한국투자증권

참여 프로젝트 상세 기술서

지원자: 최영제

목 차

인적/경력 사항

최영제 1995년 (27세) | 남 | 구직 준비중
c73801477@gmail.com
010-7380-1477
010-7380-1477
(03726) 서울 서대문구 연희로8길

[대외 활동]

기간	구분	기관	내용
2018.01~ 2019.01	외부 동아리	빅데이터 연합 동아리 ToBig's	객체 탐지 기반 한국어 지문자 번역기 / 강화학습 기반 재난 구조 Agent 개발
2017.03~ 2019.12	교내 학회	빅데이터 분석 학회 D&A	부학회장 / KOSPI 200 선물 예측 / 음원 → 악보 변환 프로젝트
2019.07~ 2019.08	인턴	CLIO	마케팅 / 텍스트 마이닝 / 데이터 분석
2021.06~ 2021.11	프로그램	타사 AI Fellowship	시계열 이상 탐지 알고리즘 개발

[자격증/수상내역]

취득일/수상일	구분	자격/대회명	발행처/기관	합격/수상
2018.07	자격증	데이터분석준전문가	한국데이터베이스진흥원	합격
2018.08	공모전	제 7회 빅데이터 분석 경진대회	UNIST	우수상
2018.12	공모전	2018 빅데이터 해커톤	우정사업본부	우수상

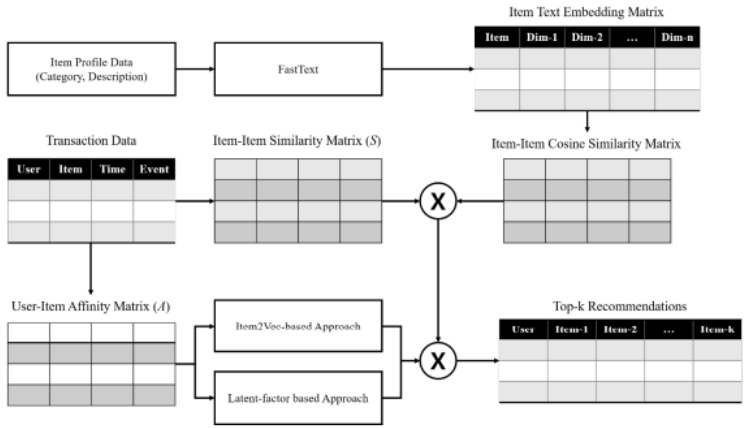
[학력]

재학기간	구분	학교명(소재지)	전공	학점
2020.03~ 현재	졸업 예정(석사)	연세대학교(서울)	산업공학	3.91/4.3
2014.03~ 2020.02	졸업	국민대학교(서울)	빅데이터 경영통계전공	3.94/4.5
2011.03~ 2014.02	졸업	광양백운고등학교	인문계	-

