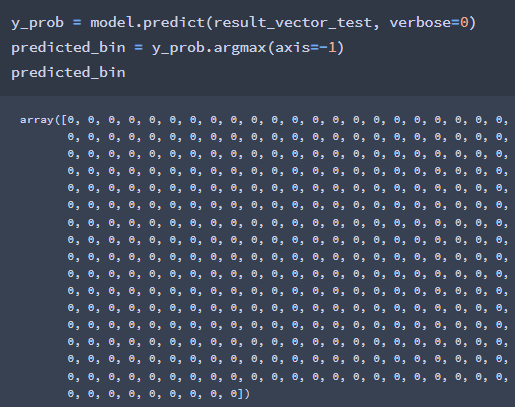
**Weekly report**

날짜 : **2023-01-05**

연구원 : **최윤석**

* **이전 수행계획**
  + **차량 이미지 분류** :
    - 차량 이미지 분류 1차 보고서 작성
  + **관계형 테이블 임베딩** :
    - 수치형 데이터를 이산화한 것을 범주형 데이터로 바꾸는 임베딩 방식을 보완하는 방법을 찾고 실험하여, 기존의 윤종찬 연구원의 연구와 **융합 가능도 척도**를 사용하여 성능 비교
    - 각 컬럼마다의 중요도를 고려해서 임베딩하는 방식과 수치형 컬럼에서 숫자의 크기가 크고 작음에 부여된 의미를 고려하는 임베딩 방식 연구
    - 나만의 딥러닝 아키텍처 그려보기, 기존의 것과 나만의 것을 분리하여 정리
    - AUC를 적용해 모델의 성능을 제대로 검증해보고, 만약 성능이 좋지 않고, 융합 불가능으로만 예측하는 경향이 보이면 학습 데이터와 라벨을 융합 가능한 것만 존재하게 교체해서 학습
* **수행결과**
* **차량 이미지 분류 :**
* 차량 이미지 분류 1차 보고서 SSD 파트 작성 완료
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + 2개의 라벨, 3개의 라벨, 10개의 라벨 전부 비슷한 성능의 모델이 나오는 것을 확인. 이는 라벨 대부분이 융합 불가능(융합 하)으로 되어있기에, 불가능으로만 예측해도 좋은 성능이 나오기에 이런 결과가 나온 것으로 생각됨.
  + 융합 가능 / 불가능 2개의 라벨로 학습한 모델을 활용해 Test Data를 predict한 결과  
    
  + 모든 입력에 대해 0으로 예측함. Test Data의 정답 라벨이 대부분 0이어서 좋은 성능을 보이는 것처럼 보이지만, 사실은 쓸모없는 모델임.
  + AUC를 적용해 모델의 성능을 파악하기로 함.

텍스트이(가) 표시된 사진

자동 생성된 설명

* + AUC score 0.5이면 해당 모델이 융합 가능 / 불가능을 구분할 수 있는 능력이 없는 것과 같음.
  + 융합 불가능 라벨의 수가 융합 가능 라벨보다 너무 많아서 이런 현상이 발생하였다고 생각해 융합 가능 라벨 95개와 융합 불가능 라벨 475개(랜덤 추출)로 1:5 비율의 새로운 학습 데이터를 생성.

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

* + 새로운 학습 데이터로 모델을 학습해도 여전히 0만을 예측하는 문제가 발생.
  + 이를 해결하기 위해 융합 가능 라벨들을 증강시키는 아이디어를 생각.

텍스트이(가) 표시된 사진

자동 생성된 설명

* + 융합 불가능한 2,405개의 데이터는 그대로 넣고 융합 가능한 95개의 데이터는 단순 증강으로 25배 증강해 융합 불가능 2,405개 융합 가능 2,375개 데이터를 학습 데이터로 모델 학습

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

* + 모델 학습 결과 테스트 데이터에 대해 Accuracy 0.922 AUC score 0.834로 준수한 성능을 가지는 것을 확인. 다만, 직접 예측한 값과 정답 라벨을 비교해본 결과 같은 테이블끼리 묶인 행렬의 융합 가능 라벨은 잘 예측하지만 서로 다른 테이블끼리 묶인 행렬의 융합 가능 라벨은 잘 예측하지 못함.
* **GAN 논문 세미나 :**
  + StackGAN 논문 읽으며, GAN 관련 사전지식(GAN, Upsampling) 공부
* **수행계획**
  + **차량 이미지 분류** :
    - SSD 소스코드 파일 정리 및 소스코드에 대한 간략한 설명 작성
  + **관계형 테이블 임베딩 :**
    - 수치형 데이터를 이산화한 것을 범주형 데이터로 바꾸는 임베딩 방식을 보완하는 방법을 찾고 실험하여, 기존의 윤종찬 연구원의 연구와 **융합 가능도 척도**를 사용하여 성능 비교
    - 각 컬럼마다의 중요도를 고려해서 임베딩하는 방식과 수치형 컬럼에서 숫자의 크기가 크고 작음에 부여된 의미를 고려하는 임베딩 방식 연구
    - 나만의 임베딩 아키텍처 / 융합 가능성 아키텍처(임베딩 아키텍처에 dependant한) 그려보기, 기존의 것과 나만의 것을 분리하여 정리 (PPT로)
    - 모델의 성능이 안 좋은 이유 찾고 해결. (데이터의 문제 / 임베딩 방식의 문제) – 데이터의 문제 : 융합 가능한 라벨들을 찾아서 추가 , 임베딩 방식의 문제 : 이전 논문에서 읽었던 Field-wise Network 방식을 적용
  + **GAN 논문 세미나 :**
    - StackGAN 논문 읽고 세미나 준비 (PPT 30분 이내 발표) , 각 논문의 핵심사항을 알아내는게 목적. 연구 동기, 배경, 관련연구 등은 생략하고 이해한 범위 내에서 핵심사항 중심으로 발표 진행. 성능 평가는 한 페이지(메인 그래프만)로 충분함. 그림과 키워드 위주로 발표.
* **기타사항**
  + **데이터 융합 프로세스**



* **융합 가능도 척도**
  + 임베딩 거리 기준 값을 정하고 기준값보다 임베딩 거리가 가까우면 융합 가능한 데이터셋이라 판단함
  + 융합 가능한 데이터셋을 사람이 하나씩 확인하여 유의미한 융합이 되는 데이터셋을 찾음
  + 을 융합 가능도 척도라 정의함



혼합형 데이터셋 예시

* **1월 초 연구 내용 정리 후 발표**