Techtonic 2018

-

Thu . Nov 15

-

SAMSUNG SDS Tower
West Campus B1F
Magellan Hall / Pascal Hall

SAMSUNG SDS
Realize your vision

Machine Learning 경진대회 Kaggle 도전기

삼성SDS 박정형

Techtonic 2018

Agenda

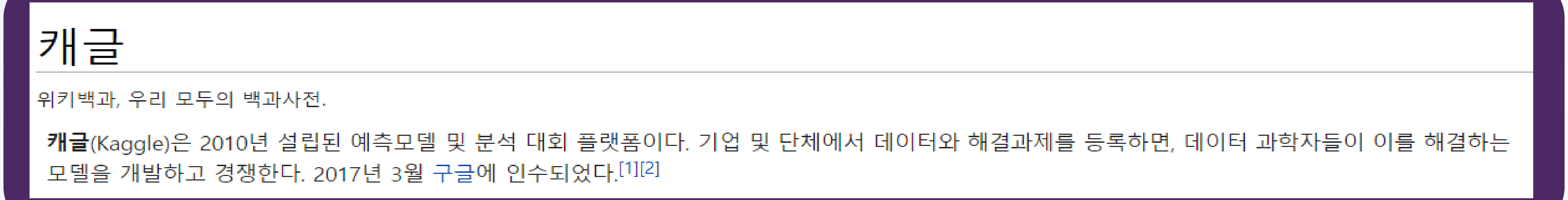
- Kaggle 소개
- Speech Recognition Challenge 참가기
- 참가 후기 및 향후 계획

Machine Learning 경진대회-Kaggle 도전기

Kaggle 소개

Kaggle 개요

도대체 Kaggle이 뭔가요?



캐글

위키백과, 우리 모두의 백과사전.

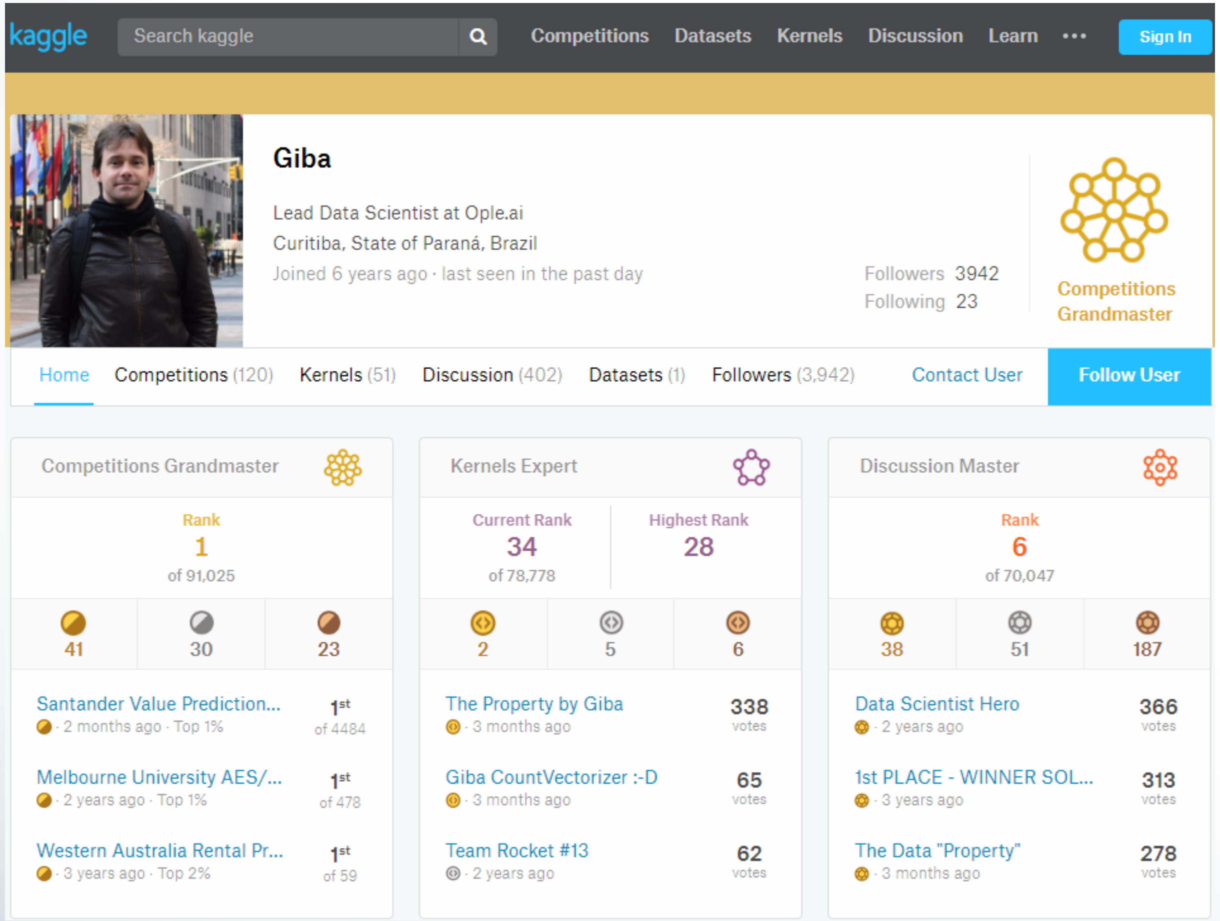
캐글(Kaggle)은 2010년 설립된 예측모델 및 분석 대회 플랫폼이다. 기업 및 단체에서 데이터와 해결과제를 등록하면, 데이터 과학자들이 이를 해결하는 모델을 개발하고 경쟁한다. 2017년 3월 구글에 인수되었다.[1][2]

안소니 골드블룸 님에 의해 만들어진 데이터 분석 경진 대회 플랫폼

과연 나는 kaggle을 통해 얻을 수 있는 게 무엇일까요?

Kaggle 개요

캐글 깨글거리는데 이거 잘하면 뭐가 좋아요?



kaggle Search kaggle Competitions Datasets Kernels Discussion Learn ••• Sign In

Giba

Lead Data Scientist at Ople.ai
Curitiba, State of Paraná, Brazil
Joined 6 years ago · last seen in the past day

Followers 3942
Following 23

Competitions Grandmaster

Home Competitions (120) Kernels (51) Discussion (402) Datasets (1) Followers (3,942) Contact User Follow User

Competitions Grandmaster

Rank
1
of 91,025

41 30 23

Santander Value Prediction... · 2 months ago · Top 1% — 1st of 4484
Melbourne University AES/... · 2 years ago · Top 1% — 1st of 478
Western Australia Rental Pr... · 3 years ago · Top 2% — 1st of 59

Kernels Expert

Current Rank 34 of 78,778 — Highest Rank 28

2 5 6

The Property by Giba · 3 months ago — 338 votes
Giba CountVectorizer :-D · 3 months ago — 65 votes
Team Rocket #13 · 2 years ago — 62 votes

Discussion Master

Rank
6
of 70,047

38 51 187

Data Scientist Hero · 2 years ago — 366 votes
1st PLACE - WINNER SOL... · 3 years ago — 313 votes
The Data "Property" · 3 months ago — 278 votes

경진대회에 많이 참여 → 높은 점수를 많이 획득 → 가입된 계정 스코어 ↑

1년전 status: Data Scientist at Airbnb
Now : Lead Data Scientist at Ople.ai

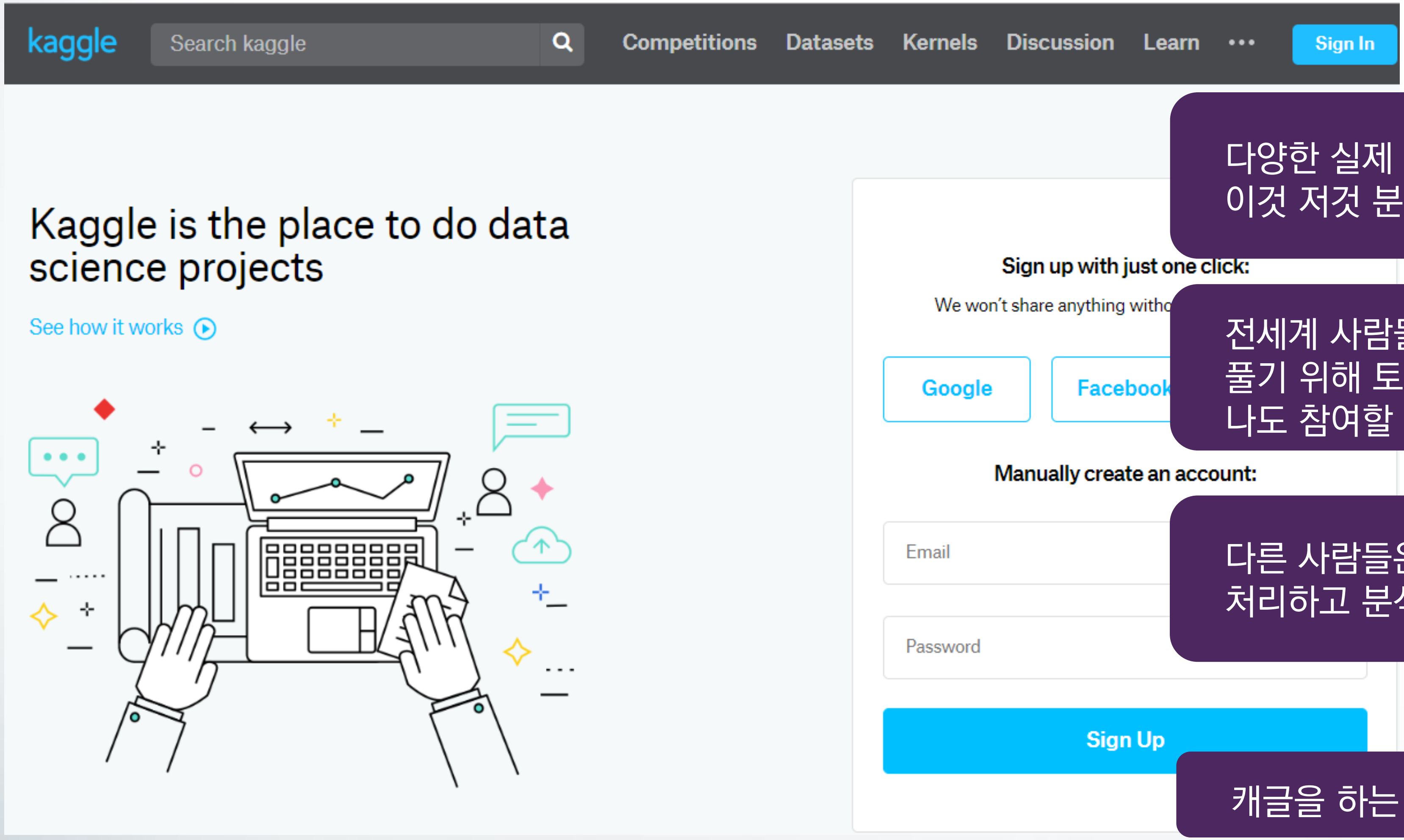
캐글 상위 등재시,
굴지의 글로벌 IT 기업

Google IBM SAMSUNG

직접 모셔 가는 경우가 다반사. WOW !

