CNN

Improvement of Classifying Pictures

CONTENTS

01

02

03

04

05

06

목표

환경 및 순서

데이터 수집

데이터 정제

모델구조

정확도 높이기

- 목표및방향

-개발환경

-진행순서

-웹 스크래핑

- -K-fold교차검증
- -Overfitting
- -Data

augmentaion

-부적절한 이미지 제거

- Model 1
- Model 2
- Model 3
- Model 4

-앙상블기법

개, 고양이 사진 분류에 있어서 AI가 분류하지 못한 사진을 추출하여 <mark>정확도를</mark> 향상시키는 것을 목표로 하고 있음.

방법 1

가중치 초기화, Conv층 설정, 러닝 레이트, 옵티마이저 설정 등

방법 2

분류하기 힘든(부적절한) 이미지 Remove 및 Relabeling

방향

방법1에서 분류하지 못한 이미지들을 방법2를 활용하여 정확도를 향상 시키면서 사람이 할 일을 줄여 주도록 한다. 운영체제

GPU 환경

IDE

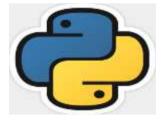
라이브러리









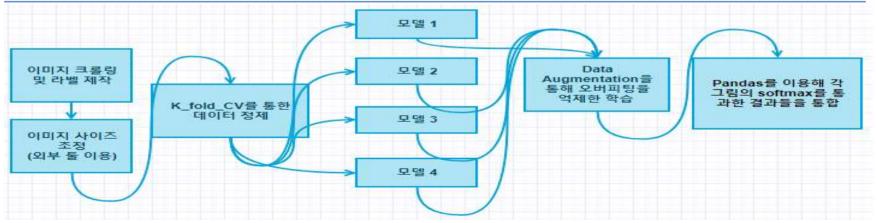






01 02 03 04 05 06

데이터 정제 데이터 학습 1 웹 스크래핑 데이터 학습 2 데이터 학습 3 전처리 전의 사진(3만장) 오버피팅 억제를 위한 만들어진 4개의 모델로 Python의 Selenium 패 정제된 데이터를 8:2로 으로 만든 95%대의 모델 부터 더 높은 정확도를 Data Augmentation € 키지를 이용한 개, 고양 나눈뒤, 미리 만들어둔 을 가지고 K_fold CV를 opencv로 구현해 train에 얻기 위해 Soft Voting 이 사진 수집 모델 4개를 이용해 학습 통해 불량 데이터 삭제 적용 을 이용



01

02

03

04

V

06

웹 스크래핑

웹 크롤링(web crawling)

자동화 봇(bot)인 웹 크롤러가 정해진 규칙에 따라 복수 개의 웹 페이지를 브라우징 하는 행위 링크를 따라 돌면서, 연결된 페이지를 가져오는 과정

웹 스크래핑(web scraping

조직적, 자동화된 방법으로 월드 와이드 웹(WWW)을 탐색하는 소프트웨어 기술. 웹 크롤러가 가져오는 하나의 웹 페이지가 있을 때, 추출하길 원하는 항목의 위치를 지정해서, 데이터로 가져오는 것

웹 스크래핑 과정

laport urillib request from bed laport BeautifulSoup from selection import webdriver from selenium mebdriver.common.keys import Keys inport time binary = 'D-Schroeedriver/chroeedriver.exe' browser - webdriver (Otrone(binary) browser.get ("http://search.dom.net/search?nil_suggest-btnik-lag80k-0906 eles - browser find elesent by id("a") A STATISTICS STATE eles send_keys("BELE") eles subsit() 多. 提出数 香香 for I in range(1 .3) browser find element by xpath("//body") send keys(Keys EMI) time.sleep(20) time.sigep(20) html = browser.page_source soup - Beautiful Soup(html . "Imi(") dof fetch list utif(): barass = [] inglist - soup find all [lag", class - thus lag"] for in in IngList: parairs append(in["src"]) return parais def fetch_detail_uritit params - fetch_Hist_url() print(parans) 8for p in parees: prist (p) # DEWE #GRE BY url lib.request universiteve(p. "d /dna langes/"+ str(a) + ".ipg") fetch_detail_url() browser.quit()

INVOKE



AUTOMATES



K-fold 교차 검증

K-fold 란?

