

머신러닝을 이용한 최적 모형 찾기



202STG01 고유정

Contents

I. Mushroom data

- 1) 자료설명
- 2) EDA
- 3) 모형탐색&최적모형 선택

II. Garbage image data

- 1) 자료설명
- 2) 모형탐색 & 최적 모형선택
- 3) 합성망이 학습한 내용 시각화

I. Mushroom data

1) 자료설명

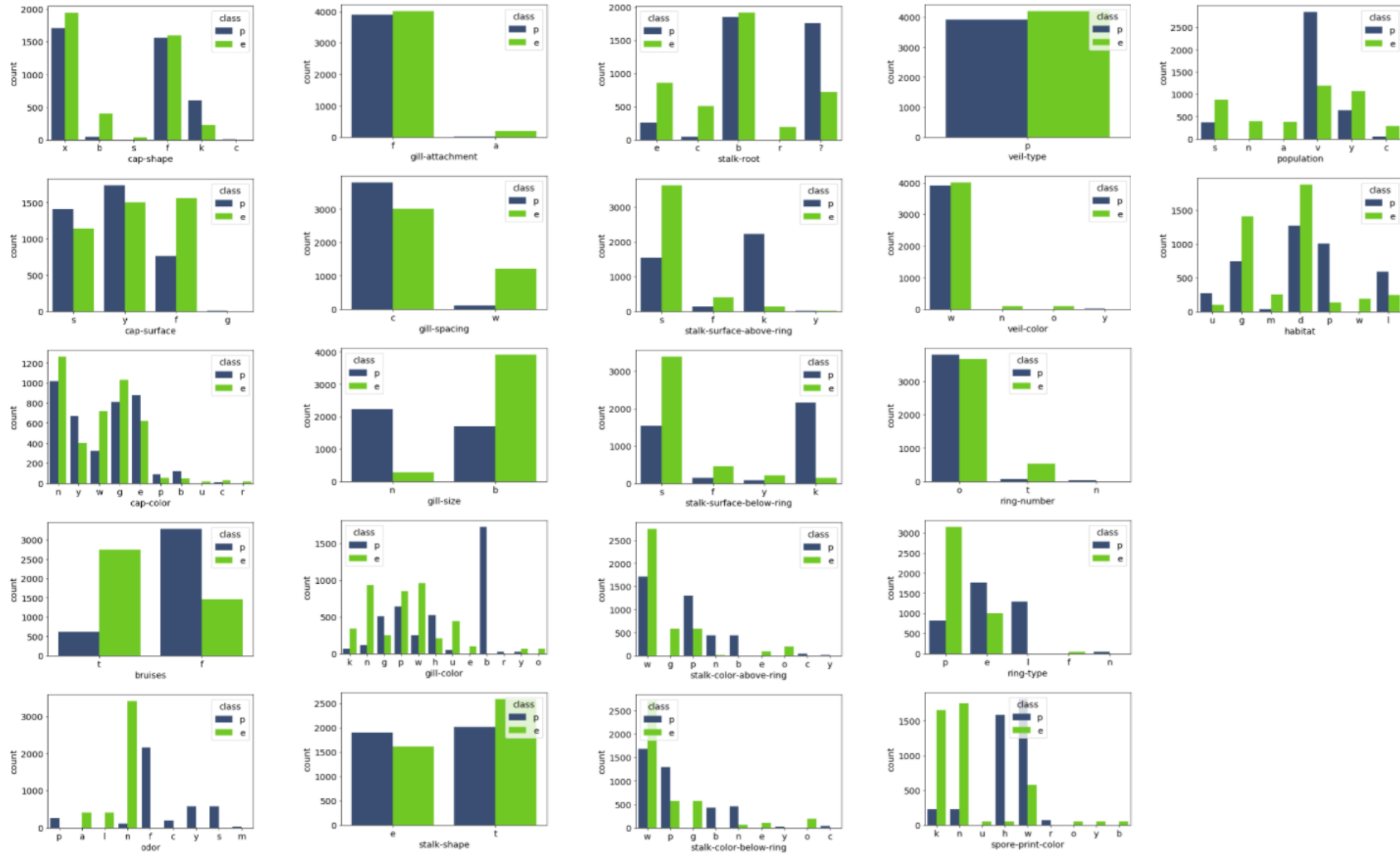
출처 : <https://www.kaggle.com/uciml/mushroom-classification>

class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	stalk-surface-above-ring	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population	habitat
p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s	u
e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	n	g
e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	n	m
p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	u
e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g

- gilled mushroom의 23가지 종류를 변수로 뒀다.
- 8124개의 gilled mushroom의 특성을 나타내고 있다
- 모든 변수는 범주형 변수다.
- 'class'를 y로 두고 나머지 변수들을 x로 두어 식용 가능한 버섯과 독성을 띄는 버섯을 구분하는 중요 특성/요인을 분류하고 최적 예측 모형을 알아보려 한다.