Kaggle 개요

Kaggle.com에 가 보았습니다.



kaggle Search kaggle Competitions Datasets Kernels Discussion Learn ••• Sign In

Kaggle is the place to do data science projects

See how it works

Sign up with just one click:

We won't share anything witho

Google Facebook

Manually create an account:

Email

Password

Sign Up

다양한 실제 데이터를 직접 눈으로 보고
이것 저것 분석해 볼 수 있을 것 같아

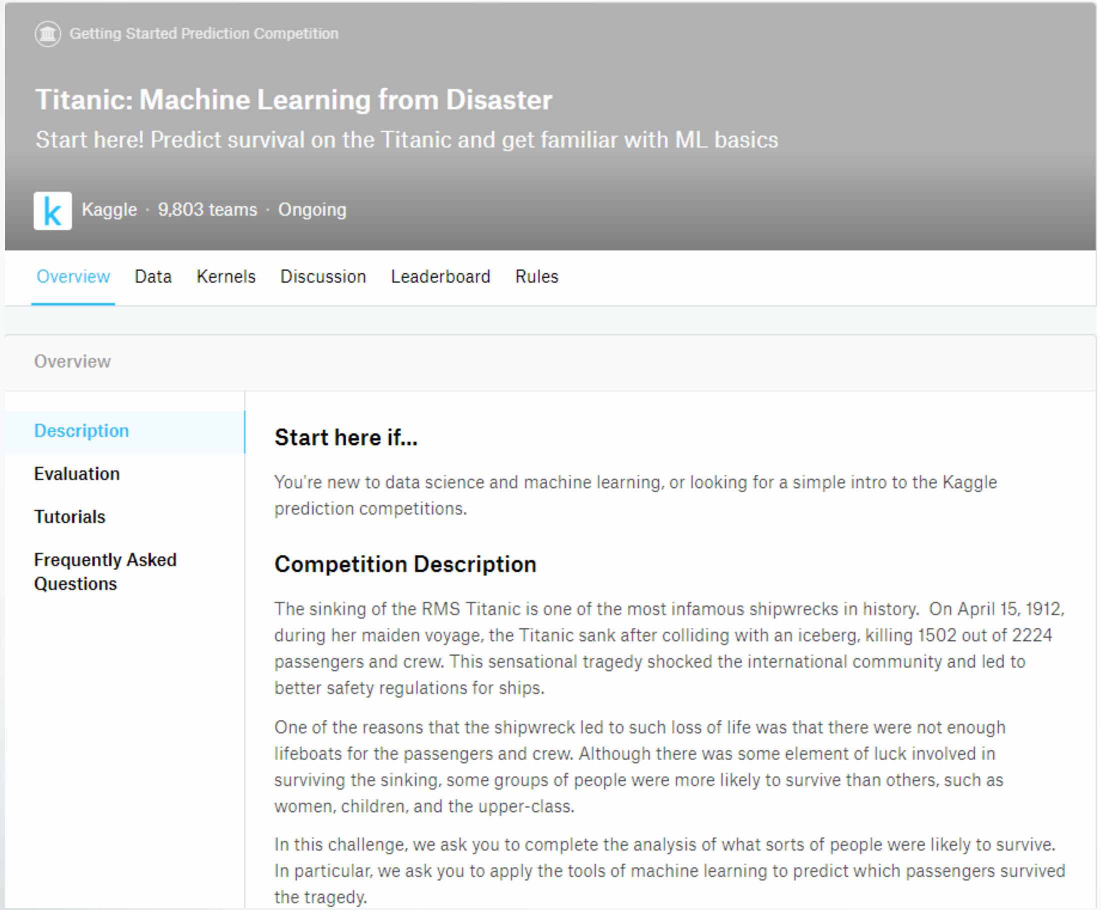
전세계 사람들이 온라인상에서 같은 문제를
풀기 위해 토론을 하는 걸 볼 수도 있네.
나도 참여할 수 있나?

다른 사람들은 파이썬으로 어떻게 데이터를
처리하고 분석하는지 한번 봐 볼까?

캐글을 하는 사람.. 캐글러 한번 되어볼까?

Kaggle competition 시작

시작은 가볍게 캐글러들이 다 도전해 본다는 그 문제



Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 9,803 teams · Ongoing

Start here if...

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.

Competition Description

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

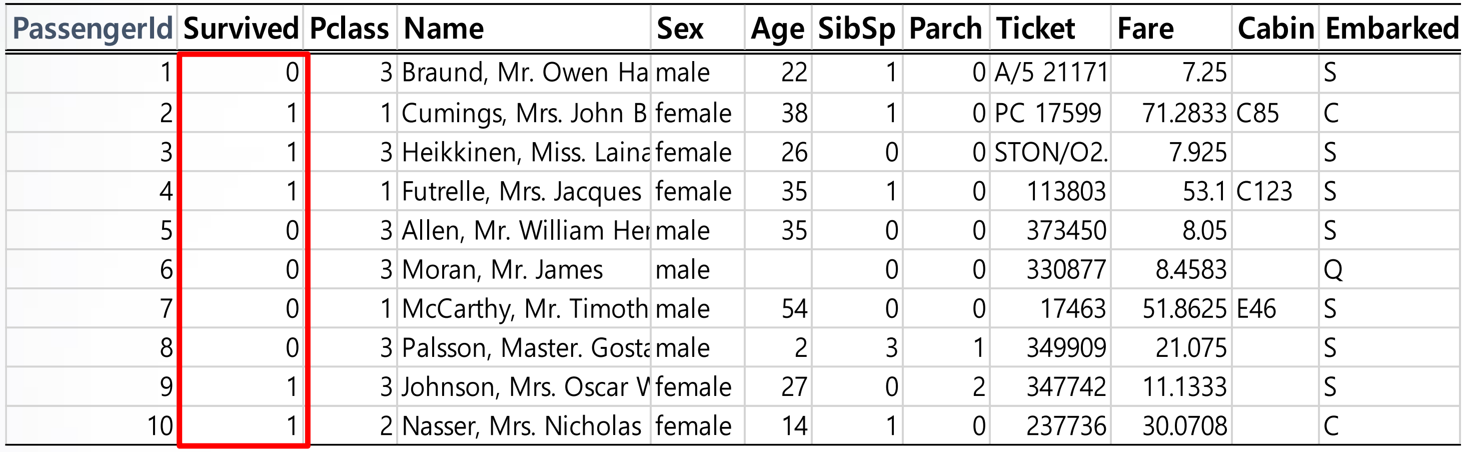
One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

9803팀이나 참여한 타이타닉 프로젝트
타이타닉호 탑승객의 생존율을 예측하라!

Titanic호 탑승객 분석

탑승객의 뭘 어떻게 분석해서 생존율을 예측하는 거지?



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Ha	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John B	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Her	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timoth	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar V	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas	female	14	1	0	237736	30.0708		C

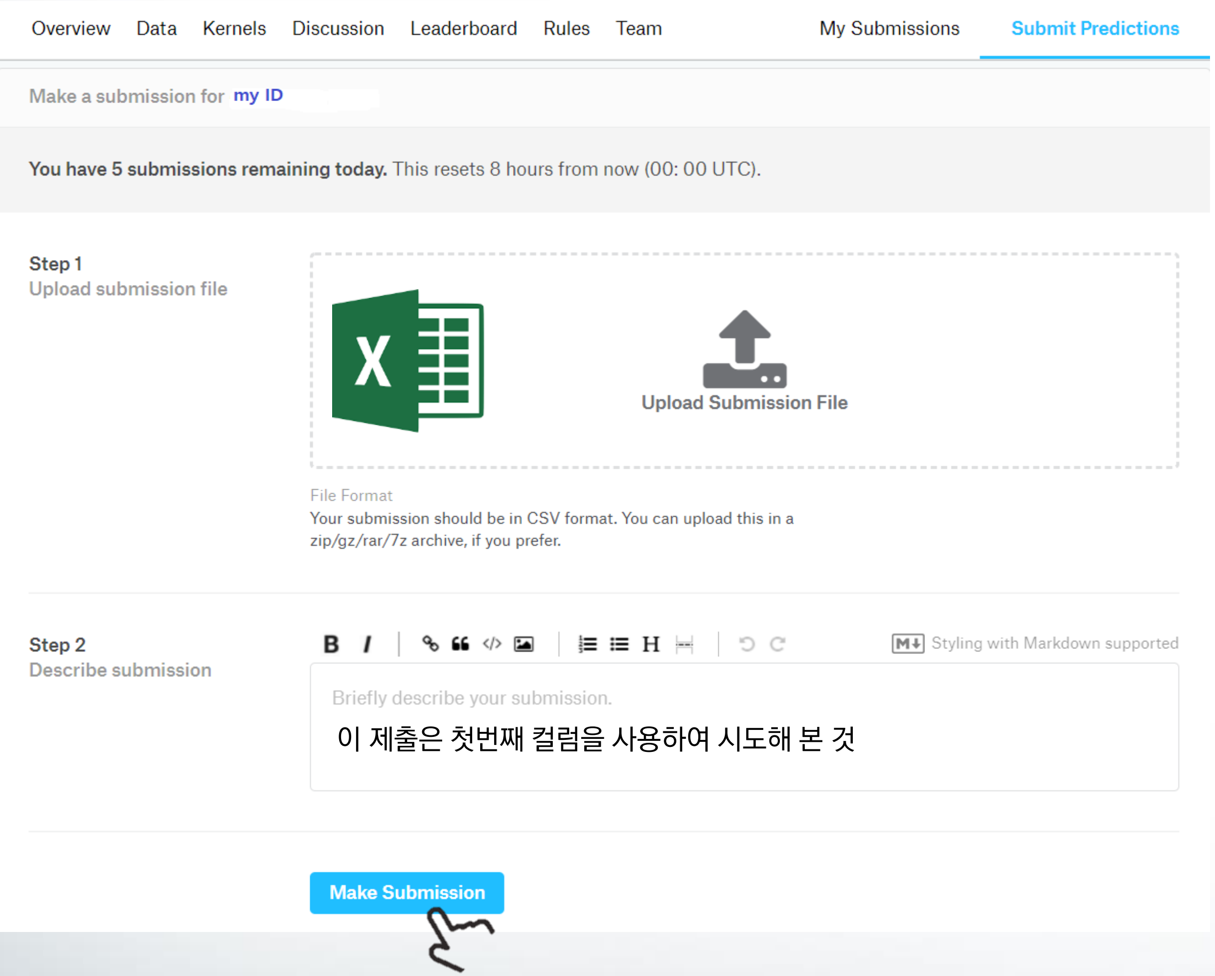
이 train 데이터를 바탕으로 survived 칸을 예측?