"K-fold"는 K개의 fold를 만들어서 검증하는 방법



데이터셋을 여러 개로 나누어 하나씩 테스트셋으로 사용하고 **나머지를 모두 합해서 학습셋으로 사용**하는 방법

장점

- 정확도 향상
- 데이터 수가 적은데 검증과 테스트에 데이터를 더 뺐기면 underfitting등 성능이 미달되는 모델이 학습됨

단점

- 시간이 오래 걸린다

K-fold 교차 검증 과정

Fold

Test 1	Train	Train	Train	Train	Test 1
Train	Test 2	Train	Train	Train	Test 2
Train	Train	Test 3	Train	Train	Test 3
Train	Train	Train	Test 4	Train	Test 4
Train	Train	Train	Train	Test 5	Test 5

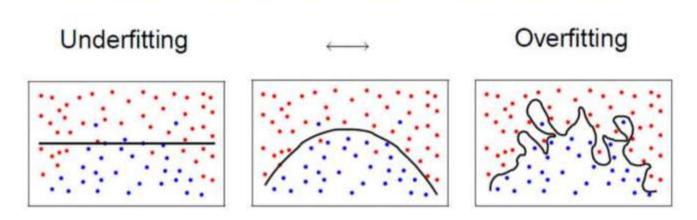
- 1. K개의 fold를 나눈다 (5개)
- 2. Train과 Test를 나눈다.
- 3. 다음 fold에서는 Test 셋을 바꿔서 지정하고, 이전 fold에서 Test 역할을 했던 set은 다시 train으로 사용한다.
- 4. 위의 방법을 K번 반복한다.

Overfitting

1) Overfitting 이란?

"Overfitting"은 공통적인 특성 이외에 지엽적인 특성도 반영하여 과하게 학습되어, 새로운 데이터에 대해 예측하지 못함.

Generalization Problem in Classification



Overfitting

- 2) Overfitting 해결방법?
 - 많은 학습 데이터를 넣기
 - 피쳐의 수를 줄이기
 - 정규화 이용

학습 데이터를 증식 시키는 방법 중에 Data augmentaion이 있다.

Data augmentaion

1) Data augmentaion 이란?

"Data augmentaion" 데이터 학습시 (학습마다) 이미지 데이터에 대조,회전 등 데이터의 변화를 줘서 단순한 원본 사진 한 장이 아닌 변화를 준 여러 사진을 학습시켜 학습량을 늘리는 방법

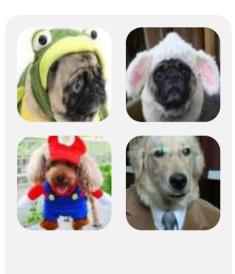


데이터 정제 01 02 03 **04** 05 06

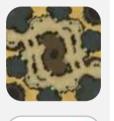
데이터 정제 과정에서 걸러진 사진들

약 600장 제거!!











모델	CNN
lr	0.001
batch	100
ор	Moment
가중치초기값	Truncated random (절단 정규분포)
dropout	0.5



1층	conv(16)
	BN
	relu
	max_pool
ĺ	dropout
	conv(32)
	BN
2층	relu
	max_pool
	dropout
	conv(64)
	BN
3층	relu
	max_pool
	dropout
	conv(128)
	BN
4층	relu
	max_pool
	dropout
5층	FC(256)
6층	FC(512)
7층	FC(1024)

모델	CNN
lr	0.0001
batch	100
ор	Adam
가중치초기값	Truncated random (절단 정규분포)
dropout	0.5



정제전정확도약93%

1층	conv(30)
	BN
	relu
	max_pool
	dropout
	conv(40)
	BN
2층	relu
	max_pool
	dropout
	conv(50)
	BN
3층	relu
	max_pool
	dropout
	conv(60)
	BN
4층	relu
	max_pool
	dropout
	conv(100)
5층	BN
9 5	relu
	dropout
6층	conv(100)
	BN
	relu
	dropout
7층	FC(200)