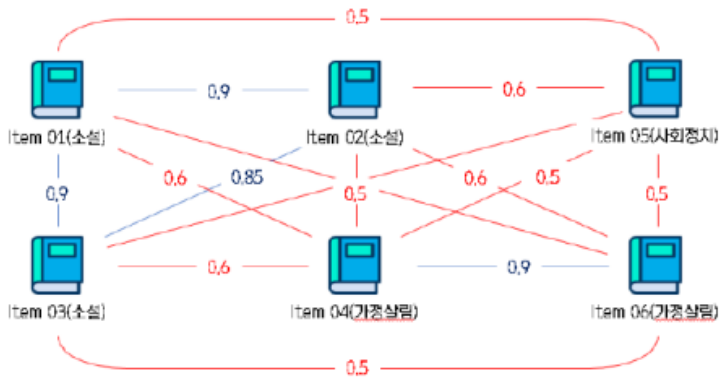
연구 논문

1) 트랜잭션 기반 추천 시스템에서 워드 임베딩을 통한 도메인 지식 반영

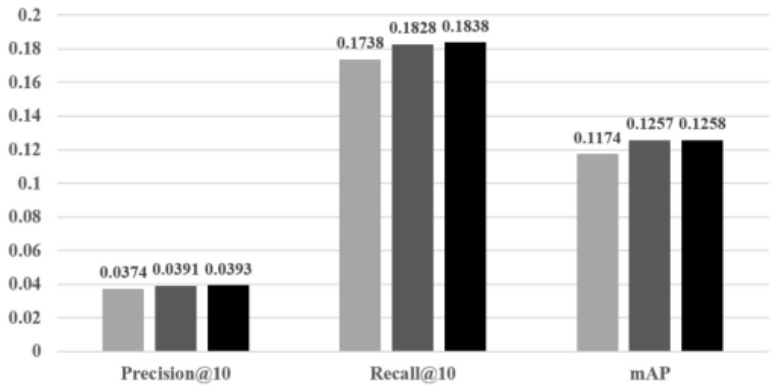
- 추천 시스템에서 **암시적 피드백**(implicit feedback)이란 구매 이력, 방문 빈도 등과 같이 0과 1로 표현되는 정보이며 해당 피드백의 경우 사용자의 선호 강도가 직접적으로 나타있지않아 상호작용이 이루어진 품목들은 모두 **동일한 선호를 가정**한다는 한계가 존재함
- 본 논문의 도메인인 도서 추천 시스템(yes24)은 EDA 결과 소비자들에게 **특정 장르의 도서만 구매하는 소비 편향**이 존재함을 발견하였으나 암시적 피드백은 이를 반영하지 못함
- 이를 해결하기 위해 사용자가 구매한 도서들의 주제, 줄거리를 fast-text를 이용하여 latent vector를 도출하고 이들 간의 **코사인 유사도(cosine similarity)를 구매 빈도의 가중치로 적용**하는 방법을 제안하였으며 이를 통해 도서 구매자의 **주소비 장르에는 높은 가중치**를, 그렇지 않은 장르에는 **낮은 가중치를 부여**함
- Simple algorithm for recommendation system (SAR) 알고리즘을 변형하였으며 기존 SAR 보다 precision, recall, mAP 지표에서 성능 개선을 보임
- 논문 1저자 및 지식경영연구(KCI) 등재



<그림 4> 제안 시스템의 구조



<그림 5> 도서간 유사도 계산 예제



<그림 9> User-Item Affinity Matrix 개선 결과

연구 논문

2) Dimension Reduction Using a DBSCAN Ensemble for High-Dimensional Semicon-ductor Manufacturing Datasets

- 본 논문은 학습에 사용할 sample의 수는 적고, 활용할 변수는 많은 Large P, Small N 상황에 적합한 feature selection & extraction 방법이며 제안 방법은 **clustering을 활용하는 feature engineering 방법**으로 기존의 feature selection, extraction 방법과는 궤를 달리함
- 제안 방법론은 입력 데이터를 전치(transpose) 후 clustering 알고리즘을 적용하는 방법을 택하는데, 전치를 수행하면 기존의 feature들은 행으로, sample들은 열로 변환되기에 **유사한 속성을 지닌 feature들끼리 군집화가 수행**됨
- 이후 구축된 cluster의 중심 벡터(centroid vector) 혹은 중심 벡터와 가장 가까운 변수를 선택하는 방법으로 extraction 및 selection을 진행함
- 총 4개의 제조 데이터에 실험 결과 RMSE 기준 두 개의 데이터에서는 가장 좋은 성능을, 나머지 데이터에서는 준수한 성능을 보임
- 논문 2저자 참여 및 투고 전