빈칸들이 있는데 어떻게 처리하면 좋을까?

PassengerID : Unique Integer. Up to 891
Survived : Survived (1) or not (0)
Pclass : ticket class 1st, 2nd, 3rd
SibSp : 같이 탑승중인 siblings, spouse 수
Parch : 같이 탑승중인 parents, children 수
Embarked : C = Cherbourg, Q = Queenstown,
S = Southampton

Submit prediction

생존자 예측을 해 봤으니 답안지 제출을 해 볼까요?



Make a submission for my ID

You have 5 submissions remaining today. This resets 8 hours from now (00: 00 UTC).

Step 1
Upload submission file

File Format
Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

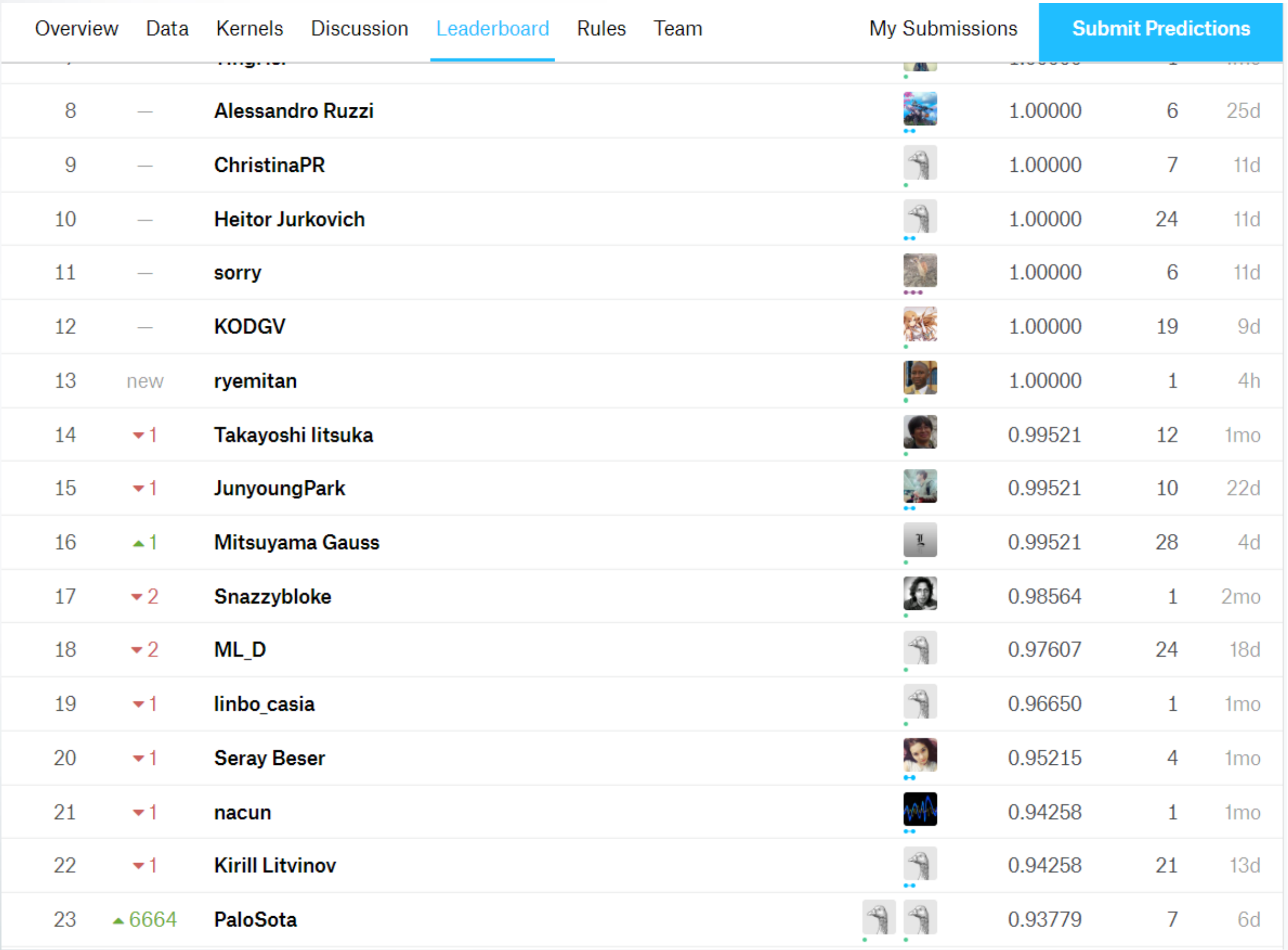
Step 2
Describe submission

Briefly describe your submission.

이 제출은 첫번째 컬럼을 사용하여 시도해 본 것

결과 확인 leaderboard

제출된 내 결과는 과연 어떨까요?



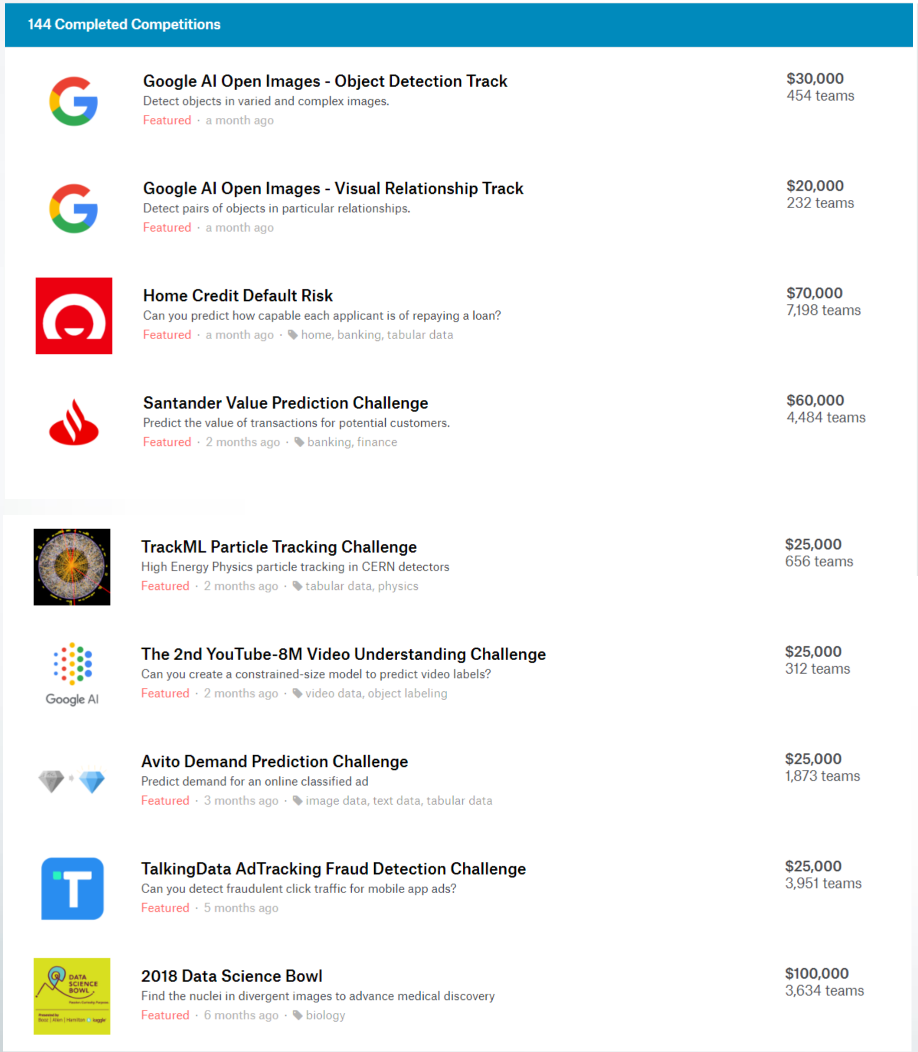
8	—	Alessandro Ruzzi	1.00000	6	25d
9	—	ChristinaPR	1.00000	7	11d
10	—	Heitor Jurkovich	1.00000	24	11d
11	—	sorry	1.00000	6	11d
12	—	KODGV	1.00000	19	9d
13	new	ryemitan	1.00000	1	4h
14	▾1	Takayoshi Iitsuka	0.99521	12	1mo
15	▾1	JunyoungPark	0.99521	10	22d
16	▴1	Mitsuyama Gauss	0.99521	28	4d
17	▾2	Snazzybloke	0.98564	1	2mo
18	▾2	ML_D	0.97607	24	18d
19	▾1	linbo_casia	0.96650	1	1mo
20	▾1	Seray Beser	0.95215	4	1mo
21	▾1	nacun	0.94258	1	1mo
22	▾1	Kirill Litvinov	0.94258	21	13d
23	▴6664	PaloSota	0.93779	7	6d

짜잔~!
실시간 채점 결과로 제출답안에 대한
전체 스코어, 현재 등수 파악 완료

단, 내가 낸 답의 정/오답을 알 순 없음..
알고싶다 ㅠ.ㅠ

Kaggle competitions

개최되었던 competitions 약 144개, 진행중인 competitions 약 8개



144 Completed Competitions

Google AI Open Images - Object Detection Track
Detect objects in varied and complex images.
Featured · a month ago
$30,000
454 teams

Google AI Open Images - Visual Relationship Track
Detect pairs of objects in particular relationships.
Featured · a month ago
$20,000
232 teams

Home Credit Default Risk
Can you predict how capable each applicant is of repaying a loan?
Featured · a month ago · home, banking, tabular data
$70,000
7,198 teams

Santander Value Prediction Challenge
Predict the value of transactions for potential customers.
Featured · 2 months ago · banking, finance
$60,000
4,484 teams

