
e-mail:{201101851, 201301669, 201101819, mnaklee, wchkang}@inu.ac.kr

The development of a blind people assistant system using deep learning techniques

KyuJin Heo, JinSook Oh, HanSaem Kim, Minhak Lee, Woochul Kang
Dept of Embedded System Engineering, Incheon National University

요 약

시각장애인의 인구비율은 전체 장애인 인구의 약 10%로 적지 않은 비율을 차지한다. 이러한 시각장애인들에게 가장 위험한 요소는 주변의 물체들이다. 하지만 현재 제시되어 있는 안전 보조 장치(보도블록 등)는 시각 장애인들 스스로가 전방에 어떠한 물체가 있는지 인지하기 어렵다. 본 논문에서는 딥러닝을 이용하여 위험물체에 대한 학습모델을 생성 한 뒤 햅틱 모션 및 음성 안내를 통하여 실시간으로 시각장애인이 위험상황을 인지할 수 있는 시스템을 제안한다.

1. 서론

최근 기계학습에 대한 관심이 증가하였고 다양한 분야에서 놀라운 성과를 내면서 기계학습은 폭 넓은 분야에 적용될 수 있음을 증명하였다. 특히 딥러닝은 기계학습의 다른 분야보다 예측력이 좋다는 평가를 받고 있고 기존 신경망 학습시의 문제였던 모수 (parameter) 최적화 문제가 해결되면서 현재 가장 각광받는 머신러닝 분석 방법으로 이슈가 되고 있다. 특히 Convolutional Neural Network (CNN)를 이용하여 만든 딥러닝 모델은 이미지 식별 분야에서 인간과 대등한 성능을 구현해냈으며, 음성인식 등의 분야에서도 Recurrent Neural Network (RNN)을 이용한 연구가 활발하게 진행 중이다 [1].

년도	시각		
	남자	여자	계
2012	150,815	101,749	252,564
2013	151,009	102,086	253,095
2014	150,843	101,982	252,825
2015	150,883	101,991	252,874

Table 1. 2012 ~ 2015년 시각 장애인 현황 [2]

특히 시각 장애인을 보조할 수 있는 기구들의 개발은 Table. 1에 나타나듯이 시각 장애인들의 인구가 많아지면서 더욱 중요해진 분야이다. 기존에 존재하는 시각 장애인들을 위한 보조 장치들의 대표적인 예로는 보도블록에 설치된 원형 블록과 음향신호기 등이 존재하는데 이것들만으로는 시각장애인 스스로 전방에 위험 요소들을 확인할 수 없고, 이 마저도 설치 안 된 곳들이 많아 현실적으로 한계가 명확하기 때문이다. 따라서 본 논문에서는 딥러닝 모델을 이용한 시각 장애인용 이미지 식별 시스템을 제안한다.

2. 전체 시스템 구성도

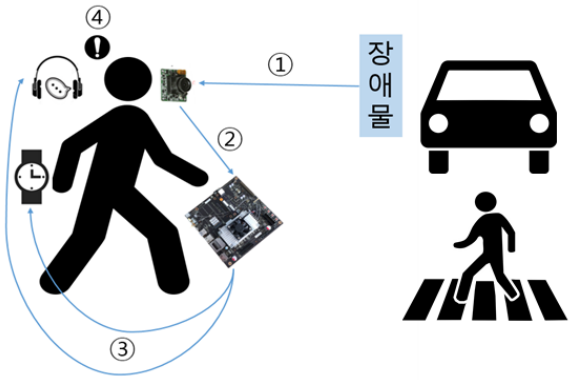

Fig. 1. 시스템의 전체 구성도

본 논문에서 제안하는 시스템의 전체적인 구성도는 Fig. 1과 같다. ①몸에 부착되어있는 카메라가 전방의 이미지를 캡처한 뒤 ②해당 이미지를 임베디드 보드로 전송시킨다. ③해당 보드에서 미리 학습한 딥러닝 모델을 토대로 해당 이미지가 어떤 물체이며 어떠한 위험도를 가지고 있는지 식별을 한다. 이때의 위험도는 상/중/하로 미리 나누어 놓은 데이터를 이용한다. ④이러한 정보를 토대로 멀티 모달 방식으로 무선 이어폰을 통한 음성 안내와 웨어러블 디바이스를 이용한 햅틱 모션을 통해 이용자에게 주변 상황을 실시간에 알리게 된다. 이미지 전송(보드 - 카메라)에는 Wi-Fi AD-HOC통신을, 위험 알림(보드 - 웨어러블 디바이스,이어폰)에는 블루투스 통신을 사용하였다. 본 논문에서는 딥러닝을 통한 이미지식별 부분에 집중하여 설명하도록 한다.

