



Researcher 장상범, 전자공학과 ( jangbbam@ajou.ac.kr )  
황보찬, 전자공학과 ( halamadrid@ajou.ac.kr )

Professor 선우명훈, 전자공학과

## ABSTRACT

- 인공지능 기술은 딥러닝 및 강화학습의 등장과 함께 크게 발달하기 시작했다. 이 덕분에 최근에는 음성 인식, 이미지 인식 등 다양한 분야에서 인공지능 기술이 사용되고 있다.
- 특히 Object Detection 은 주로 Faster R-CNN 계열이 좋은 성능을 내고 있었다. 하지만 Faster라는 단어와 맞지 않게 10프레임도 안 나왔으나 YOLO가 등장하여 45 프레임을 보여주었다.
- ( TITAN-X 기준의 정확도와 프레임  
YOLOv2 : 76.8mAP , 67 FPS, Faster R-CNN : 73.2mAP 7 FPS )
- 현재까지 국내에서는 인공지능 시스템의 구현이 시뮬레이션 수준에서 멈추어 있는 경우가 많다. 국내에 실물 구현의 예가 부족하기 때문에 본 프로젝트를 계획하였다.

## OBJECTIVES

- 이 연구에서는 딥러닝을 바탕으로 하여 물체인식과 표정검출을 만족하는 Filter수와 Classes수를 구하고 그 값을 바탕으로 YOLO Darknet 에 적용했을때, 목표로 했던 6가지의 표정검출을 만족하는지 여부를 실시간 WEP CAM simulation을 통해서 알아보았다.