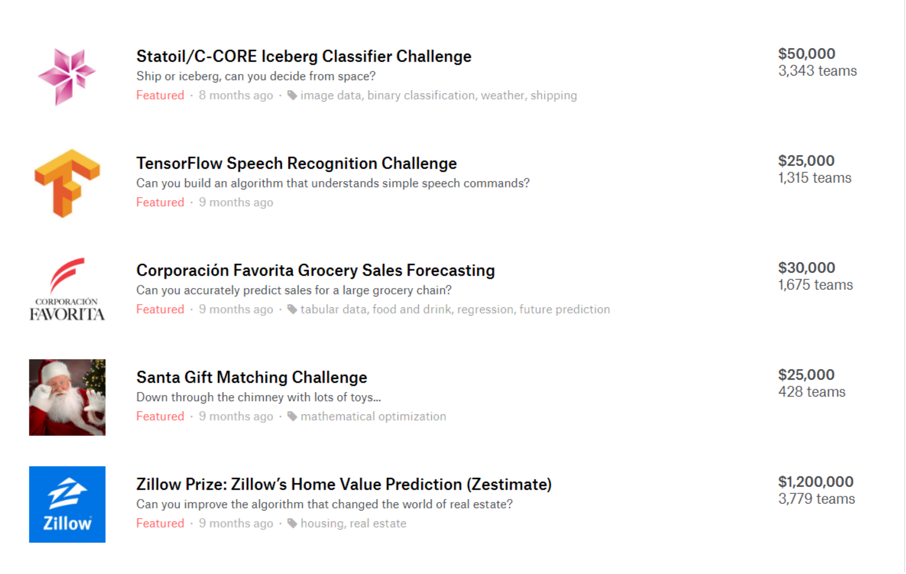
TrackML Particle Tracking Challenge
High Energy Physics particle tracking in CERN detectors
Featured · 2 months ago · tabular data, physics
$25,000
656 teams

The 2nd YouTube-8M Video Understanding Challenge
Can you create a constrained-size model to predict video labels?
Featured · 2 months ago · video data, object labeling
$25,000
312 teams

Avito Demand Prediction Challenge
Predict demand for an online classified ad
Featured · 3 months ago · image data, text data, tabular data
$25,000
1,873 teams

TalkingData AdTracking Fraud Detection Challenge
Can you detect fraudulent click traffic for mobile app ads?
Featured · 5 months ago
$25,000
3,951 teams

2018 Data Science Bowl
Find the nuclei in divergent images to advance medical discovery
Featured · 6 months ago · biology
$100,000
3,634 teams



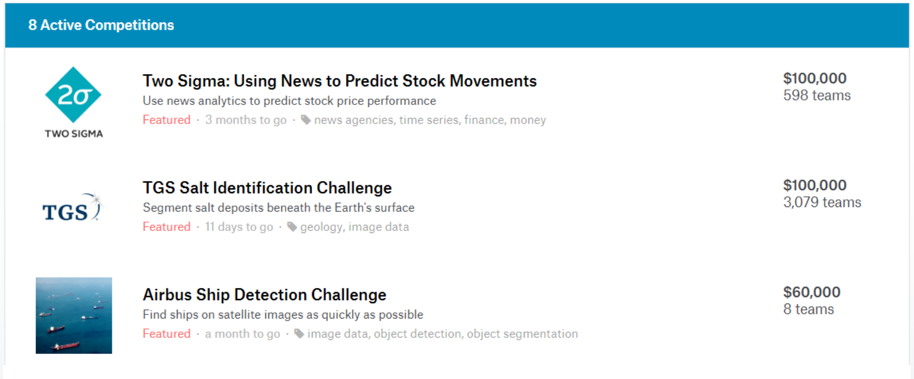
Statoil/C-CORE Iceberg Classifier Challenge
Ship or iceberg, can you decide from space?
Featured · 8 months ago · image data, binary classification, weather, shipping
$50,000
3,343 teams

TensorFlow Speech Recognition Challenge
Can you build an algorithm that understands simple speech commands?
Featured · 9 months ago
$25,000
1,315 teams

Corporación Favorita Grocery Sales Forecasting
Can you accurately predict sales for a large grocery chain?
Featured · 9 months ago · tabular data, food and drink, regression, future prediction
$30,000
1,675 teams

Santa Gift Matching Challenge
Down through the chimney with lots of toys...
Featured · 9 months ago · mathematical optimization
$25,000
428 teams

Zillow Prize: Zillow's Home Value Prediction (Zestimate)
Can you improve the algorithm that changed the world of real estate?
Featured · 9 months ago · housing, real estate
$1,200,000
3,779 teams



8 Active Competitions

Two Sigma: Using News to Predict Stock Movements
Use news analytics to predict stock price performance
Featured · 3 months to go · news agencies, time series, finance, money
$100,000
598 teams

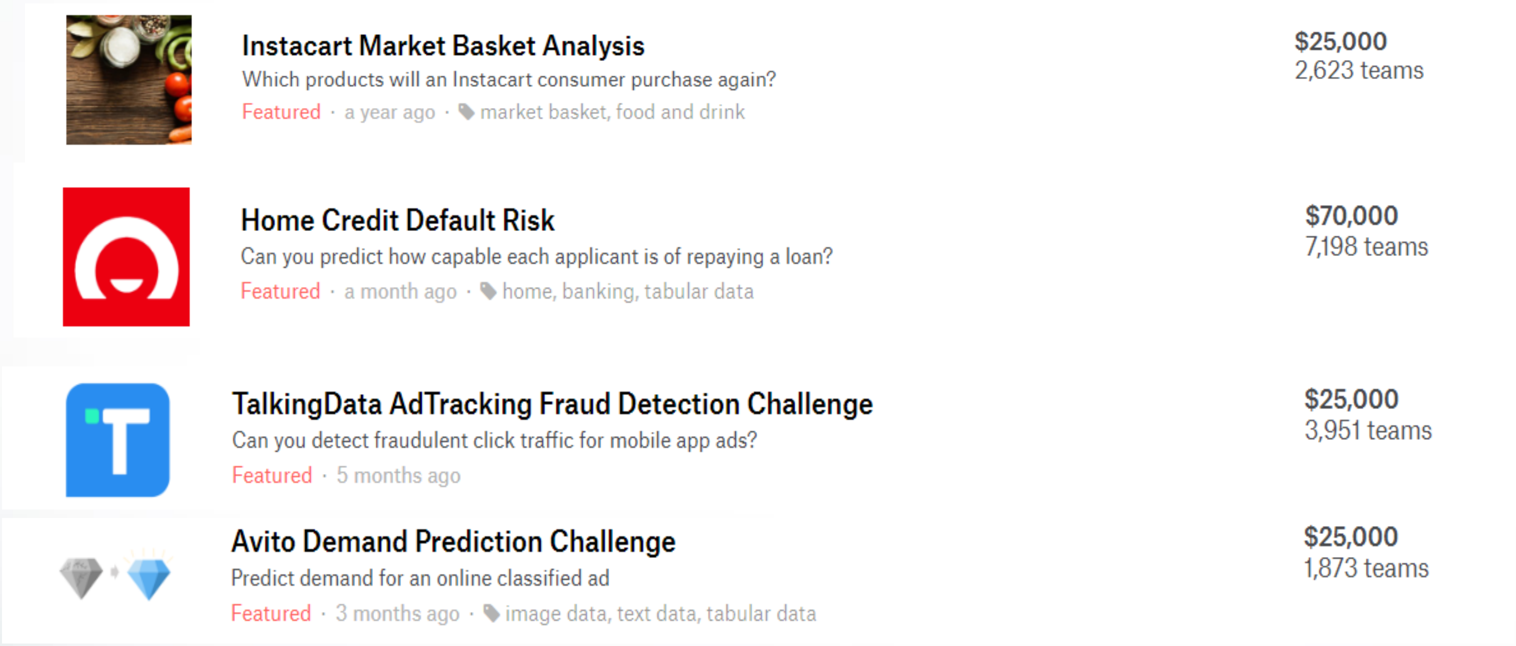
TGS Salt Identification Challenge
Segment salt deposits beneath the Earth's surface
Featured · 11 days to go · geology, image data
$100,000
3,079 teams

Airbus Ship Detection Challenge
Find ships on satellite images as quickly as possible
Featured · a month to go · image data, object detection, object segmentation
$60,000
8 teams

참가해 본 kaggle competitions

업무 후 잠깐씩 시간을 내서 도전해 봤던 competitions

Machine learning 기반 tabular data 분석



Instacart Market Basket Analysis
Which products will an Instacart consumer purchase again?
Featured · a year ago · market basket, food and drink
$25,000
2,623 teams

Home Credit Default Risk
Can you predict how capable each applicant is of repaying a loan?
Featured · a month ago · home, banking, tabular data
$70,000
7,198 teams

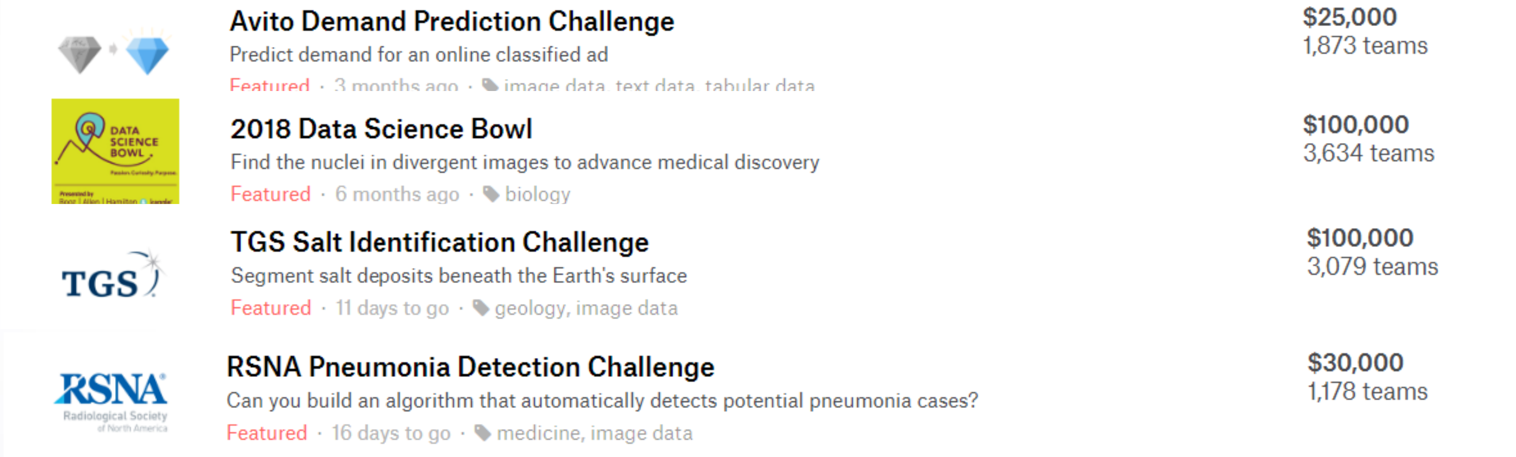
TalkingData AdTracking Fraud Detection Challenge
Can you detect fraudulent click traffic for mobile app ads?
Featured · 5 months ago
$25,000
3,951 teams