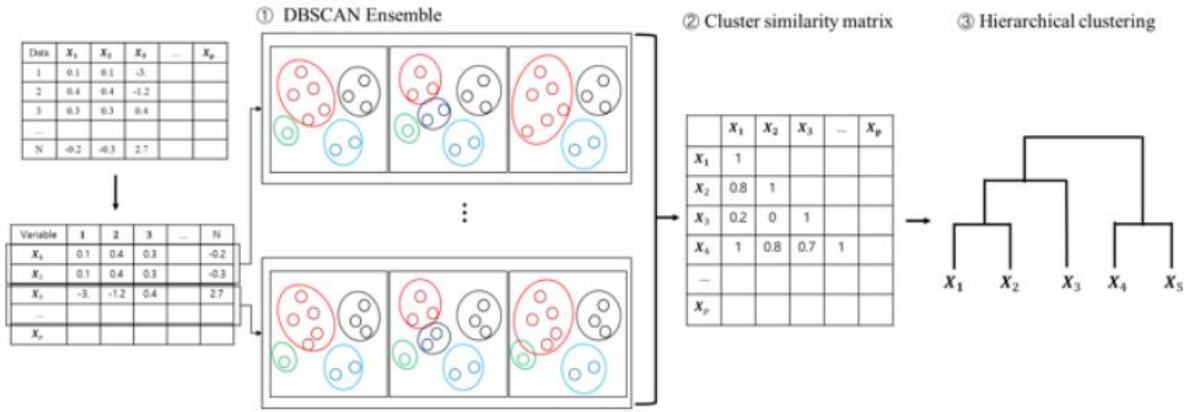


Fig. 1. DBSCAN ensemble based feature selection procedure

	Data 1	Data2	Data3	Data4
All feature	87.86	4416.05	17404.56	45.11
SR	89.96	8914.22	16755.28	31.32
RFE	55.76	**192.87**	4458.62	25.11
GA	60.11	624.85	16976.88	22.42
PCA	52.23	238.06	4835.53	12.01
AE	55.27	232.59	**4188.28**	11.97
KC(FS)	54.28	268.81	6101.94	13.87
KC(FE)	53.30	257.05	5295.96	12.14
Proposed Method(FS)	**49.65**	229.10	4451.42	**11.95**
Proposed Method(FE)	50.69	230.34	4453.17	12.07

연구 논문

3) SAFE: Unsupervised image feature learning by self-attention based feature extraction networks

- 본 논문은 비지도 이미지 특징 추출에서 **관행적으로 사용되는** convolutional autoencoder **(CAE)**, convolutional variational autoencoder **(CVAE)의 단점을 파악**하고 이를 **개선한 모델 구조를 제안**함
- 기존의 encoder-decoder 구조의 단점은 encoder로부터 추출된 특징이 불명확하더라도 decoder의 성능이 우수할 경우 입력 이미지의 복원이 완벽하게 가능함. 즉, **autoencoder의 복원 성능과 encoder의 특징 추출 능력은 반드시 비례하지 않음**
- 이를 해결할 수 있는 self-attention based feature extraction networks (SAFE)를 제안하였으며 제안 구조는 아래 그림과 같이 3개의 encoder가 존재, 이들의 외적(outer-product)과 attention 가중치 반영을 통해 입력 이미지를 복원함. 이에 따라 복원 시 **decoder의 영향력이 매우 적어지기에 복원 성능과 encoder의 특징 추출 능력을 비례**시킬 수 있어 **더 좋은 feature를 얻을 수 있음**
- 논문 1저자(졸업 논문) 및 투고 전

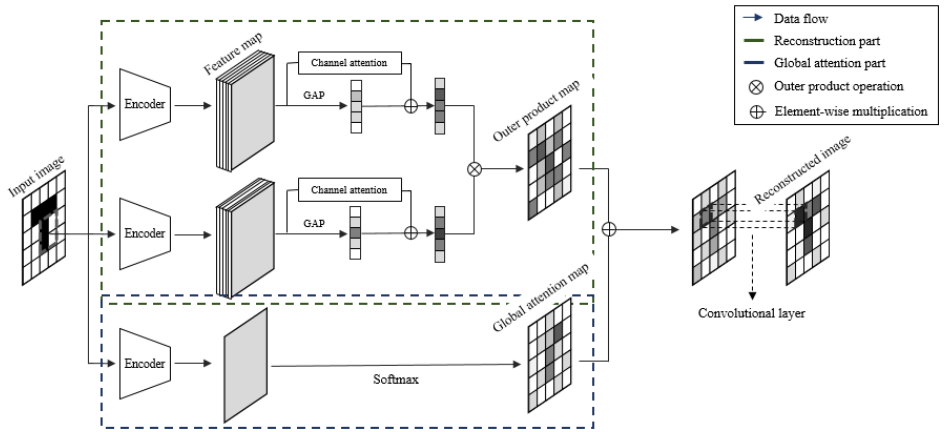


Fig.1. The proposed architecture for the self-attention based feature extraction networks (SAFE). GAP means global average pooling.