모델	CNN
lr	0.001
batch	100
ор	Adam
가중치초기값	He
dropout	0.5



Conv(8) BN relu max_pool Conv(16) BN relu max_pool Conv(16) BN relu max_pool Conv(32) BN relu max_pool Conv(64) BN relu max_pool Conv(64) BN relu max_pool Conv(128) BN relu max_pool Conv(128) BN relu Conv(128) BN relu Conv(128) Conv(
relu max_pool conv(16) BN relu max_pool relu max_pool conv(32) BN relu max_pool conv(64) BN relu max_pool conv(128) BN relu dropout 6층 FC(200)	1층	conv(8)
relu max_pool conv(16) BN relu max_pool conv(32) BN relu max_pool conv(64) BN relu max_pool conv(64) BN relu max_pool conv(128) BN relu dropout 6층 FC(200)		BN
2층		relu
2층 BN relu max_pool conv(32) 3층 BN relu max_pool conv(64) 4층 BN relu max_pool conv(64) 4층 BN relu max_pool conv(128) 5층 BN relu dropout 6층 FC(200)		max_pool
relu max_pool conv(32) BN relu max_pool conv(64) 4층 BN relu max_pool conv(128) BN relu dropout 6층 FC(200)		conv(16)
relu max_pool conv(32) BN relu max_pool conv(64) 4층 BN relu max_pool conv(128) BN relu dropout 6층 FC(200)	9초	BN
S층 Conv(32) BN relu max_pool conv(64) BN relu BN relu max_pool conv(128) BN relu dropout 6층 FC(200)	25	relu
3층 BN relu max_pool conv(64) 4층 BN relu max_pool conv(128) 5층 BN relu dropout 6층 FC(200)		max_pool
### Red		conv(32)
relu max_pool conv(64) BN relu max_pool conv(128) BN relu dropout 6층 FC(200)	2.5	BN
## Conv(64) ## BN ## relu ## max_pool ## conv(128) ## BN ## relu ## dropout ## 6층 ## FC(200)	ve	relu
4층 BN relu max_pool conv(128) BN relu dropout 6층 FC(200)		max_pool
relu max_pool conv(128) BN relu dropout 6층 FC(200)		conv(64)
relu max_pool conv(128) BN relu dropout 6층 FC(200)	/ 츠	BN
conv(128) BN relu dropout 6층 FC(200)	40	relu
5층 BN relu dropout 6층 FC(200)		max_pool
5층 relu dropout 6층 FC(200)	5층	conv(128)
relu dropout 6층 FC(200)		BN
6층 FC(200)		relu
		dropout
7층 FC(200)	6층	FC(200)
, , ,	7층	FC(200)

모델	CNN
lr	0.001
batch	100
ор	Adam
가중치초기값	He
dropout	0.8



1층	conv(20)
	BN
	leaky_relu
	max_pool
	dropout
	conv(40)
	BN
2층	leaky_relu
	max_pool
	dropout
	conv(60)
	BN
3층	leaky_relu
	max_pool
	dropout
	conv(80)
	BN
4층	leaky_relu
	max_pool
	dropout
5층	conv(100)
	BN
	leaky_relu
	dropout
6층	FC(200)

앙상블 기법 (Boosting)

앙상블이란?

여러 모델을 사용하여 데이터를 학습 시키고, 모든 모델의 예측 결과를평균하여 예측

Bagging

각 부스트랩 자룔를 모델링 한 후 결합하여 최종의 예측 모형을 산출하는 방법

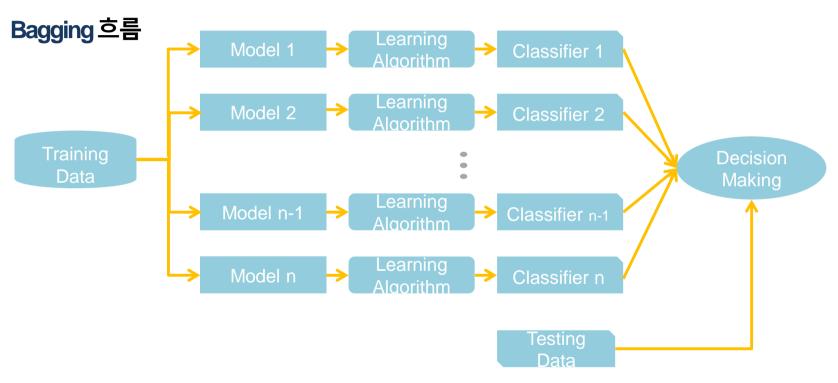
특징

여러번의 복워 누샘플링을 통해 예측 모형의 분산을 줄여 줌으로써 예측력을 향상

목적

예측 모형의 변동성을 감소시키 위해 사용

Bagging



모델	CNN
lr	0.001
batch	100
ор	Adam
가중치초기값	He
dropout	0.8



정제후정확도약96%

1층	conv(20)
	BN
	leaky_relu
	max_pool
	dropout
	conv(40)
	BN
2층	leaky_relu
	max_pool
	dropout
	conv(60)
	BN
3층	leaky_relu
	max_pool
	dropout
	conv(80)
	BN
4층	leaky_relu
	max_pool
	dropout
5층	conv(100)
	BN
	leaky_relu
	dropout
6층	FC(200)

THANK YOU