Avito Demand Prediction Challenge
Predict demand for an online classified ad
Featured · 3 months ago · image data, text data, tabular data
$25,000
1,873 teams

참가해 본 kaggle competitions

풀어볼 만한 문제가 정말 끝이 없는 듯

Deep learning 기반 image data 분석



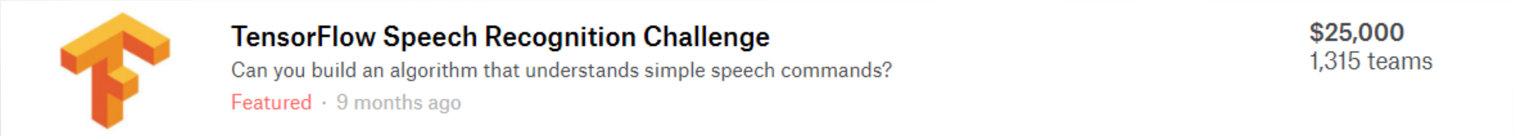
Avito Demand Prediction Challenge
Predict demand for an online classified ad
Featured · 3 months ago · image data, text data, tabular data
$25,000
1,873 teams

2018 Data Science Bowl
Find the nuclei in divergent images to advance medical discovery
Featured · 6 months ago · biology
$100,000
3,634 teams

TGS Salt Identification Challenge
Segment salt deposits beneath the Earth's surface
Featured · 11 days to go · geology, image data
$100,000
3,079 teams

RSNA Pneumonia Detection Challenge
Can you build an algorithm that automatically detects potential pneumonia cases?
Featured · 16 days to go · medicine, image data
$30,000
1,178 teams

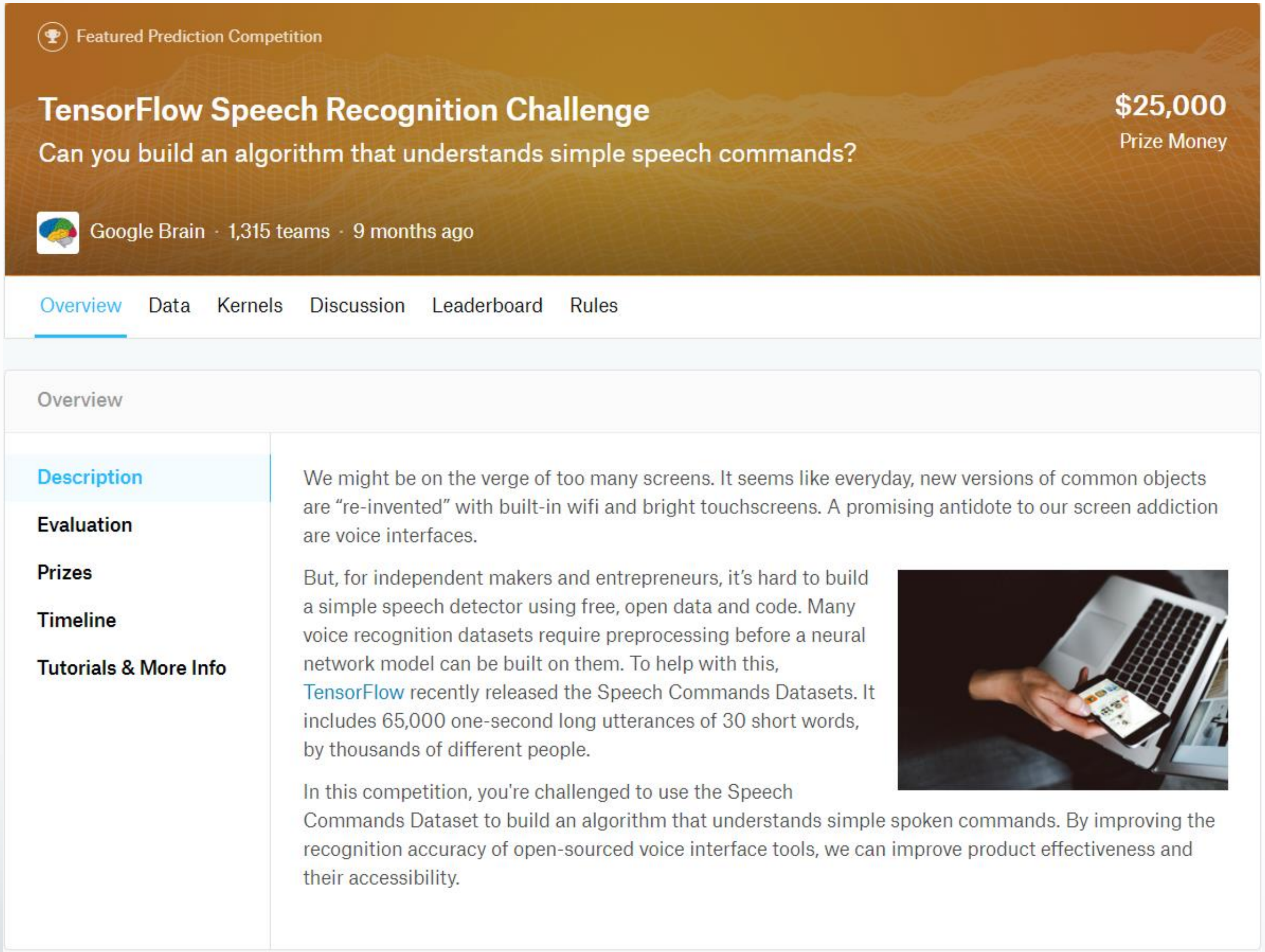
Deep learning 기반 speech data 분석



TensorFlow Speech Recognition Challenge
Can you build an algorithm that understands simple speech commands?
Featured · 9 months ago
$25,000
1,315 teams

Speech recognition competition 참가기

Competition 문제 정의



Featured Prediction Competition

TensorFlow Speech Recognition Challenge

Can you build an algorithm that understands simple speech commands?

$25,000
Prize Money

Google Brain · 1,315 teams · 9 months ago

Overview

We might be on the verge of too many screens. It seems like everyday, new versions of common objects are "re-invented" with built-in wifi and bright touchscreens. A promising antidote to our screen addiction are voice interfaces.

But, for independent makers and entrepreneurs, it's hard to build a simple speech detector using free, open data and code. Many voice recognition datasets require preprocessing before a neural network model can be built on them. To help with this, TensorFlow recently released the Speech Commands Datasets. It includes 65,000 one-second long utterances of 30 short words, by thousands of different people.

In this competition, you're challenged to use the Speech Commands Dataset to build an algorithm that understands simple spoken commands. By improving the recognition accuracy of open-sourced voice interface tools, we can improve product effectiveness and their accessibility.

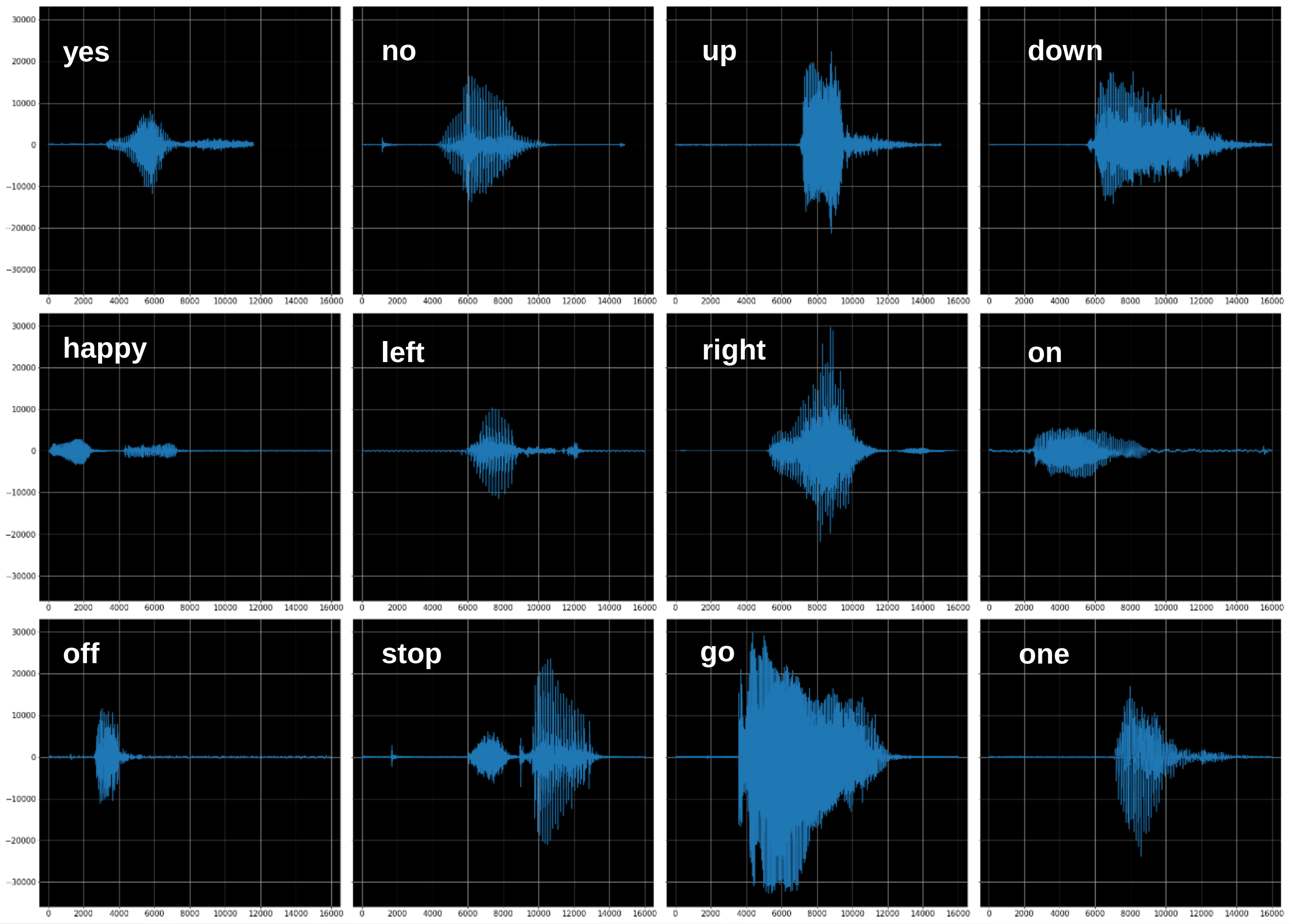
우리가 풀어야 할 문제 :
짧은 음성 명령어를 잘 인식할 수
있는 알고리즘 (모델)을 구성하기

평가방법 :
입력된 음성을 12개의 class중
하나로 잘 구별되었나 정확도 산출

명령어 클래스 :
Yes, no, up, down, left, right, on,
off, stop, go, silence, unknown

Speech recognition competition 참가기

우리가 분석해야 할 Data는?