3. 딥러닝을 이용한 이미지 식별 학습 모델

3-1. 물체인식을 위한 신경망 모델링

본 프로젝트의 가장 핵심이 되는 기술은 시각 장애인이 판단할 수 없는 외부 상황을 이미지 분석을 통하여 시각 장애인에게 알려줄 수 있는 것이다. 이미지 분석을 위해 사용되는 기술들 중에서 현재 가장 최적화된 결과를 보이는 방법은 딥러닝 기술이다. 이러한 결과를 추출하기 위해 CNN을 이용하여 임의의 필터를 거친 뒤 사진의 특징들을 추출하여 데이터로 만들었으며 이를 위한 데이터 집합들은 IMAGENET에서 배포하는 이미지들을 필요에 따라 선별하여 사용하였다 [3]. 이러한 데이터를 학습하기 위해 GoogLeNet[4]이라는 오픈 소스 아키텍쳐를 사용하였다. 식별 가능한 물체의 수는 71가지이며 ***학습용 이미지 집합***은 90026장, ***검증용 이미지 집합***은 3550장으로 구성되어 있다. CNN이란 많은 양의 이미지 데이터를 학습이라는 과정을 통하여 신경망의 가중치 값들을 점진적으로 조절해 나가는 과정이다. 이는 Stochastic Gradient Descent (SGD)라는 최적화 과정을 반복적으로 적용해나가는 것으로 많은 양의 부동소수점 연산을 요구하게 되는데 이를 해결하기 위해 6.14 Tflops의 연산능력을 제공하는 Nvidia TITAN X라는 Graphic Processing Unit(GPU)을 활용하였다.

3-2. 인셉션 모듈

기본적으로 인셉션 모듈을 구현하였을 경우 1x1 convolution, 3x3 convolution, 5x5 convolution, 3x3 max pooling을 나란히 놓는 구조를 고안해냈으나 이러한 구조는 망의 깊이가 깊어지는 GoogLeNet과 같은 구조에서는 연산량이 급격하게 증가하게 되는 결과를 가져오게 된다. 따라서 Fig. 2와 같이 3x3 convolutions과 5x5 convolutions앞에 1x1 convolution을 두고, 이를 통해 특징 맵의 차원을 줄이게 되면 특징 추출을 위한 여러 값들을 가지면서 동시에 연산량의 균형을 맞출 수 있게 된다[4].

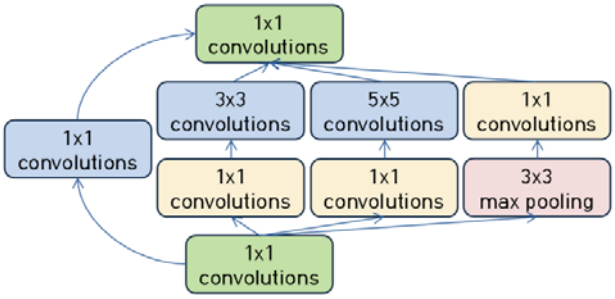

Fig. 2. 인셉션 모듈

3-3. 데이터 집합의 제작 과정

이미지들의 학습을 위하여 기본적으로 이미지들을 하나의 집합으로 만들어야 할 필요가 있었고 NVIDIA사에서 제작한 DIGIT이라는 툴을 이용하여 데이터집합을 제작하였다. 데이터 집합은 크게 2가지가 필요하다. 1차적인 학습만을 위한 ***학습용 이미지 집합***이 필요하고, 이것이 얼마나 정확한 값을 가지는지 확인하는 ***검증용 이미지 집합***이 필요하다. ***학습용 이미지 집합***에 ***검증용 이미지 집합***이 포함될 경우 학습의 결과가 저조하게 나오는 것을 확인하였다. 이러한 문제를 극복하고자 2개의 이미지들은 완전히 다른 이미지들로만 구성하였다. ***학습용 이미지 집합***은 90026장의 이미지와 71개의 클래스로 구성되어 있으며 ***검증용 이미지 집합***은 3550장의 이미지와 71개의 클래스로 구성되어 있다. 이 이미지들의 레이블을 지정하기 위한 클래스명과 의미가 매칭되어 있는 텍스트 파일을 따로 작성하여 준비하였다.