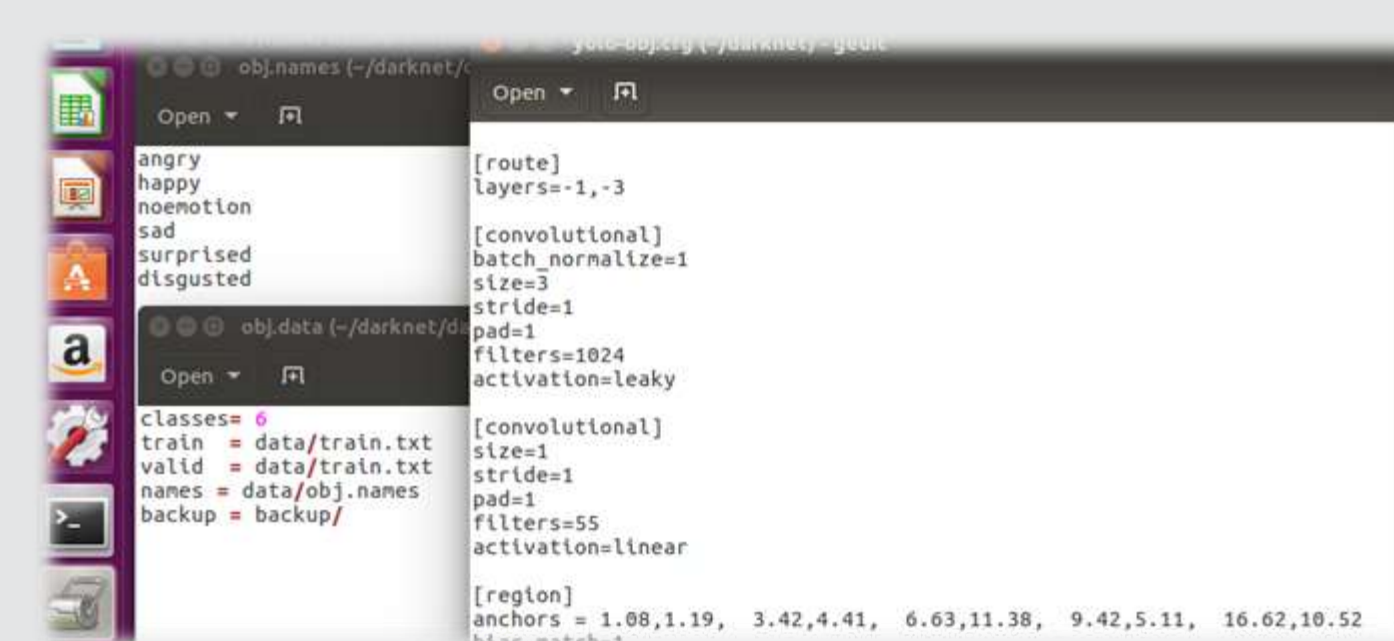
## METHODOLOGY

### 1. YOLO v3 환경

layer	filters	size	input	output
0 conv	32	3 x 3 / 1	416 x 416 x 3	-> 416 x 416 x 32
1 max	2	2 x 2 / 2	416 x 416 x 32	-> 208 x 208 x 32
2 conv	64	3 x 3 / 1	208 x 208 x 32	-> 208 x 208 x 64
3 max	2	2 x 2 / 2	208 x 208 x 64	-> 104 x 104 x 64
4 conv	128	3 x 3 / 1	104 x 104 x 64	-> 104 x 104 x 128
5 conv	64	1 x 1 / 1	104 x 104 x 128	-> 104 x 104 x 64
6 conv	128	3 x 3 / 1	104 x 104 x 64	-> 104 x 104 x 128
7 max	2	2 x 2 / 2	104 x 104 x 128	-> 52 x 52 x 128
8 conv	256	3 x 3 / 1	52 x 52 x 128	-> 52 x 52 x 256
9 conv	128	1 x 1 / 1	52 x 52 x 256	-> 52 x 52 x 128
10 conv	256	3 x 3 / 1	52 x 52 x 128	-> 52 x 52 x 256
11 max	2	2 x 2 / 2	52 x 52 x 256	-> 26 x 26 x 256
12 conv	512	3 x 3 / 1	26 x 26 x 256	-> 26 x 26 x 512
13 conv	256	1 x 1 / 1	26 x 26 x 512	-> 26 x 26 x 256
14 conv	512	3 x 3 / 1	26 x 26 x 256	-> 26 x 26 x 512
15 conv	256	1 x 1 / 1	26 x 26 x 512	-> 26 x 26 x 256
16 conv	512	3 x 3 / 1	26 x 26 x 256	-> 26 x 26 x 512
17 max	2	2 x 2 / 2	26 x 26 x 512	-> 13 x 13 x 512
18 conv	1024	3 x 3 / 1	13 x 13 x 512	-> 13 x 13 x 1024
19 conv	512	1 x 1 / 1	13 x 13 x 1024	-> 13 x 13 x 512
20 conv	1024	3 x 3 / 1	13 x 13 x 512	-> 13 x 13 x 1024
21 conv	512	1 x 1 / 1	13 x 13 x 1024	-> 13 x 13 x 512
22 conv	1024	3 x 3 / 1	13 x 13 x 512	-> 13 x 13 x 1024
23 conv	1024	3 x 3 / 1	13 x 13 x 1024	-> 13 x 13 x 1024
24 conv	1024	3 x 3 / 1	13 x 13 x 1024	-> 13 x 13 x 1024
25 route	16			
26 reorg		/ 2	26 x 26 x 512	-> 13 x 13 x 2048
27 route	26 24			
28 conv	1024	3 x 3 / 1	13 x 13 x 2048	-> 13 x 13 x 1024
29 conv	55	1 x 1 / 1	13 x 13 x 1024	-> 13 x 13 x 55
30 detection				

Loading weights from 6\_emo\_160000.backup...Done!

### 2. Yolo Darknet 딥러닝 학습 방법



- (1) Obj.names와 obj.data 에서 학습할 표정과 사진파일의 경로를 설정.
- (2) yolo-obj.cfg의 filters값과 classes값을 데이터에 맞게 설정  
Filter = 5 x ( Class +5 )  
Filter = 5 5, classes=6 으로 학습
- (3) YOLO mark  
- 이미지 파일에 Bounding Box를 그려줌으로서 Box 좌표를 지정해줌