I. Mushroom data

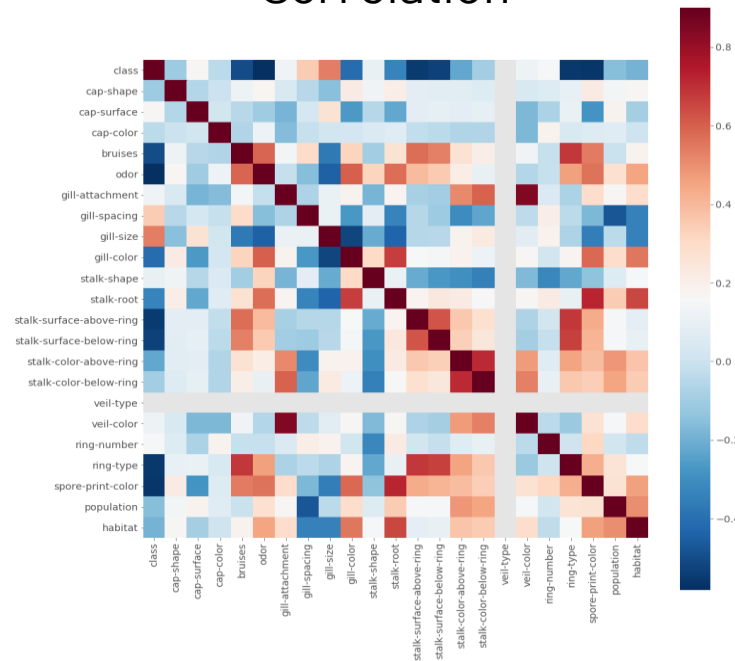
2) EDA



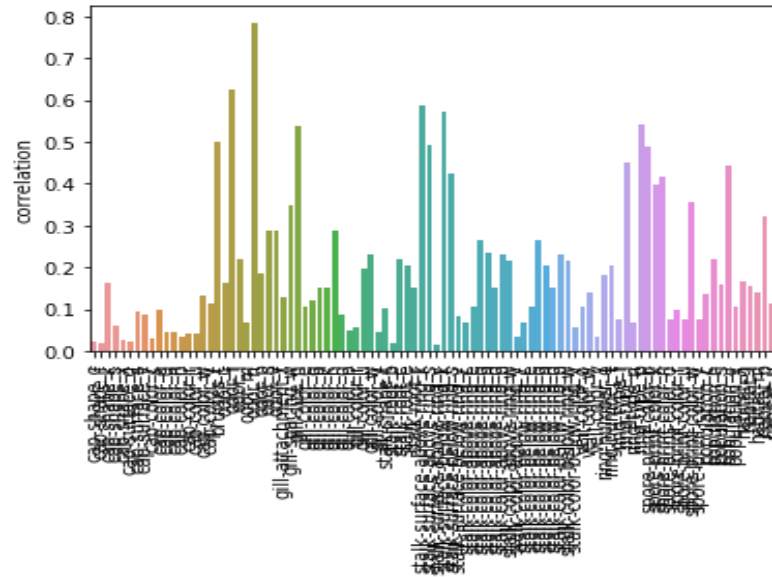
I. Mushroom data

2) EDA

Correlation



Feature importance



Feature selection

Feature	correlation
odor_n	0.785557
odor_f	0.623842
stalk-surface-above-ring_k	0.587658
stalk-surface-below-ring_k	0.573524
ring-type_p	0.540469
gill-size_n	0.540024
bruises_t	0.501530
stalk-surface-above-ring_s	0.491314
spore-print-color_h	0.490229
ring-type_l	0.451619
population_v	0.443722
stalk-surface-below-ring_s	0.425444
spore-print-color_n	0.416645