3-4. 이미지 학습 과정

3-3에서 제작한 데이터집합을 바탕으로 이미지들의 학습을 진행하였다. 아키텍쳐로는 GoogLeNet을 사용하였으며, IMAGENET에서 선별한 위험요소들의 이미지들을 입력 데이터로 활용하였다. 학습에 사용한 GPU는 NVIDIA TITAN X 1대로 하였으며 학습시간은 약 12시간이 소요되었다.

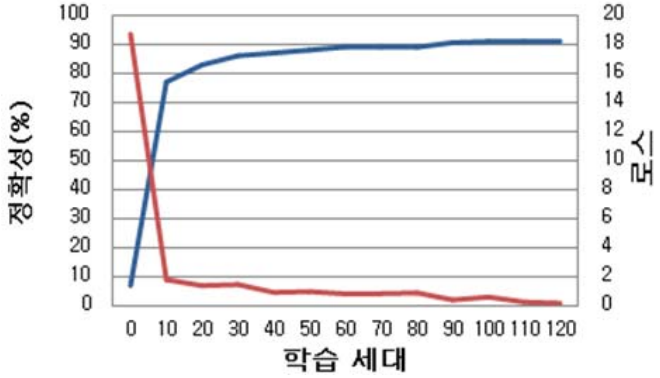

Fig. 3 트레이닝 그래프

Fig. 3.는 학습이 반복되면서 나타나는 정확도와 로스(Loss)를 그래프로 나타낸 값이다. X축은 세대를 나타내며, 오른쪽의 Y축은 로스, 왼쪽의 Y축은 정확도를 나타낸다. 약 60세대(약 40만 번의 학습)가 진행 된 이후부터는 정확도가 90%에 육박하여 가장 최적의 정확성을 나타낸다. 90%이상의 정확도를 나타내는 모델 중 가장 로스가 적은 모델을 선택하여 개발을 진행하였다.

3-5. 이미지 식별 과정

제안하는 시각장애인 보조 기구 시스템의 동작 순서는 Fig. 4를 통하여 설명한다. 우선 이미지를 받아오기 전 모델 인스턴스를 먼저 생성한다. 이미지를 받을 때 마다 매번 생성할 경우 프로그램의 동작 속도가 느려지고 오버헤드가 상당히 커지는 부분이 발생하여 이를 줄이기 위하여 초기에 인스턴스를 생성하고 이것을 반복해서 사용하는 방식으로 진행하였다. 이미지를 로딩하고 나면 기존에 만들었던 모델을 이용하여 이미지 식별을 해야 하는데 그에 앞서 모델을 제작할 때 사용하였던 이미지의 사이즈대로 재설정을 해줘야하기에 이 작업을 해준다. 특징점을 추출하기 위해 생성한 모델에 입력된 이미지를 적용시켜 CNN을 통한 특징점 추출을 하게 된다. 모든 이미지들은 batch_Size에 따라 Convolutions 연산을 반복하여 지역 특징 점으로부터 전체 특징 점을 추출해내고 그 값을 바탕으로 전체 클래스에 대한 스코어를 가지게 된다. 모든 스코어들은 100을 만

점으로 상대적인 값을 가지게 되며 이 정보를 바탕으로 Top-5 개의 값을 결정하게 된다.

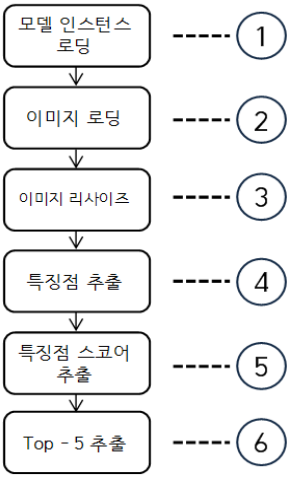

Fig. 4. 이미지 식별 과정

Fig. 5 은 GoogLeNet을 이용하여 이미지의 특징을 추출하는 한 과정이다. 지역 특징점으로부터 pooling과정 등을 거쳐 전체 특징점를 추출하게 되고 최종 레이어에서 패턴 특징을 찾아내어 해당 이미지가 모델에 어떤 클래스에 해당하는지 알아내게 된다.

