AI Research Scientist

Master of Mathematics, Kwangwoon University
ydy89899@gmail.com
Doyeon Yoon

Projects

NLP projects

1. Text Classification
2. KLUE : Relation Extraction
3. KLUE : Dialogue State Tracking(DST)

Another projects

4. Sensor data Anomaly detection
5. Medical / CT-images Segmentation
6. Deep Knowledge Tracing(DKT)
7. Image Classification

ETC.

Projects

NLP projects

1. Text Classification
2. KLUE : Relation Extraction
3. KLUE : Dialogue State Tracking(DST)

Another projects

4. Sensor data Anomaly detection
5. Medical / CT-images Segmentation
6. Deep Knowledge Tracing(DKT)
7. Image Classification

ETC.

1. Senticle : News data-based stock price prediction

POSTECH PIRL AI/Big Data Advanced Course / 18.09 ~ 18.10 / Team

Overview

- √ Implementation of model & app implementation to predict next-day stock price through news data of a specific company(Binary Text Classification)

Dataset

- √ Crawl 5 years news data

Model

- √ **Architecture** : 1DCNN
- √ **Tokenizer** : Soynlp
- √ **Embedding** : RandomVector, ~~Fasttext, Word2Vec~~
- √ **Skills** : Tensorflow V1, Android, LIME, NLP

Role

- √ Preprocessing(Labeling, Tokenizing)
- √ Modeling

Result

- √ Inferene Acc : 73%
- √ Backtesting results were not good.
- √ Run the app

Link

- √ **Github :** *https://github.com/ydy8989/senticle-proj*
- √ **App play :** *https://youtu.be/syQfQGFAAZ0*

2. KLUE : Relation Extraction Competitions

Boostamp AI Tech, NAVER Connect Foundation / 21.04.12 ~ 21.04.23 (2W) / Solo

Overview

- √ A task to classify the relationship between two entities in sentences

Dataset

- √ **Input** : sentence, entity1, entity2

	sentence	entity_01	entity_02	label
0	영국에서 사용되는 스포츠 유틸리티 자동차 브랜드로는 랜드로버(Land Rover)...	랜드로버	자동차	17
1	선거에서 민주당은 해산 전 의석인 230석에 한참 못 미치는 57석(지역구 27석,...	민주당	27석	0
2	유럽 축구 연맹(UEFA) 집행위원회는 2014년 1월 24일에 열린 회의를 통해 ...	유럽 축구 연맹	UEFA	6
3	용병 공격수 차디의 부진과 시즌 초 활약한 강수일의 침체, 시즌 중반에 영입한 세르...	강수일	공격수	2
4	탐험행 왕은 1237년에서 1247년 사이 수코타이의 왕 파준 씨 인트라팃과 쓰엉 ...	탐험행	파준 씨 인트라팃	8

- √ **Output** : One of 42 relation classes

{'관계_없음': 0, '인물:배우자': 1, '인물:직업/직함': 2, '단체:모회사': 3, '인물:소속단체': 4, '인물:동료': 5, '단체:별칭': 6, '인물:출신성분/국적': 7, '인물:부모님': 8, '단체:본사_국가': 9, '단체:구성원': 10, '인물:기타_친족': 11, '단체:창립자': 12, '단체:주주': 13, '인물:사망_일시': 14, '단체:상위_단체': 15, '단체:본사_주(도)': 16, '단체:제작': 17, '인물:사망_원인': 18, '인물:출생_도시': 19, '단체:본사_도시': 20, '인물:자녀': 21, '인물:제작': 22, '단체:하위_단체': 23, '인물:별칭': 24, '인물:형제/자매/남매': 25, '인물:출생_국가': 26, '인물:출생_일시': 27, '단체:구성원_수': 28, '단체:자회사': 29, '인물:거주_주(도)': 30, '단체:해산일': 31, '인물:거주_도시': 32, '단체:창립일': 33, '인물:종교': 34, '인물:거주_국가': 35, '인물:용의자': 36, '인물:사망_도시': 37, '단체:정치/종교성향': 38, '인물:학교': 39, '인물:사망_국가': 40, '인물:나이': 41}

- √ **Datasets :** *https://klue-benchmark.com/tasks/70/data/description*

Model

- √ **MLM** : XLM-RoBERTa, ~~Koelectra, Bert~~
- √ **Preprocessing** : EDA(Easy Data Augmentation), Back Translation, TEM(Typed Entity Marker)
- √ **Skills** : Pytorch, Huggingface, Tensorboard

Link

- √ **Github :** *https://github.com/bcaitech1/p2-klue-ydy8989*
- √ **Notion :** *https://www.notion.so/whydo/KLUE-Relation-Extraction-4708eefe61f849ac8771806898e97333*

Rank

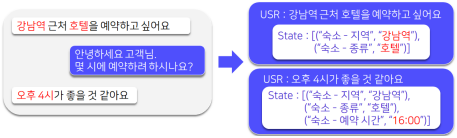
- √ **ACC** : 79.9% | **Public LB** : 46th Place (No Private LB)

3. KLUE : Dialogue State Tracking(DST) Competitions

Boostamp AI Tech, NAVER Connect Foundation / 21.04.26 ~ 21.05.21(4W) / Team

Overview

- √ A task that infers the pair of SLOT and VALUE to be predicted in the Dialogue System every turn.