독성버섯의 결정적인 요인

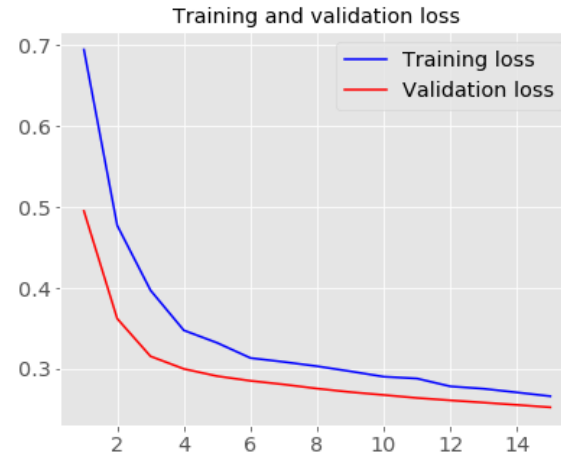
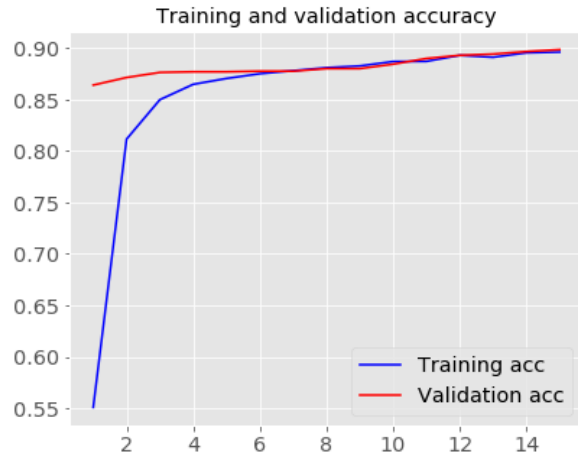
Odor_f : foul, Stalk-surface-above-ring_k : silky, Stalk-surface-below-ring_k : silky
 gill-size_n : narrow, spore-print-color_h : chocolate, ring-type_l : large
 population_v : several

식용가능한 버섯의 결정적 요인

Odor_n : none 무향, ring-type_p : pendant, bruise_t : 멍이 없는 경우
 stalk-surface-above-ring_s: smooth, stalk-surface-below-ring_s : smooth
 spore-print-color_n : brown

I. Mushroom data

3) 모형탐색&최적모형 선택



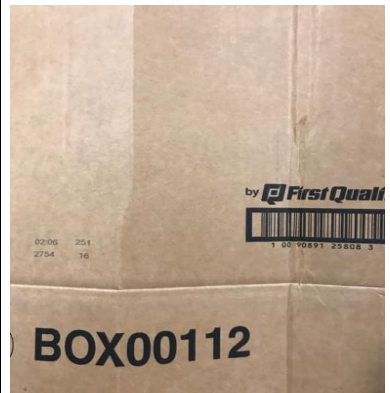





Model	Accuracy
Decision Tree	0.971
Random Forest	0.974
Support Vector Machine	0.978
Logistic regression	0.879
KNN	0.985
XGboost	0.985
CNN	0.896

- ➔ 총 7가지의 모형을 fitting 시켜 accuracy score를 비교해보았다.
- ➔ 과적합을방지하기위하여 GridSearchCV를 이용하여 하이퍼파라미터를 튜닝하여 모델을 적합시켰다.
- ➔ accuracy를 따져본 결과 점수가 제일 높은 SVM 모델을 최적모형으로 선택했다.

II. Garbage image data

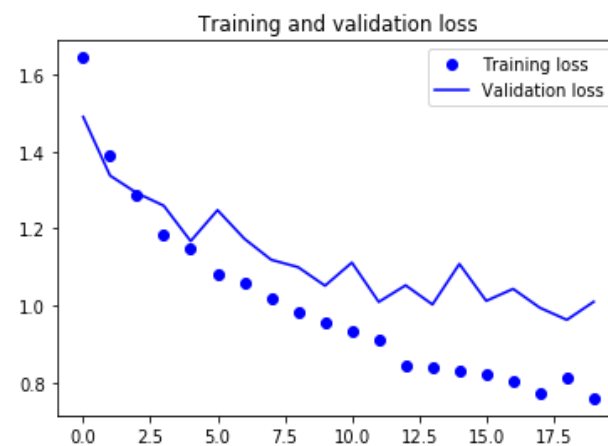
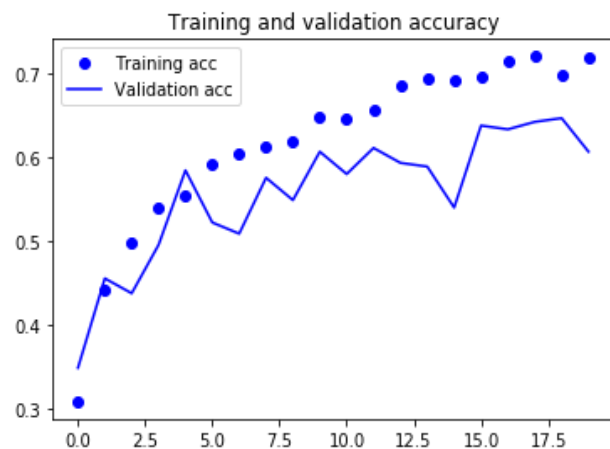
1) 자료설명