Unique 한 명령어는 몇 개 일까요?

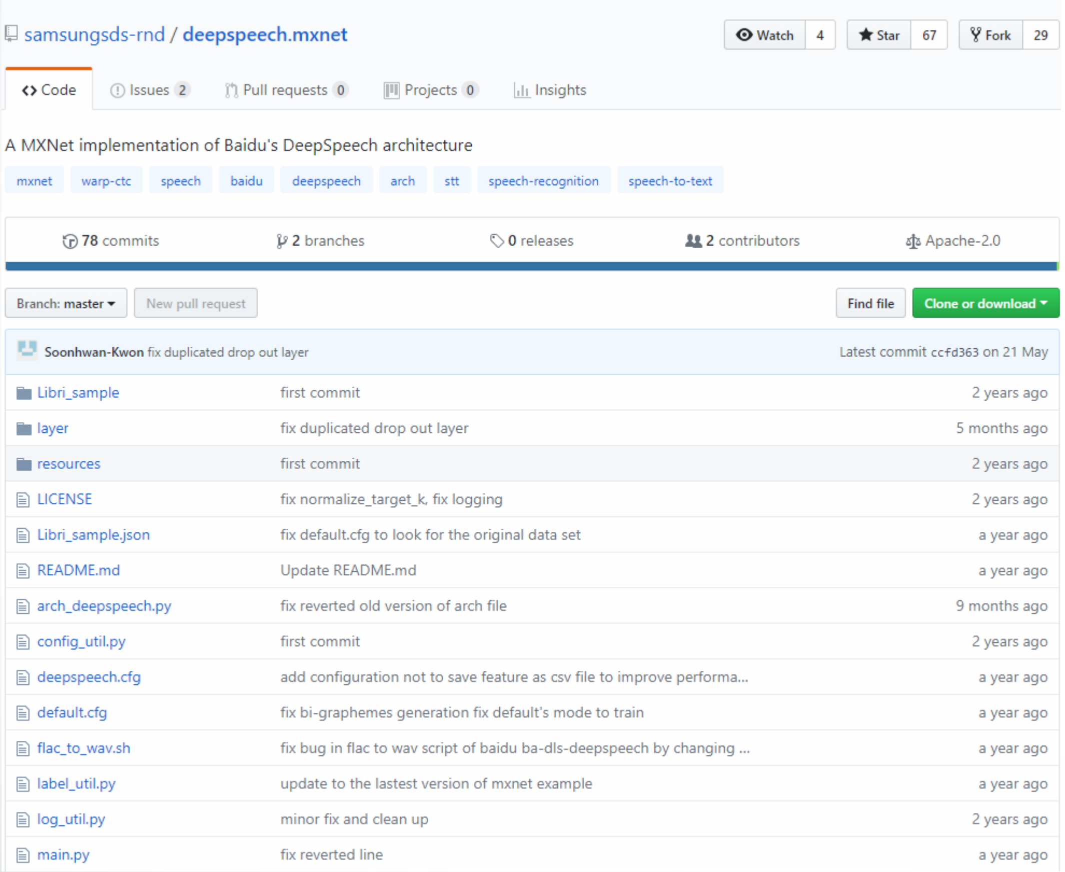
Train data : 71427 건

- **Yes, no, up, down, left, right, on, off, stop, go** (23682)
- **Unknown** (zero, one, two, three, four, five, six, seven, eight, nine, bed, bird, cat, dog, happy, house, marvin, Sheila, tree, wow) (41039)
- **Silence** (2001)
- Noise: white noise, dish wash (4705)

10가지 명령어, unknown class, silence class
Q. 노이즈는 왜 있을까?

Speech recognition model 만들기

문제에 맞게 Model을 골라서 뜯고, 쌓고, 학습 시작 !!



samsungsds-rnd / deepspeech.mxnet

Watch 4 | Star 67 | Fork 29

<> Code | Issues 2 | Pull requests 0 | Projects 0 | Insights

A MXNet implementation of Baidu's DeepSpeech architecture

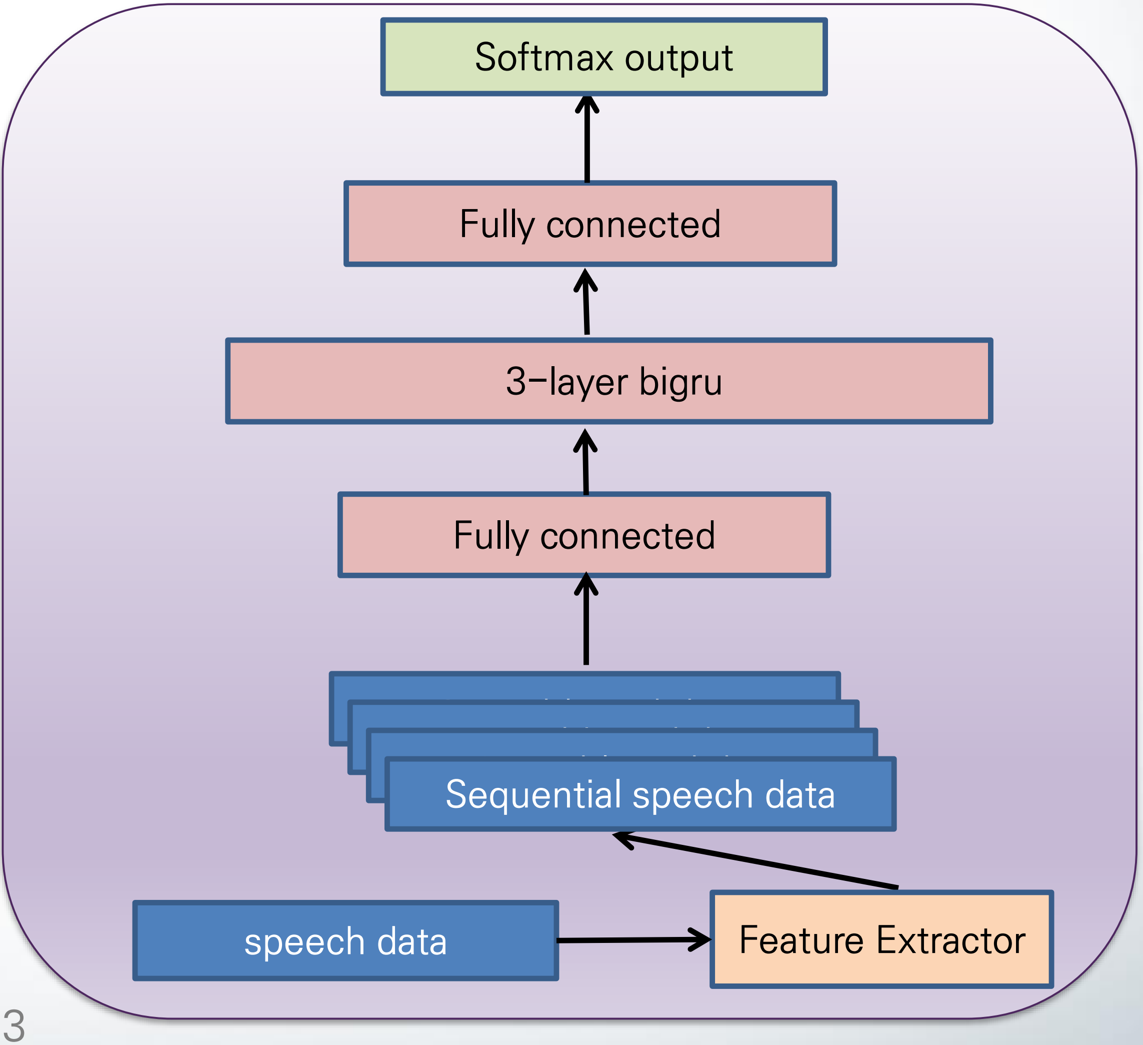
mxnet | warp-ctc | speech | baidu | deepspeech | arch | stt | speech-recognition | speech-to-text

78 commits | 2 branches | 0 releases | 2 contributors | Apache-2.0

Branch: master ▾ | New pull request | Find file | Clone or download ▾

Soonhwan-Kwon fix duplicated drop out layer — Latest commit ccfd363 on 21 May

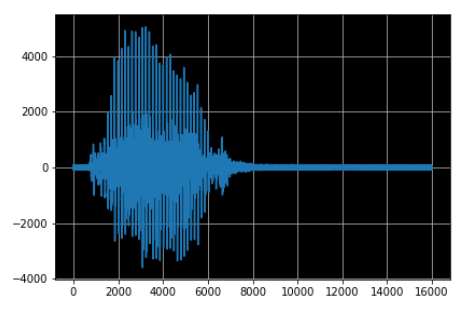
Libri_sample	first commit	2 years ago
layer	fix duplicated drop out layer	5 months ago
resources	first commit	2 years ago
LICENSE	fix normalize_target_k, fix logging	2 years ago
Libri_sample.json	fix default.cfg to look for the original data set	a year ago
README.md	Update README.md	a year ago
arch_deepspeech.py	fix reverted old version of arch file	9 months ago
config_util.py	first commit	2 years ago
deepspeech.cfg	add configuration not to save feature as csv file to improve performa…	a year ago
default.cfg	fix bi-graphemes generation fix default's mode to train	a year ago
flac_to_wav.sh	fix bug in flac to wav script of baidu ba-dls-deepspeech by changing …	a year ago
label_util.py	update to the lastest version of mxnet example	a year ago
log_util.py	minor fix and clean up	2 years ago
main.py	fix reverted line	a year ago



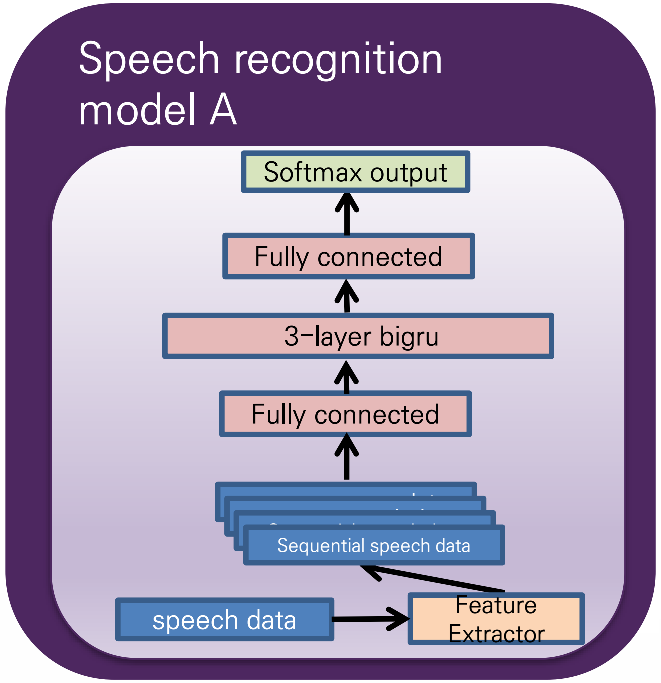
음성 인식 모델의 마지막 layer를 우리의 문제에 맞게 디자인하고
Transfer learning 을 적용해 볼까요?