### 3. Simulation YOLO v3 Structure

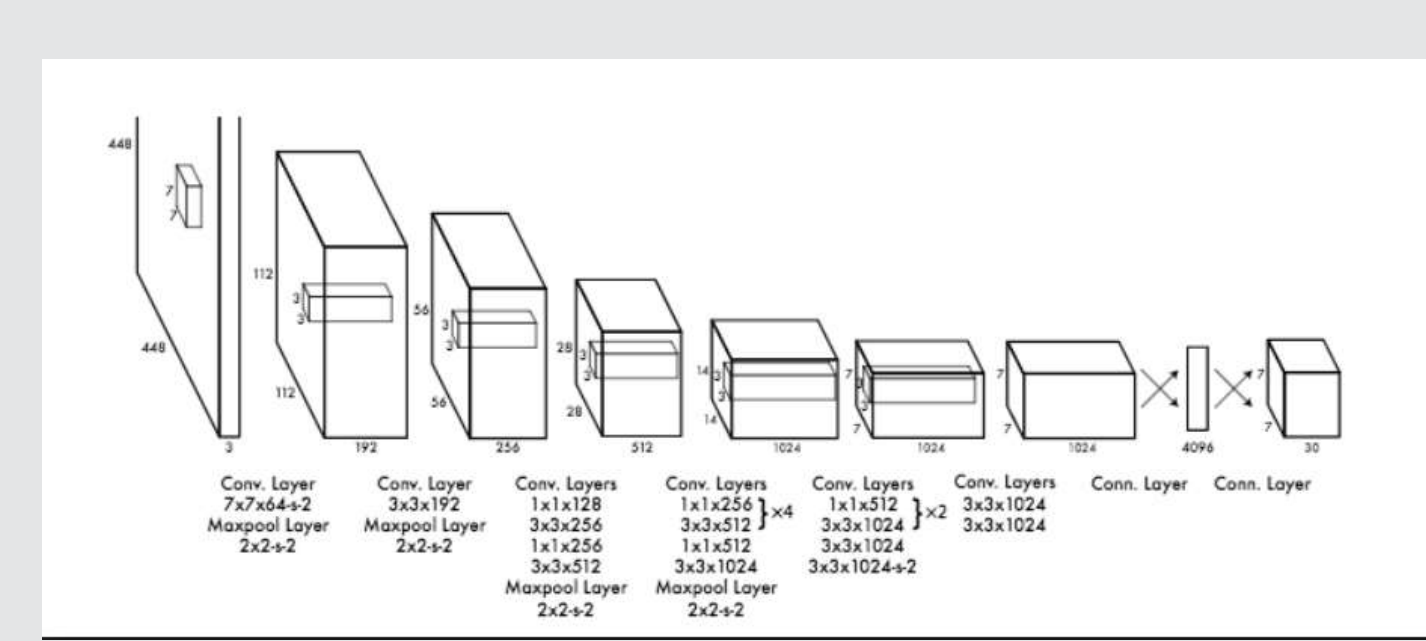


Fig 3. Simulation의 Flow Chart

Convolution Layer : 3x3 1x1 의 두종류로 구성  
Pooling Layer : Network Size를 줄임.  
Skip Connection : Network의 정확도를 올리기 위해 이전 단계 Layer와 현재단계 Layer를 포함하여 연산  
FC Layer : 최종 결과값을 구하기 위해 Fully Connected Layer를 이용