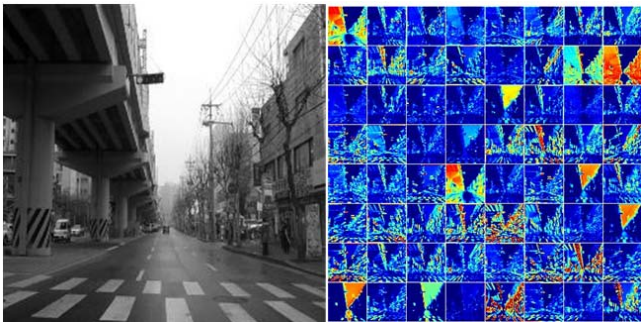

Fig. 5. 이미지의 특징점 추출 과정

4. 데이터 통신 구성 및 결과

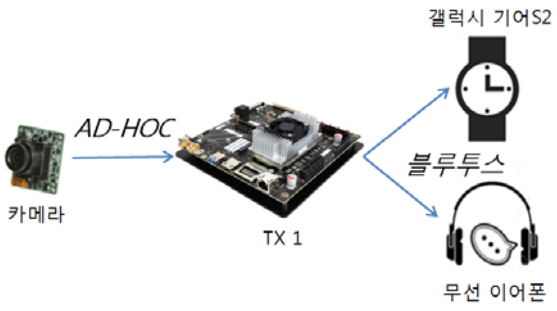

Fig. 6. 데이터 통신 구성도

본 논문에서 제안하는 시스템의 통신 구성은 Fig. 6과 같다. 딥러닝 모델을 통한 인식은 1 Tflops의 연산능력을 제공하는 Nvidia Jetson TX1 임베디드 보드를 사용하였다.

외부 상황에 대한 카메라 영상 캡쳐는 상대적으로 소형인 라즈베리 파이 카메라를 이용하였다. 시각 장애인이 항상 네트워크 구축을 위한 공유기를 들고 다닐 수 없기에 Wi-Fi AD-HOC 통신을 이용하여 카메라 모듈이 설치된 라즈베리파이와 Jetson TX1 간 1:1 통신을 구현 하여 이미지의 전송이 가능하도록 하였다.

또한 이미지 식별 결과를 다수의 웨어러블 디바이스에 블루투스 통신을 통해 전달함으로써, 안전성을 높이도록 하였다. 음성 안내는 블루투스 헤드폰으로 전송되며, 미리 녹음해둔 mp3 파일을 순차적으로 재생하는 방식으로 구현하였다. 전방 물체의 경우 식별 된 결과를 토대로 해당 물체의 음성파일을 재생하도록 하였다. 그리고 갤럭시 기어 S2 웨어러블 디바이스를 통해 전방 물체에 따라 햅틱 모션을 다르게 주어 시각 장애인에게 음성뿐만 아니라 모션으로도 알림이 가능하도록 하였다.

5. 이미지 식별 기능 개발 환경 및 테스트 결과

5-1. 개발 환경

GPU 용도	사용 GPU	아키텍쳐
		코어 수
이미지 학습	Nvidia TITAN X	Maxwell
		3702개
이미지 식별	Nvidia TX 1	Maxwell
		256개

Table. 2. GPU 환경[5]

학습 프레임	DIGIT(Caffe)
학습 아키텍쳐	GoogeLeNet
통신 방식	AD-HOC, 블루투스
학습용 이미지 갯수	90026
검증용 이미지 갯수	3550
물체의 클래스 갯수	71