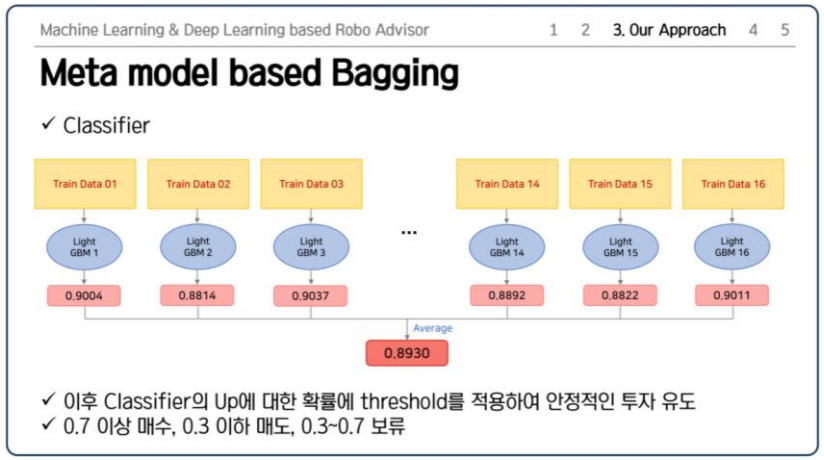
Fig. 8. Visualization of extracted features in the MNIST dataset. The left is CAE's feature space and the right is SAFE's.

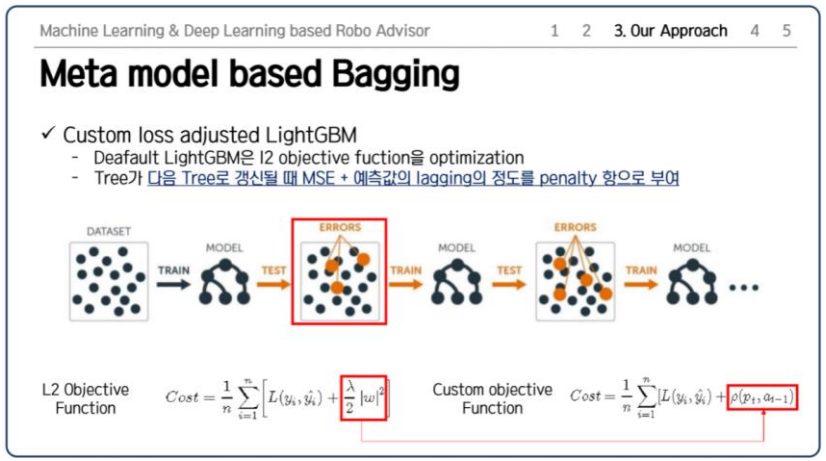
프로젝트 – 시계열 예측

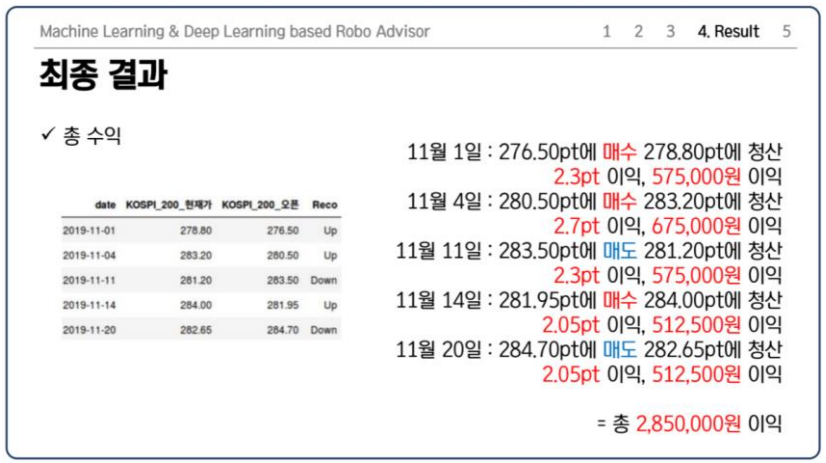
KOSPI 200 선물 예측

발표 자료 링크:
https://ds.kookmin.ac.kr/community/?board_name=Community&order_by=fn_pid&order_type=desc&board_page=2&vid=1