Dataset overview

- √ JSON Format, The state to be predicted consists of a pair of "Domain-Slot-Value"
- √ **Datasets :** *https://klue-benchmark.com/tasks/73/data/description*

Dataset I/O

- √ **Input** : 1 turn of user and system utterance within Dialogue
- √ **Output** : State pairs of user utterances classified as "Domain-Slot-Value"

Domain-Slot Value

{ "숙소-가격대" : ["저렴", "적당", "비싼", "none", "dontcare"],
"숙소-지역" : ["동쪽", "서쪽", "남쪽", "북쪽", "none", "dontcare"],
"숙소-주차가능" : ["yes", "no", "none", "dontcare"],
...}

- Domain - 5 Classes
- Slot - 45 Classes
- Value - It changes according to the data.

Metric

- √ Joint Goal Accuracy » Slot Accuracy » Slot F1 Score

3. KLUE : Dialogue State Tracking(DST) Competitions

Boostamp AI Tech, NAVER Connect Foundation / 21.04.26 ~ 21.05.21(4W) / Team

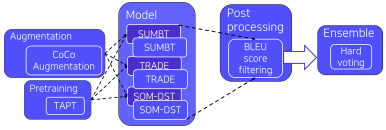
Model

- √ **Architecture :**
  - Ontology based : SUMBT
  - Open Vocab based : TRADE, SOM-DST
- √ **Preprocessing :**
  - CoCo Augmentation
  - TAPT
- √ **Post-Preprocess :** BLEU score filtering(Idea)
- √ **Ensemble :** HardVoting
- √ **Skills :** Pytorch, Tensorboard, Huggingface

Role

- √ Build SOM-DST and fine tuning
- √ SOM-DST + CoCo / TAPT

Pipeline



Link

- √ **Team Github :** *https://github.com/bcaitech1/p3-dst-teamed-st*
- √ **Notion :** *https://www.notion.so/whydo/Dialogue-State-Tracking-81883b2d7c0246c7b2d3ea2cb766ba62*

Rank

- √ **JGA** : 0.8344 | **Public LB** : 1st Place
- √ **JGA** : 0.7355 | **Private LB** : 1st Place

Projects

NLP projects

1. Text Classification
2. KLUE : Relation Extraction
3. KLUE : Dialogue State Tracking(DST)

Another projects

4. Sensor data Anomaly detection
5. Medical / CT-images Segmentation
6. Deep Knowledge Tracing(DKT)
7. Image Classification

ETC.

4. Samsung Semiconductor Smart Interlock

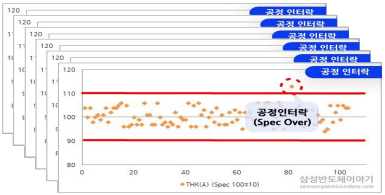
Hbee Co. AI team, Machine learning engineer / 19.09 ~ 20.03 (6M)

Overview

- √ A project to detect anomalies in time-series sensor data in the semiconductor process and classify them into 7 classes in detail

Dataset

- √ Sensor logs for the past 30 days from the time of the anomaly prognosis
- √ **INPUT** : Timeseries raw data 2000
- √ **OUTPUT** : True(3 Classes) / False(4 Classes)



Pipeline

- √ Two-way ensemble(image data + raw data)
  - ▶ Plotting image : SGAN
  - ▶ Raw data : Stacked Auto-encoder + Linear regression
- √ ensemble model ⇔ Tensorflow serving ⇔ Flask

Model

- √ **Architecture** : SGAN, Auto-Encoder, Linear Regression
- √ **Metric importance** : Precision » f1-score
- √ **Skills** : Keras, Flask, Tensorflow Serving

Result

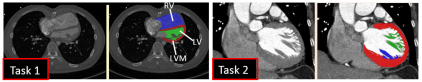
- √ **Precision** : 99% / **F1-Score** : 83%
- √ Applied to all Samsung Electronics semiconductor process lines

5. Cardiac CT images Segmentation Competition

POSTECH AIRL Intern / 18.11 ~ 18.01 (2.5M) / Team

Overview

- √ Cardiac Segmentation task competition



Dataset

- √ 3D Images(.mha format)
- √ Sample 1 case, Train 100 cases, Test 100 cases(*Train data, Test data : Blind(in server)*)