출처: <https://www.kaggle.com/asdasdasdas/garbage-classification>

cardboard	glass	metal	paper	plastic	trash
					

II. Garbage image data

2) 모형탐색 & 최적 모형선택

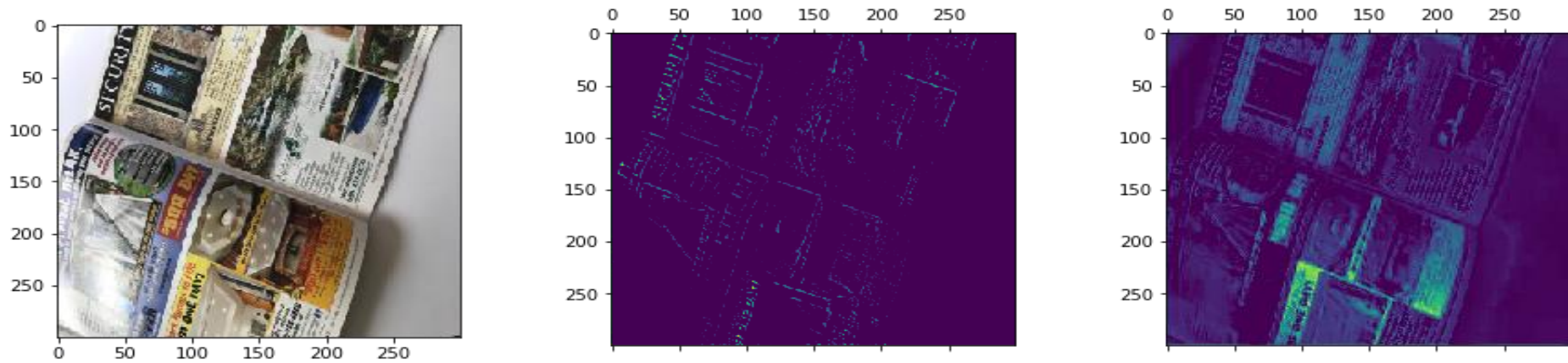


→ ImageDataGenerator 이용한 데이터 보강, dropout으로 조정, rotation range = 40

→ 데이터 보강후에 적합시킨 CNN 모형이 조금 더 accuracy가 높은 0.75이므로 최적모형으로 선택

II. Garbage image data

3) 합성망이 학습한 내용 시각화