- 해당 프로젝트는 2019년 교내 빅데이터 학회 D&A의 제 2회 컨퍼런스로 진행하였으며 2010.01-2019.10까지의 학습 데이터를 활용해 2019.11월 **한달 간의 KOSPI 200 선물 가격을 예측**하고자 하였음
- 활용한 데이터는 investing.com에서 수집할 수 있는 지수, 환율, 주식, 원자재, 채권 등 시계열 데이터와 네이버 증권에서 수집한 증시 뉴스, 투자 전략 report 등 텍스트 데이터를 함께 사용하였음
- 이후 TA-LIB을 통한 기술적 지표 파생변수 추가, Fast-text 이용 텍스트 임베딩 변수 추가, 시계열 예측 시 발생하는 time-lagging 해결 방안 탐색, feature & sample bagging을 통한 ensemble 예측 등을 통해서 최종 예측 모형을 구축함(세부 내용은 우측 상단의 발표 자료를 참조)







date	KOSPI_200_현재가	KOSPI_200_오픈	Reco
2019-11-01	278.80	276.50	Up
2019-11-04	283.20	280.50	Up
2019-11-11	281.20	283.50	Down
2019-11-14	284.00	281.95	Up
2019-11-20	282.65	284.70	Down

프로젝트 – 시계열 예측

2018 우정사업본부 빅데이터 해커톤 – 우수상

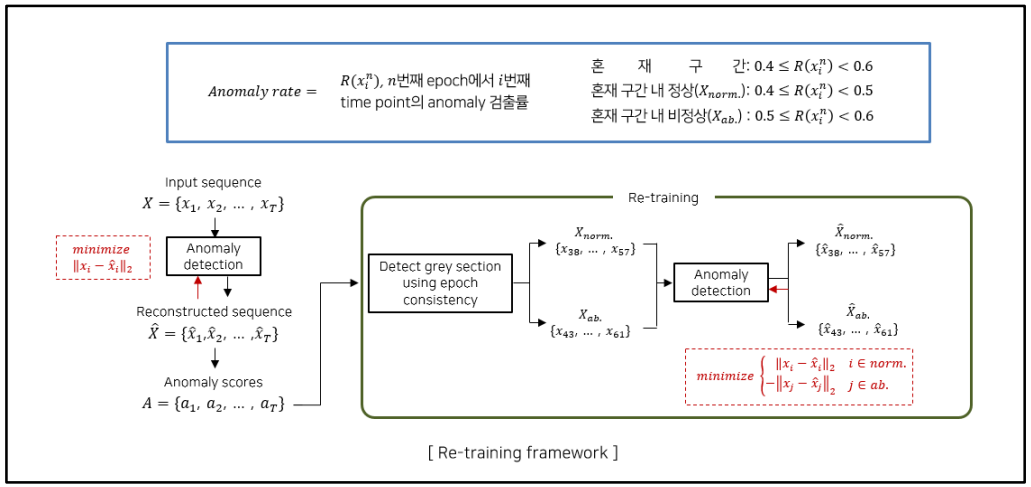
- 사용 툴 : R, Python
- 프로젝트 목적 : Regression, 2017년 하반기, 2018년 하반기 연휴기간(연휴일 전,후 3일) 택배 물량 예측
- 프로젝트 기간 : 2018.12.20 ~ 2018.12.21

