**Weekly report**

날짜 : **2022-11-24**

연구원 : **최윤석**

* **이전 수행계획**
* **차량 이미지 분류 :**
* **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
  + 혼합형 테이블 임베딩 관련 논문 공부
* **수행결과**
* **차량 이미지 분류 :** 
  + SSD 모델 학습 및 성능측정
    - 트라콤 머신 용량 부족으로 공간 확보 중 SSD 파일을 잘못 삭제하여 SSD 모델 재생성
      * 10만 epoch로 진행 – mAP는 0.92로 약간 낮아진 것을 확인  
        텍스트이(가) 표시된 사진

        자동 생성된 설명

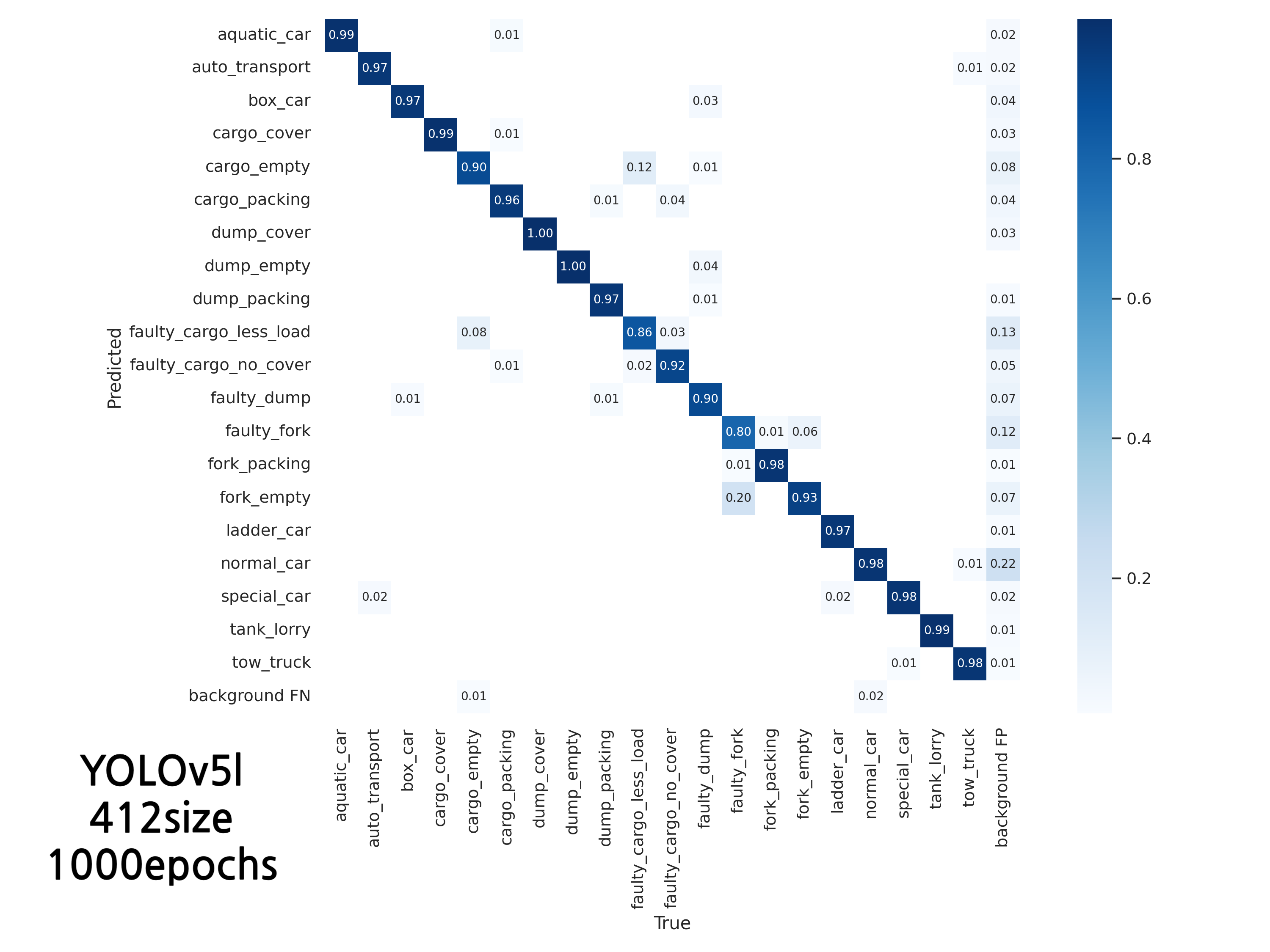
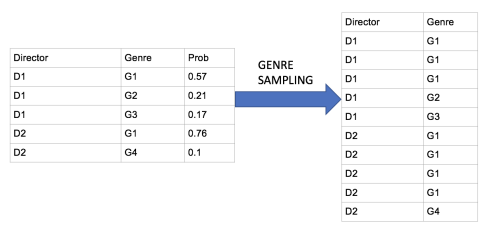


Figure 1. 김민중 연구원의 YOLO 모델 confusion matrix

* Epoch 10만개로 학습한 모델의 Average\_precision이 0.9보다 안 좋게 나온 라벨(4, 5, 9, 10, 12, 14)이 김민중 연구원의 confusion matrix에서 성능이 안 좋게 나온 라벨(4, 9, 12)과 비슷함.
* **관계형 테이블 임베딩** :
  + 텍스트이(가) 표시된 사진

    자동 생성된 설명윤종찬 연구원의 코드를 사용해 270여개의 관계형 테이블 데이터에서 융합 가능도 척도를 측정함.

Figure . 임베딩 벡터의 코사인 유사도가 0.975 이상이고 카테고리가 다른 두 테이블들만 조사해 융합가능도 척도를 계산한 결과

* + - 270 \* 270 개의 코사인 유사도를 구하고 그 값이 0.975 이상인 데이터 쌍이 29개, 그 중 카테고리가 다른 데이터 쌍이 9개 그 중 융합이 가능하다고 판단한 쌍이 1개로 결과는 0.11 정도로 좋지 않았음
  + On Embeddings in Relational Databases 논문 학습
    - 논문은 IMDB 영화 데이터셋을 여러가지 임베딩 방식을 사용하여 성능 비교 진행
    - 그 중 Genre Sampling 방식을 관계형 테이블 임베딩에 활용할 방안 모색중
    - Genre Sampling은 임베딩의 효율을 높이기 위해 의미적 레벨에 좀 더 가중치를 두는 방식을 사용. 아래의 그림에서 수치형 데이터 prob의 차이를 W2V가 인식하지 못하기에 prob의 크기만큼 레코드 수를 넣어, 같은 레코드에 동시에 사용되는 경우를 늘려 W2V 가중치를 두는 방식을 사용
* **수행계획**
* **차량 이미지 분류** :
  + 20만 epoch로 모델 학습 후 평가
* **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
  + 캐글에서 융합이 가능한 데이터셋을 찾아 데이터셋에 추가

* **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