Speech recognition model 결과

주어진 train data를 바탕으로 모델 학습을 진행



Test data 🔈



Speech recognition model A

Softmax output

Fully connected

3−layer bigru

Fully connected

Sequential speech data

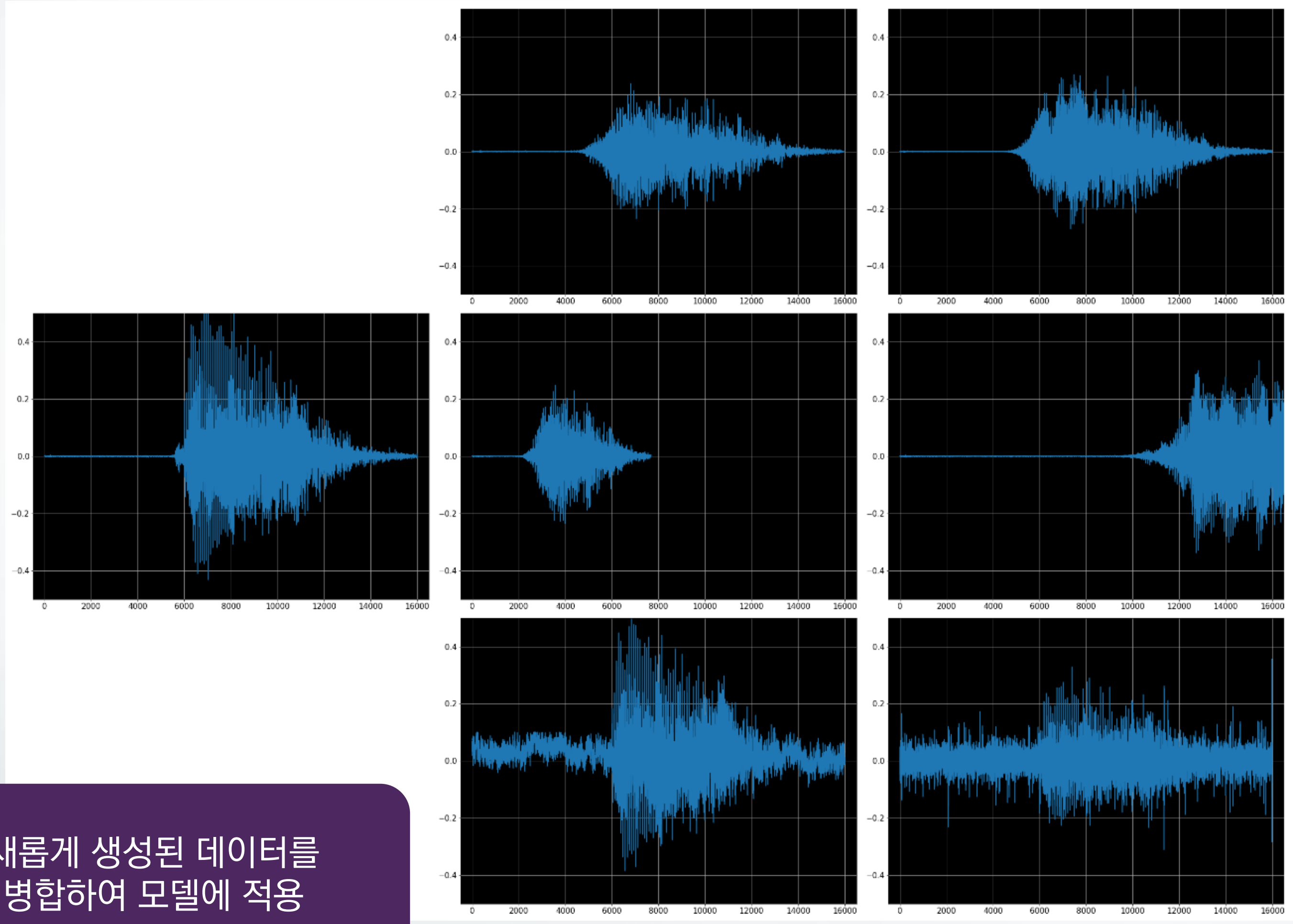
speech data

Feature Extractor

Unknown class

Speech recognition model A의 문제

모델 학습 1차 결과 Train에서 학습한 적 없었던 데이터가 Test에 다수 등장!! 해결책은??



