**Weekly report**

날짜 : **2022-12-08**

연구원 : **최윤석**

* **이전 수행계획**
* **차량 이미지 분류 :**
  + 실험결과 정리, 각 실험마다 mAP / 수행시간(모델 학습 , 테스트)을 테이블 형태로 정리
* **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
  + 임베딩 방식에 대해 관련 논문을 읽고 임베딩 아이디어 얻기
  + 딥러닝 모델을 생성하여 Hidden layer의 파라미터를 새로운 임베딩 벡터로 만들기
* **수행결과**
* **차량 이미지 분류 :** 
  + SSD 모델 성능 표

|  |  |  |
| --- | --- | --- |
|  | **Epoch 10만** | **Epoch 20만** |
| **mAP** | 0.920 | 0.953 |
| **Train Time** | 28시간 | 56시간 |
| **Detection Time** | 2초 / 1장 | 2초 / 1장 |

* 모델 학습 마지막까지도 loss 가 크게 흔들려, learning rate 초기값은 기본 0.004 에서 0.008로 늘리고 1만 Epoch마다 learning rate에 0.9를 곱해 작아지게 설정하여 Epoch 20만으로 학습중
* **관계형 테이블 임베딩** :
  + 768차원의 최종 테이블 임베딩 벡터 2개를 묶어 (2, 768) 형태의 새로운 배열을 생성하여 딥러닝 모델의 입력 데이터로 사용
  + 총 테이블 69개 중 50개의 테이블에서 2개씩 묶어 만든 새로운 배열 2,500개를 DNN 모듈의 Train 데이터로, 19개의 테이블을 2개씩 묶어 만든 새로운 배열 361개를 Test 데이터로 사용.

텍스트이(가) 표시된 사진

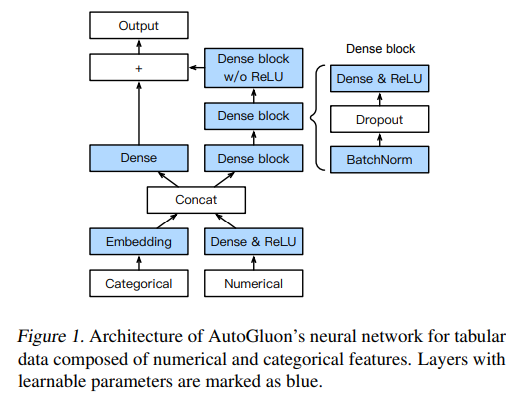
자동 생성된 설명

* + 묶인 두 벡터의 라벨 값은, 직접 테이블을 보고 융합 가능도에 따라 1 – 10의 값을 사용. 1이면 융합이 어렵고 의미가 없는 테이블, 10이면 융합이 쉽고 유의미한 테이블.

텍스트이(가) 표시된 사진

자동 생성된 설명

* + **점수 기준** : 1점을 기본으로 특정 컬럼을 통해 융합이 가능하면 +4점 불가능하면 +0점. 융합한 테이블이 유의미한 정보를 가지는 정도에 따라 +0 ~ +5점.
  + 딥러닝 모델은 DNN 구조, 4개의 Hidden layer를 사용하였고, 최종텍스트이(가) 표시된 사진

    자동 생성된 설명적으로 10개의 값 중 하나를 출력하는 분류 모델임.
  + 임베딩 아이디어를 얻기 위해 AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data 논문 읽기
    - Categorical 데이터를 임베딩할 때 논문에서는 각 Categorical 데이터가 가지는 Unique값에 비례해서 임베딩 차원을 늘려서 임베딩. 이렇게 하면 다른 Categorical 데이터와 의미가 섞이지 않고 구분되게 임베딩할 수 있다고 함.
* **수행계획**
* **차량 이미지 분류** :
  + 실험결과 정리, 각 실험마다 mAP / 수행시간(모델 학습 , 테스트)을 테이블 형태로 정리
* **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
  + 딥러닝 모듈에 입력과 테스트 데이터의 융합 가능성 라벨 작성
  + Network On Network for Tabular Data Classification in Real – world Applications 논문 읽기
  + 융합 가능도 평가 결과에 대한 다른 사람들과의 공유 및 절충
  + 나만의 딥러닝 아키텍처 그려보기, 기존의 것과 나만의 것을 분리하여 정리
* **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**