→ Test data에 해당하는 이미지 paper380.jpg 입력

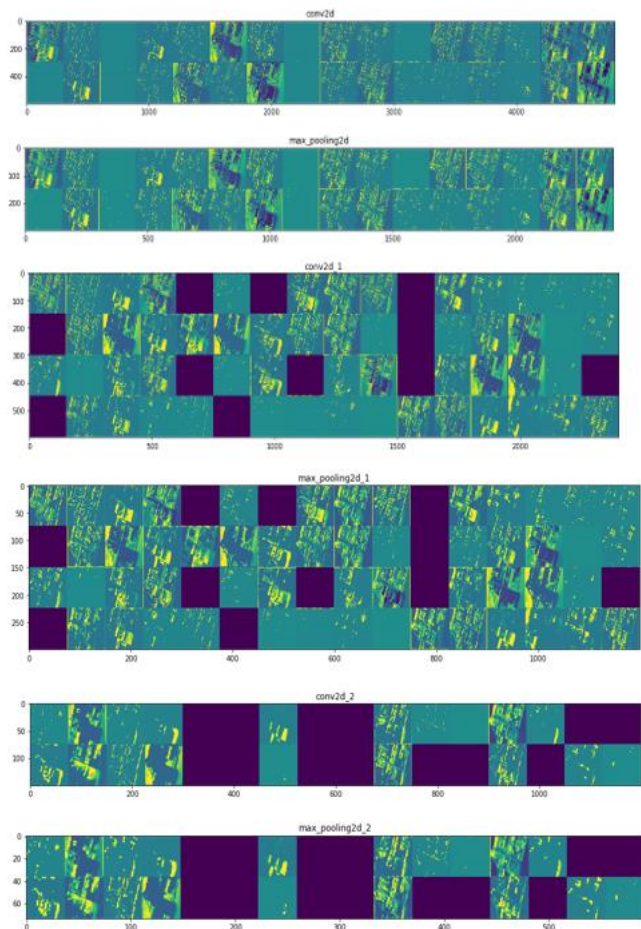
→ 32 채널, 300 x 300 특징지도

→ 순차적으로 채널 시각화

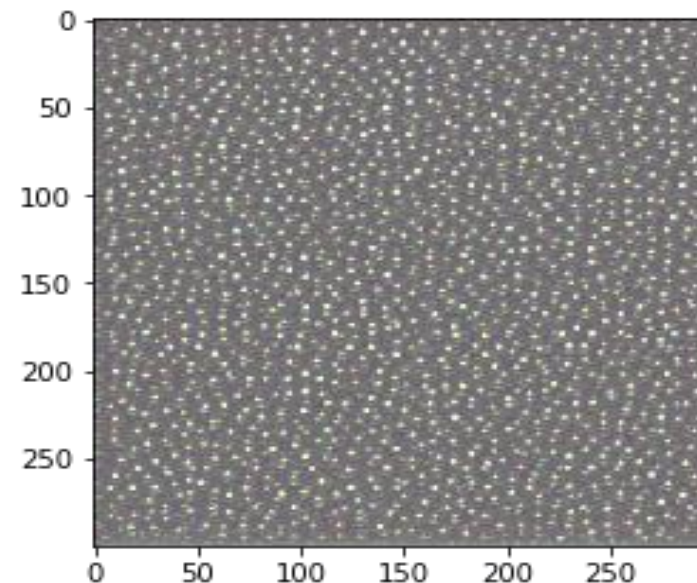
II. Garbage image data

3) 합성망이 학습한 내용 시각화

zip함수를 이용하여 특성 맵을 큰 그리드에 채운 각 활성화 채널



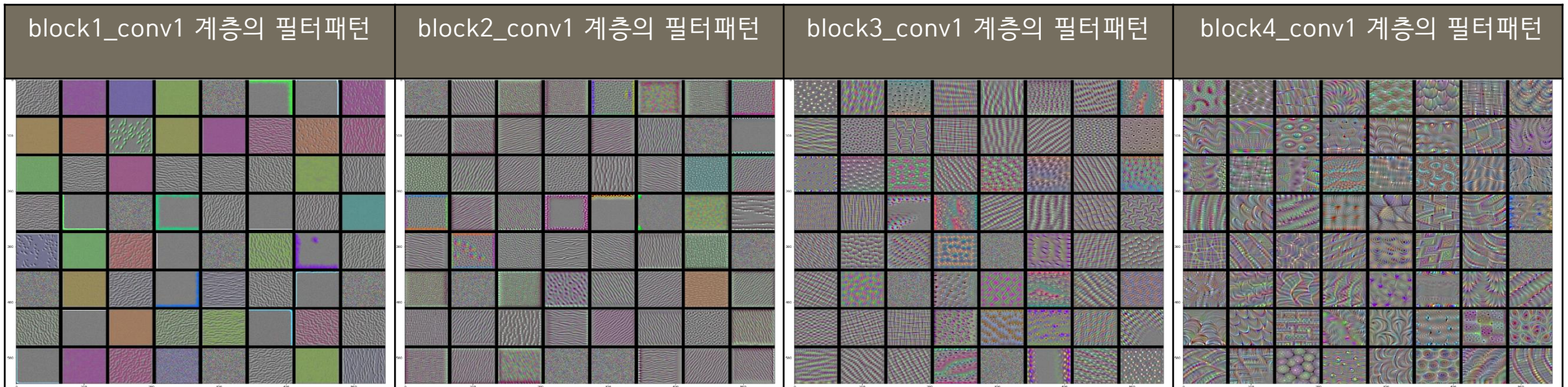
필터 시각화



II. Garbage image data

3) 합성망이 학습한 내용 시각화

계층内の 모든 필터 응답 패턴의 격자망 생성



- 감사합니다 -