- [데이터 수집]: 우정사업본부에서 제공한 데이터는 17,18년도 상반기 구별로 집계된 시계열 운송 데이터임. 이에 구글맵 크롤링, 구별 인구 census 정보, 주변 시설(아파트, 병원, 산업단지 유무 등) 등 공간정보를 포함한 내용과 **네이버 검색어 트렌드**, Twiiter API를 활용한 **SNS 정보**, 기상특보 등의 **데이터를 크롤링하여 자체 수집 후 분석을 진행**.
- [EDA, 텍스트 마이닝]: 해커톤에서 수행해야하는 과업은 **공휴일 인근의 택배량을 예측**하는 것. 주어진 데이터를 EDA를 통해 파악한 결과 공휴일 당일을 제외한 주변 3일간 택배량은 평소와 다름을 확인. 또한 **30대~40대 여성 거주 인구**가 높을 수록 택배량이 많아지는 양의 상관관계를 도출. 내부 데이터 외, 네이버 검색어 키워드와 함께 EDA를 한 결과 택배 배송량과 "택배", "택배 지연", "이사", "군입대" 등 **특정 키워드와 높은 상관관계**를 갖는 것을 발견. 이에 택배량을 간접적으로 유추할 수 있는 대리지표로 선정하고 이를 독립변수로 선정.
- [택배 배송량 예측 모델링]: 앞서 언급한 데이터 및 독립 변수들을 이용하여 예측을 진행. 가장 흥미로웠던 변수는 **"우체국 택배", "우체국 택배 배송조회" 키워드의 검색량 변수**였으며 이는 **택배 배송량과 놀라울 정도로 일치**하였음. SNS 변수를 추가했을 때 RMSE가 2407.33에서 1940.92으로 큰 폭의 성능 향상을 보여주었음. 예측 모델로는 **XGBoost 모델 기반의 Bagging 알고리즘**을 사용하였으며 이는 Bagging 알고리즘의 기본 estimator를 의사결정나무 대신에 XGBoost로 삼도록 하는 모델임. Bagging 앙상블을 적용하기 전과 비교하여 약 80~100 정도의 RMSE 성능 향상 효과가 있었음.
- 키워드: R, Python, Regression, Web Crawling, textmining, XGBoost based Bagging

프로젝트 – 시계열 이상탐지

타사 AI Fellowship 3기 – Smart Factory 서비스를 위한 진동/압력/온도 센서의 Anomaly Detection 개발

- 해당 프로젝트는 진동 도메인의 시계열 센서 데이터가 들어올 때 **조기 이상 탐지 알고리즘 구축을 목표**로 함
- 이에 딥러닝 기반 최신 시계열 이상 탐지 알고리즘 모델 research 진행, 데이터 전처리 및 모델 구조 개선, 학습 프레임 워크 변화 등을 통한 성능 개선을 수행 중에 있음
- 오토 인코더 기반의 DAGMM, USAD 모델과 적대적 신경망(Generative adversarial networks, GAN) 기반의 GANomaly를 기본 네트워크로 설정 후 **제안 학습 프레임 워크**로 상기 모델들을 학습 시 **우측 하단의 표와 같은 성능 향상**이 있었음
- 현재는 모든 데이터, 알고리즘에서 동일한 성능 향상이 있도록 **일반화 성능 향상을 목표로 개발을 수행 중**



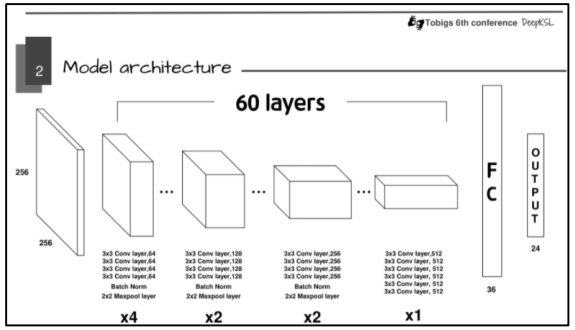
[Re-training framework]

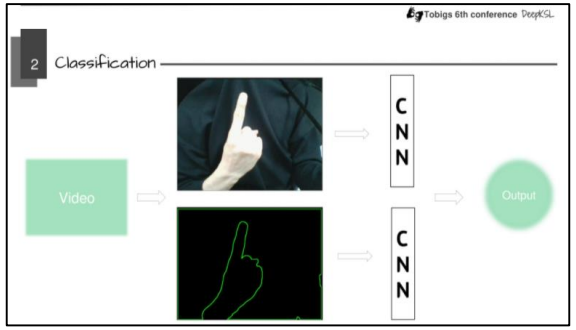
	Measure	DAGMM '18		GANomaly '18		USAD '20	
		Baseline	**IDEA**	Baseline	**IDEA**	Baseline	**IDEA**
Bearing 1	Time point (826)	820	825	821	824	820	820
	False Alarm	26	2	11	8	17	14
	Score	32	**3**	16	**10**	23	**20**
Bearing 2	Time point (1274)	1171	1150	1330	1341	1101	1279
	False Alarm	276	271	202	207	80	99
	Score	**379**	405	**258**	274	243	**204**