Table. 3. 기타 환경

5-2. 테스트 결과

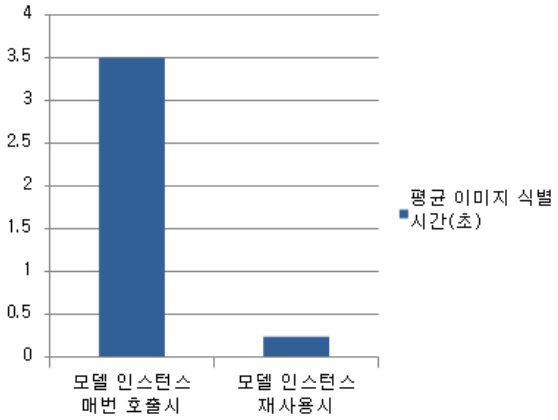

Fig. 7. 모델 인스턴스 재사용 유무에 따른 성능 비교

Fig. 7을 보면 모델 인스턴스의 재사용 유무에 따른 평균 이미지 식별 시간을 나타내고 있다. 이는 이미지를 식별 할 때마다 매번 하는 인스턴스의 로딩은 필요치 않은 메모리의 낭비와 시간의 소모를 증명하였다. 이 경우 평균 약 3.5초의 식별 시간이 소요된 반면 한 번의 로딩 후 이를 재사용 하여 식별할 경우 약 0.24초로 10배 이상의 성능향상을 확인 할 수 있다. 본 시스템은 보행자의 실시간 상황 인지를 위한 것이므로, 0.24초의 응답시간은 도로에서 일어나는 상황들을 인식하여, 위험을 알리기에 충분히 빠른 응답시간이라고 판단한다.

Fig. 8. 이미지 식별 결과

이미지 식별 결과는 Fig. 8 과 같이 나타난다. 예시로 횡단보도의 사진을 넣었을 때 68%의 스코어로 횡단보도를 식별하였으며 동시에 24%의 스코어로 신호등을 식별하였다.

그 외의 다른 이미지들을 입력으로 주었을 때의 결과는 Table. 4와 같다. cab, convertible은 자동차 등의 의미로 주차된 자동차들을 각각 46%, 23%의 스코어로 식별하였고, 사람들이 모여 있는 사진을 입력으로 주었을 경우 human이라는 값을 62%라는 높은 값으로 식별하는 결과를 보였다.

사진	식별 결과	스코어
	cab, convertible	46%, 23%
	Human	62%
	crosswalk, traffic sig	68%, 24%

Table. 4. 그 외 이미지의 식별 결과

5. 결론

본 논문에서는 시각 장애인들에게 도움이 되고자 이들에게 위험이 될 수 있다고 판단되는 물체들이 전방에 있는 지 알려주는 시스템을 제안하였다. 이 시스템은 패턴을 추출하여 식별하기 때문에 학습된 물체라면 어떠한 모습으로 존재하더라도 식별이 가능하다는 장점을 가지고 있다. 이를 위해 GoogLeNet이라는 아키텍쳐를 이용하였다. 특히 학습에 있어서는 기존에 오픈소스로 공개되어있는 AlexNet[7] 등의 여러 아키텍쳐를 사용해 보았으나 비교적 정확도가 높게 학습된 것을 선택하였다. 본 논문의 실험결과 GPU가 내장된 임베디드 보드를 사용할 경우 약 0.24초의 응답시간으로 주변 상황에 대한 영상 인식이 가능함을 확인하였다. 이는 보행자시에 일어날 수 있는 상황에 대한 충분한 대처시간을 시각장애인에게 부여함으로써 보행 안전성을 높일 수 있다. 또한, 본 논문은 인식 결과를 다수의 웨어러블 기기를 통해 알림으로써, 더욱 높은 보행 안전성을 보장하게 된다.

향후 연구에서는 현 아키텍쳐의 성능 개선시켜 정확도와 로스값이 각각 최적의 값을 가지고, 이를 통하여 보다 더욱 정확한 식별 값을 추출하는 연구를 제시할 것이다.

참고문헌

[1] Alex Krizhevsky, Hya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, 2012

[2] *KOSIS,* http://kosis.kr

[3] *IMAGENET,* http://image-net.org/

[4] Hyeong-Joong Yoo, “Deep Convolution Neural Networks in Computer VIsion : a Review,” *IEIE Transactions on Smart Processing and Computing,* Vol.4, No. 1, February 2015

[5] Kim. JiWon, Ha. JungWoo, Lee ChanKyu, Kim JungHee, “Deep Learning algorithms and applications,“ *Journal of KIISE*, Vol. 33, No. 8, pp. 85-86, 2015

[6] *NVIDIA,* www.nvdia.com

[7] Mi-Young Lee, Jin-Kyu Kim, Byung-Jo Kim, Ju-Yeob Kim, Joo-Hyun Lee. "Analysis of Deep Learning Framework for Visual Perception." *대한전자공학회 학술대회,* (2016.6): 1808-1811.