## RESULTS

### 1. YOLO Darknet Network

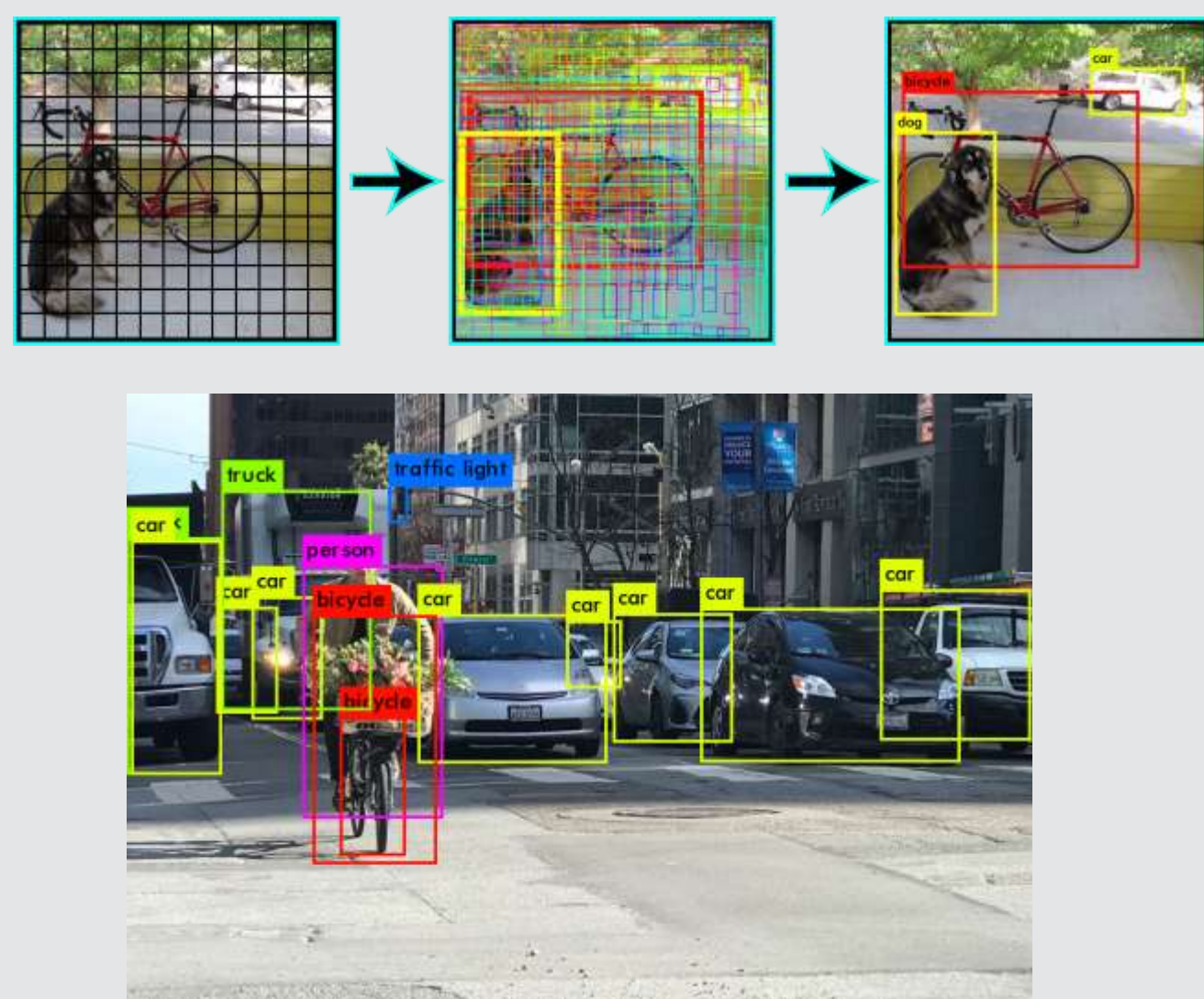


Fig 5. 객체인식에 특화된 YOLO 딥러닝

### 2. YOLO Darknet 학습 결과 [시행 시간 : 48hour]

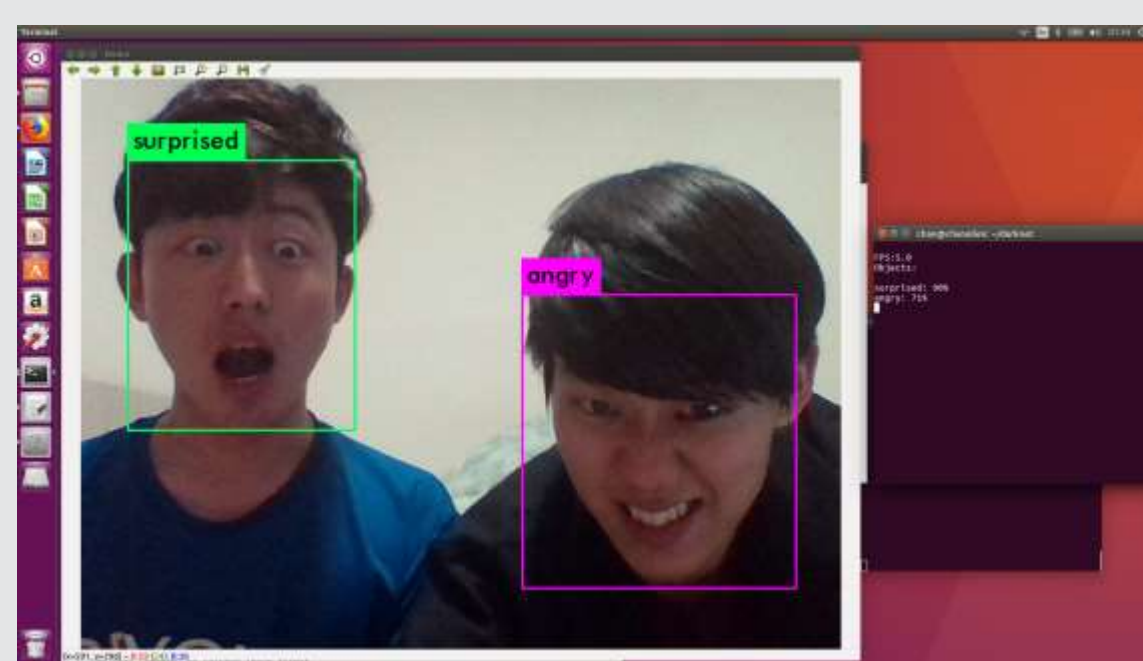
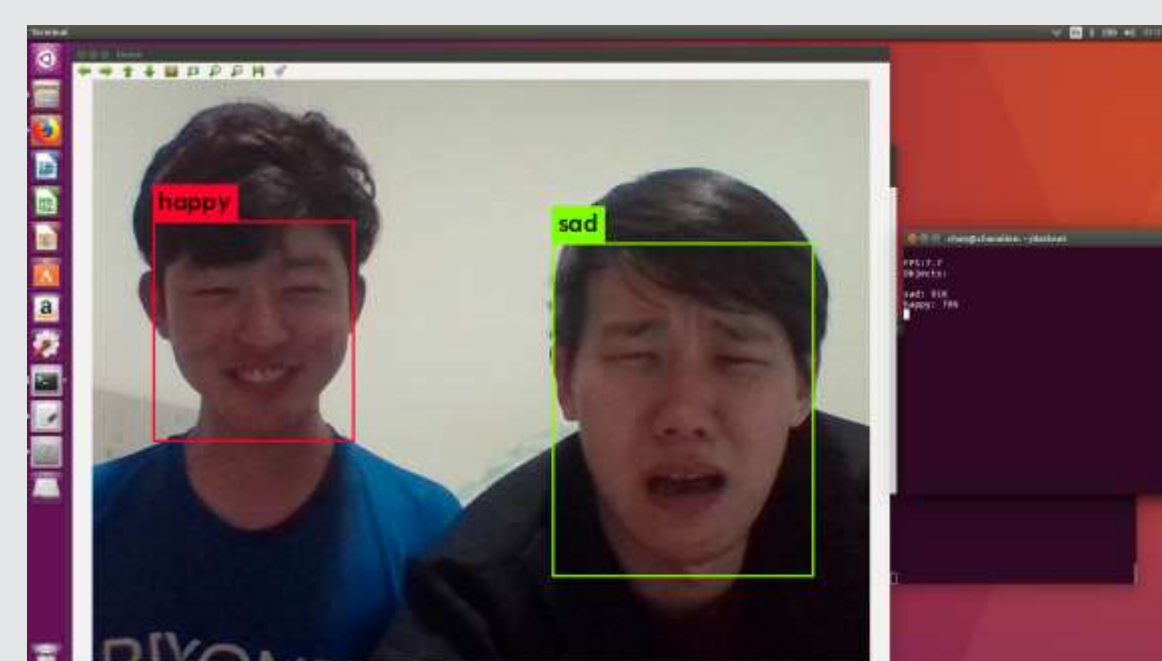


Fig 7. 동시에 검출되는 두 개의 표정

Surprised 인식률	Angry 인식률
90%	71%



Happy 인식률	Sad 인식률
79%	91%

- Yolo Darknet 결과를 바탕으로 검출된 표정들 출력값에 따라서 x,y,w,h 값 및 정확도를 기반으로 bounding box 를 표시 한 결과

- 한 번에 한 개의 표정이 아닌, 두 개 이상의 표정을 동시에 검출 가능

- 현재 90, 79%등의 결과가 나타났지만, 실시간으로 검출을 하기 때문에 정확도에 다소 변동이 크게 나타남

- MAX 20~30 Frame 정도의 상당히 높은 Frame rate를 가짐

학습 Step 수	Class	Obj	No Obj
5000	0.7335	0.724	0.185
10000	0.8522	0.852	0.115
22000	0.9997	0.907	0.046

## CONCLUSIONS

- 이 연구에서는 YOLO v3 Structure를 바탕으로 하여 Convolution layer =22 개, Max pooling =5 개, route 2번을 하는 스트럭처와 수정을 통해 구할 수 있었다.
- 표정당 약 100장, 총 636장의 Image를 이용하여 여섯가지의 표정을 학습하였다.
- (Happy, Sad, Angry, Disgusted, No emotion, Surprised)
- 그 값을 바탕으로 웹캠에서 실시간 얼굴 디텍션과 평균 90%이상의 표정 검출률을 보였고 총 표정의 검출이 성공하였다.