음성 Pitch 조절

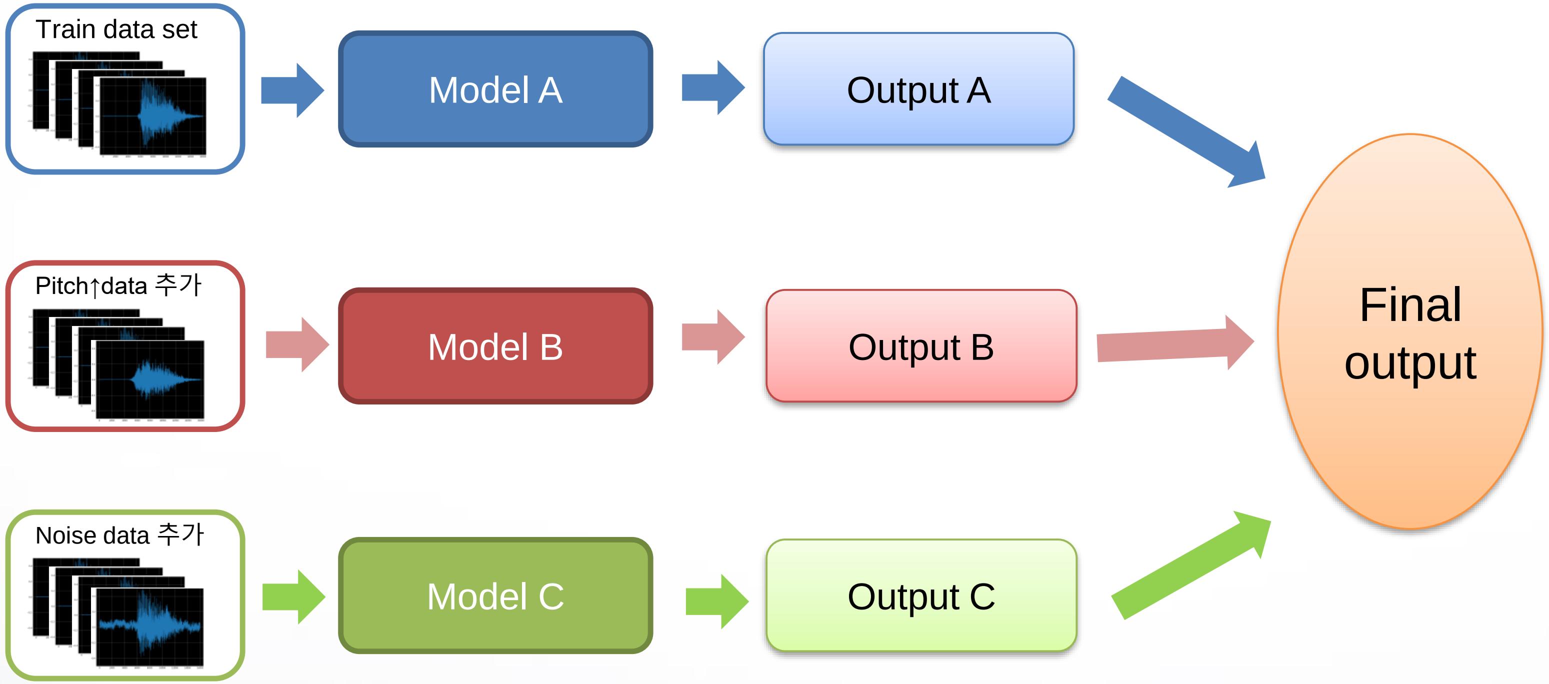
음성 속도 조절

주변 소음 입히기

새롭게 생성된 데이터를
병합하여 모델에 적용

Speech recognition models

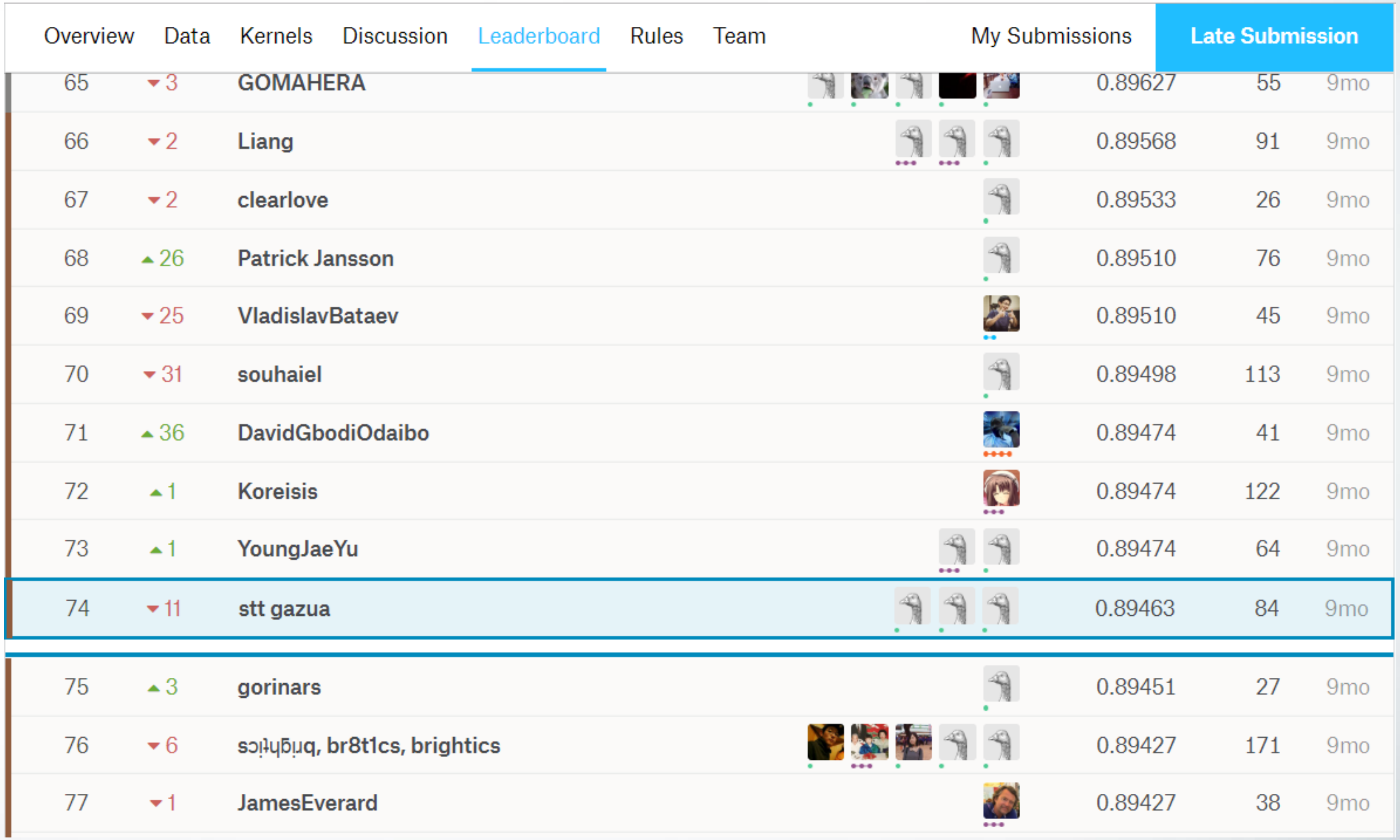
새롭게 추가되어진 데이터들 기반으로 여러 model을 만들어 ensemble을 해 보자



Tensorflow speech recognition challenge

Prediction 결과는?

정확도: 0.89463
등수 : bronze Level (top 6% over 1315 teams)



65	▼3	GOMAHERA	0.89627	55	9mo
66	▼2	Liang	0.89568	91	9mo
67	▼2	clearlove	0.89533	26	9mo
68	▲26	Patrick Jansson	0.89510	76	9mo
69	▼25	VladislavBataev	0.89510	45	9mo
70	▼31	souhaiel	0.89498	113	9mo
71	▲36	DavidGbodiOdaibo	0.89474	41	9mo
72	▲1	Koreisis	0.89474	122	9mo
73	▲1	YoungJaeYu	0.89474	64	9mo
74	▼11	stt gazua	0.89463	84	9mo
75	▲3	gorinars	0.89451	27	9mo
76	▼6	sɔıʇɥƃıɹq, br8t1cs, brightics	0.89427	171	9mo
77	▼1	JamesEverard	0.89427	38	9mo

Competition 참가 후기

Speech recognition challenge 도전을 통하여 배운 점은

유사한 소리(on, off, up) 에 대한 분류도를 높이기 위한 speech data augmentation

- 음성의 speed, pitch 조절
- Noise 추가 : 화이트 노이즈, 외부 소음, 물방울 소리, random distributed noise

Model ensemble 시도

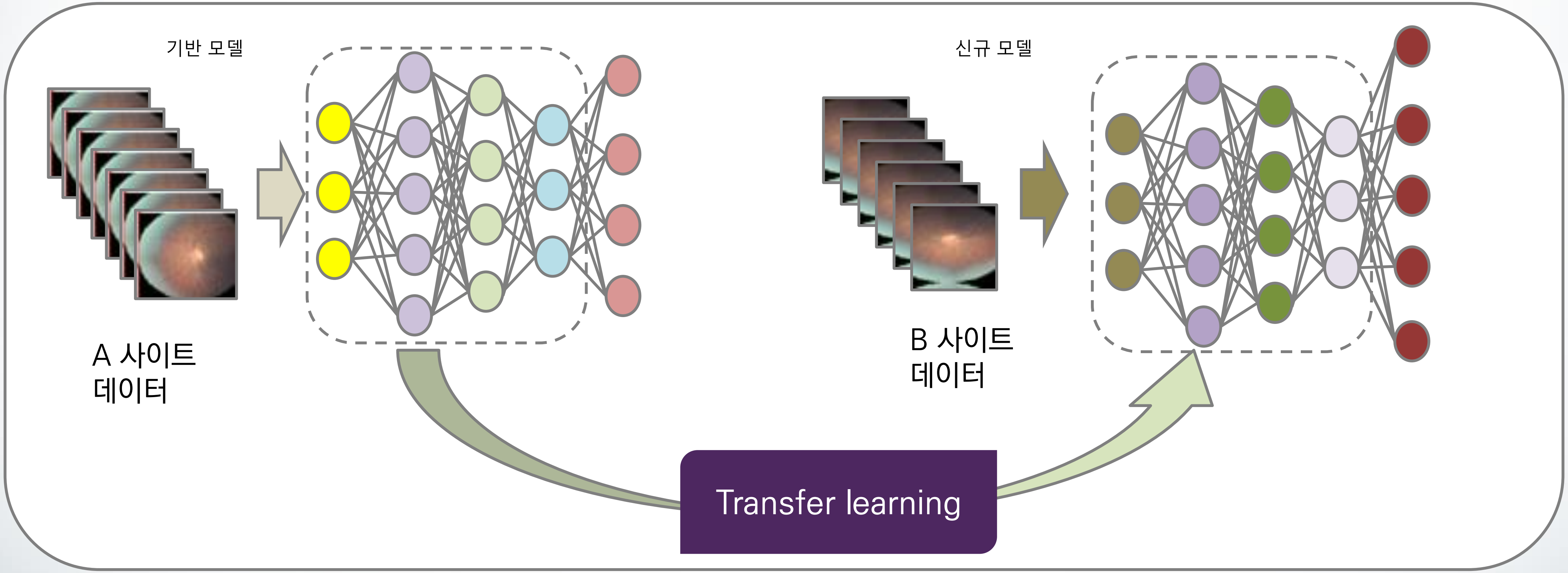
- Transfer learning 기반 여러 모델 개발. 각 모델 별 최적의 classification 결과 도출을 시도
- 10개의 모델에 대한 ensemble 도전 : random vote, average value, weighted sum etc.

업무 종료 후에 캐글러로 변신

시간 제약 상 더 많은 모델을 디자인하여 학습해 볼 수 있었었으면 좋았을 것 같다는 아쉬움이 남지만
Top 6% 로 종료

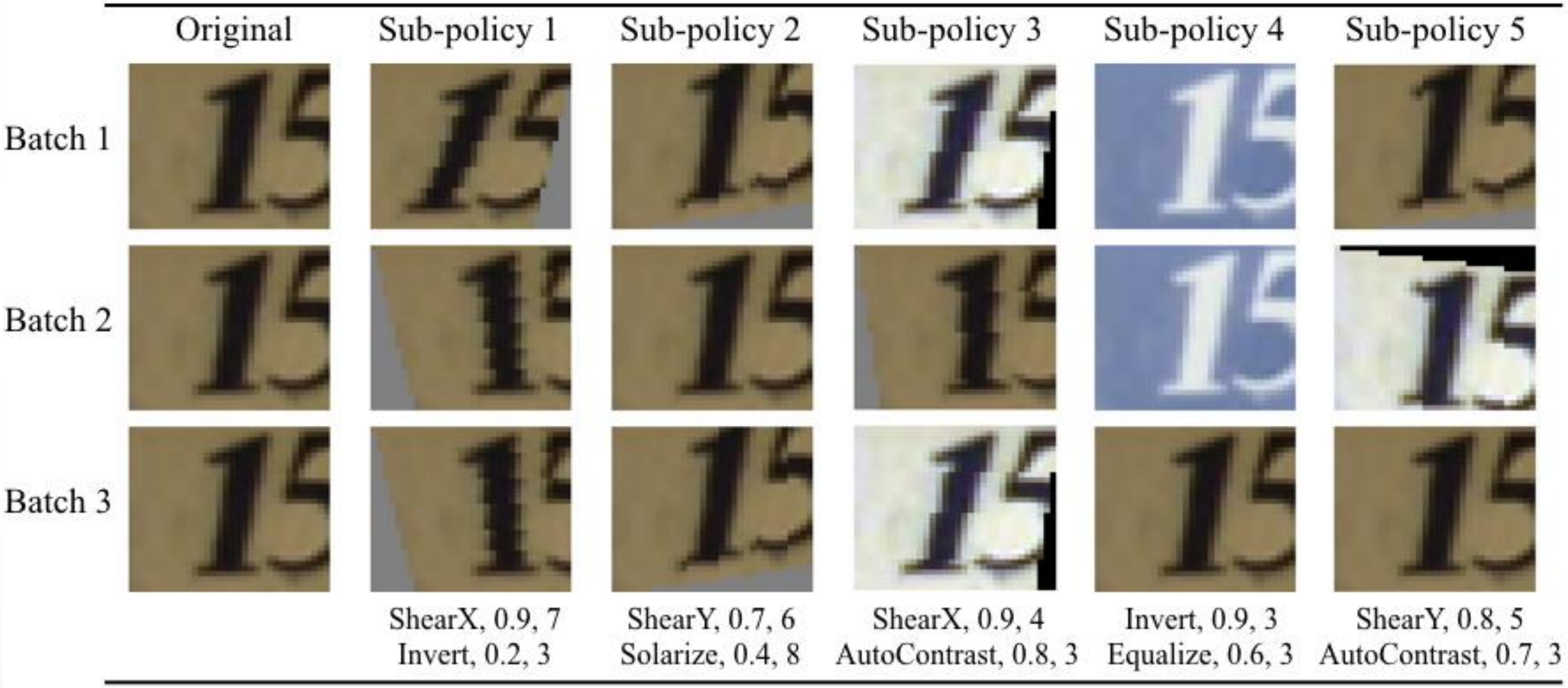
향후 계획

Kaggle competition 참가를 통해 얻은 Transfer learning 기술 개발



향후 계획

다양한 데이터 대상의 augmentation 기술 개발



Ekin Cubuk. Et al, AutoAugment: Learning Augmentation Policies from Data, google

Q & A

Thank you