기타 프로젝트

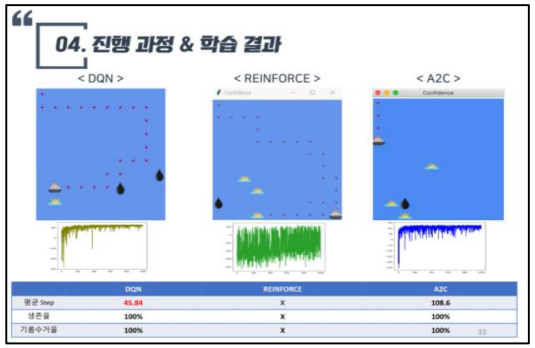
빅데이터 연합 동아리 ToBig's – 제 6회, 7회 딥러닝 컨퍼런스 참여

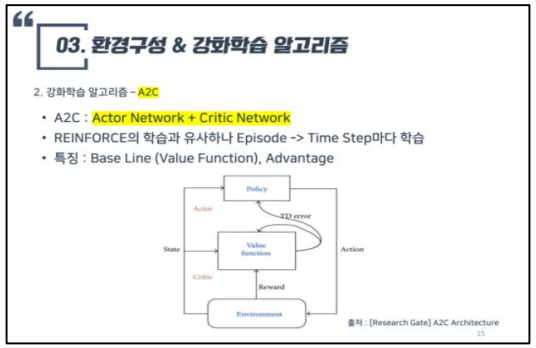
- 2018.01-2019.01 동안 전국 빅데이터 연합 동아리 투빅스 활동을 진행하며 총 2차례의 딥러닝 컨퍼런스를 수행하였음
- 제 6회 투빅스 딥러닝 컨퍼런스: YOLO와 CNN을 활용한 한국어 지문자 번역기 구축 프로젝트 링크: http://www.datamarket.kr/xe/index.php?mid=board_pdzw77&page=3&document_srl=44513

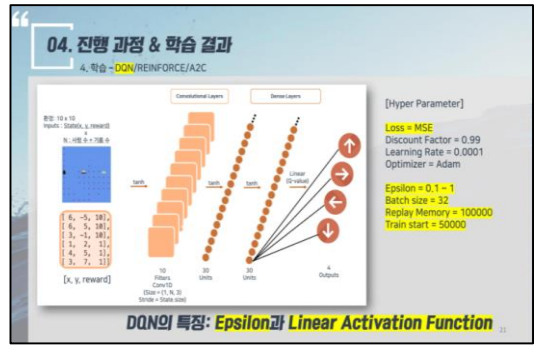




- 제 7회 투빅스 딥러닝 컨퍼런스: 강화학습 기반의 재난 구조 에이전트 구축 프로젝트 링크: http://www.datamarket.kr/xe/index.php?mid=board_pdzw77&page=2&document_srl=50423







기타 프로젝트

(학부) 데이터 분석 프로젝트

- 2017.09-2019.12 동안 대학교 재학 중 수업, 교내 빅데이터 분석 학회 D&A에서 하기 프로젝트를 수행하였음
  - (D&A 학회, 2017) 인터넷 이용 기록 분석을 통한 이용자 성별 연령 예측 프로젝트
  - (학부 수업, 2018) 서울시 부동산 가격 예측 프로젝트
  - (학부 수업, 2018) 노래 가사와 앨범커버를 이용한 힙합/비힙합 분류
  - (학부 수업, 2018) Kaggle – Human protein atlas image classification
  - (D&A 학회, 2018) 제 1회 D&A 빅데이터 컨퍼런스 – Automatic music transcription with machine learning
  - (학부 수업, 2019) CLIO – 마케팅 애널리틱스 산학 협동 – 텍스트 마이닝 기반 추천 시스템 구축, 하계 인턴 연계

(대학원) 산학 과제

- 2019.06-2021.08 동안 연구실에서 총 4차례의 산학과제를 수행하였음
  - (산학과제, 2019) SK hynix – 객체 탐지 활용 반도체 불량 탐지 모델 개발
  - (산학과제, 2020) 삼성전자 – 제품별 수요 예측 모델 개발
  - (산학과제, 2020) 국방과학연구소(ADD) – 강화학습을 통한 AI 조종사 개발 및 기종 변경을 위한 전이 학습기법 연구
  - (산학과제, 2020) LG innotek – 이미지 & 정형 데이터 활용 수율 예측 모델 개발

Thank you