Model

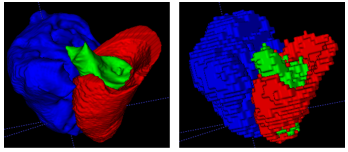
- √ **Architecture** : 3D-UNet
- √ **Preprocessing** : Patching(Resize, Crop, Augmentation), Resampling(Voxel spacing), Intensity Windowing
- √ **Skills** : Tensorflow, Keras, SimpleITK, Docker, Segmentation

Role

- √ Preprocessing(Resampling, Intensity windowing)
- √ Fine tuning

Result

- √ **Dice Coefficient** : 73%(winner:79%)



Ground truth Prediction

Link

- √ **Github :** *https://github.com/ydy8989/Cardiac_Segmentation*

6. Deep Knowledge Tracing(DKT)

Boostamp AI Tech, NAVER Connect Foundation / 21.05.24 ~ 21.06.15(4W) / Team

Overview

- √ A task that tracks the personalized knowledge state through the user's (student) problem-solving information

Dataset

- √ **Input** : Problem solving data for 7442 users

	userID	assessmentItemID	testId	answerCode	Timestamp	KnowledgeTag
0	0	A060001001	A060000001	1	2020-03-24 00:17:11	7224
1	0	A060001002	A060000001	1	2020-03-24 00:17:14	7225
2	0	A060001003	A060000001	1	2020-03-24 00:17:22	7225
...	...	...	...	...	...	...
2266581	7441	A030071005	A030000071	0	2020-06-05 06:50:21	438
2266582	7441	A040165001	A040000165	1	2020-08-21 01:06:39	8836

- √ **Output** : Answer to the problem

Model

- √ **Architecture :** LGBM, SAINT, LastNQuery, ~~GKT~~
- √ **Preprocessing :** Feature Engineering
- √ **Skills :** Pytorch, Tensorboard, Transformer, Bert
- √ **Metric :** AUROC

Link

- √ **Team Github :** *https://github.com/bcaitech1/p4-dkt-decayt*
- √ **Team Notion :** *https://www.notion.so/Home-b263b1f24c3147ac9f8f2544178d66f6*

Rank

- √ **AUROC** : 0.842 | **Public LB** : 2nd Place
- √ **AUROC** : 0.845 | **Private LB** : 4th Place

7. Mask Image Classification

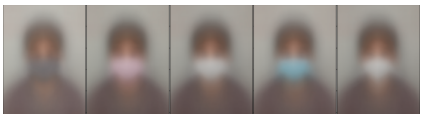
Boostamp AI Tech, NAVER Connect Foundation / 21.03.29∼21.04.08(2W) / Solo

Overview

- √ Image classification task according to gender / age / mask wearing types

Dataset

- √ Images for 2700 people(7 images per 1 person → 5 wears, 1 no wear, 1 half wear)
- √ **Input** : 384x512 sized masked face image



- √ **Output** : 18 Classes
  - No wear / wear / half wear
  - Gender
  - Ages(30 or less / 30-60 / 60 or more)

Model

- √ **Architecture :** EfficientNet b4, ~~ResNet~~
- √ **Preprocessing :** Augmentation, Label filtering
- √ **Skills :** Pytorch, Tensorboard, Stratified kfold
- √ **Metric :** F1-score

Link

- √ **Github :** *https://github.com/bcaitech1/p1-img-ydy8989*

Rank

- √ **F1-score** : 0.6800 | **Public LB** : 158th Place
- √ **F1-score** : 0.6738 | **Private LB** : 152th Place

Projects

NLP projects

1. Text Classification
2. KLUE : Relation Extraction
3. KLUE : Dialogue State Tracking(DST)

Another projects

4. Sensor data Anomaly detection
5. Medical / CT-images Segmentation
6. Deep Knowledge Tracing(DKT)
7. Image Classification

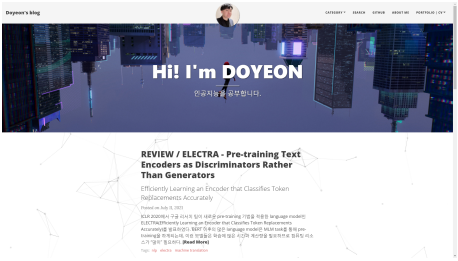
ETC.

ETC.

BLOG

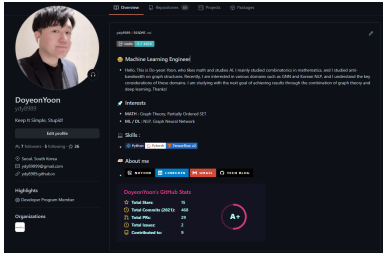
- ✓ We are organizing thesis reviews and learning contents on our blog. Details of the above project can be found through the link below.

  https://ydy8989.github.io/



Github

- ✓ *https://github.com/ydy8989/*



Notion

- ✓ *https://www.notion.so/whydo/Doyeon-Yoon-05603016086c4b3ca954cf2b6c64e46f*

Thank You!