**Weekly report**

날짜 : **2022-12-01**

연구원 : **최윤석**

* **이전 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
* **수행결과**
* **차량 이미지 분류 :** 
  + SSD 모델 학습 및 성능측정
    - 트라콤 머신 용량 부족으로 공간 확보 중 SSD 파일을 잘못 삭제하여 SSD 모델 재생성
      * 20만 epoch로 진행 – mAP는 0.953으로 약간 낮아진 것을 확인  
        텍스트이(가) 표시된 사진

        자동 생성된 설명

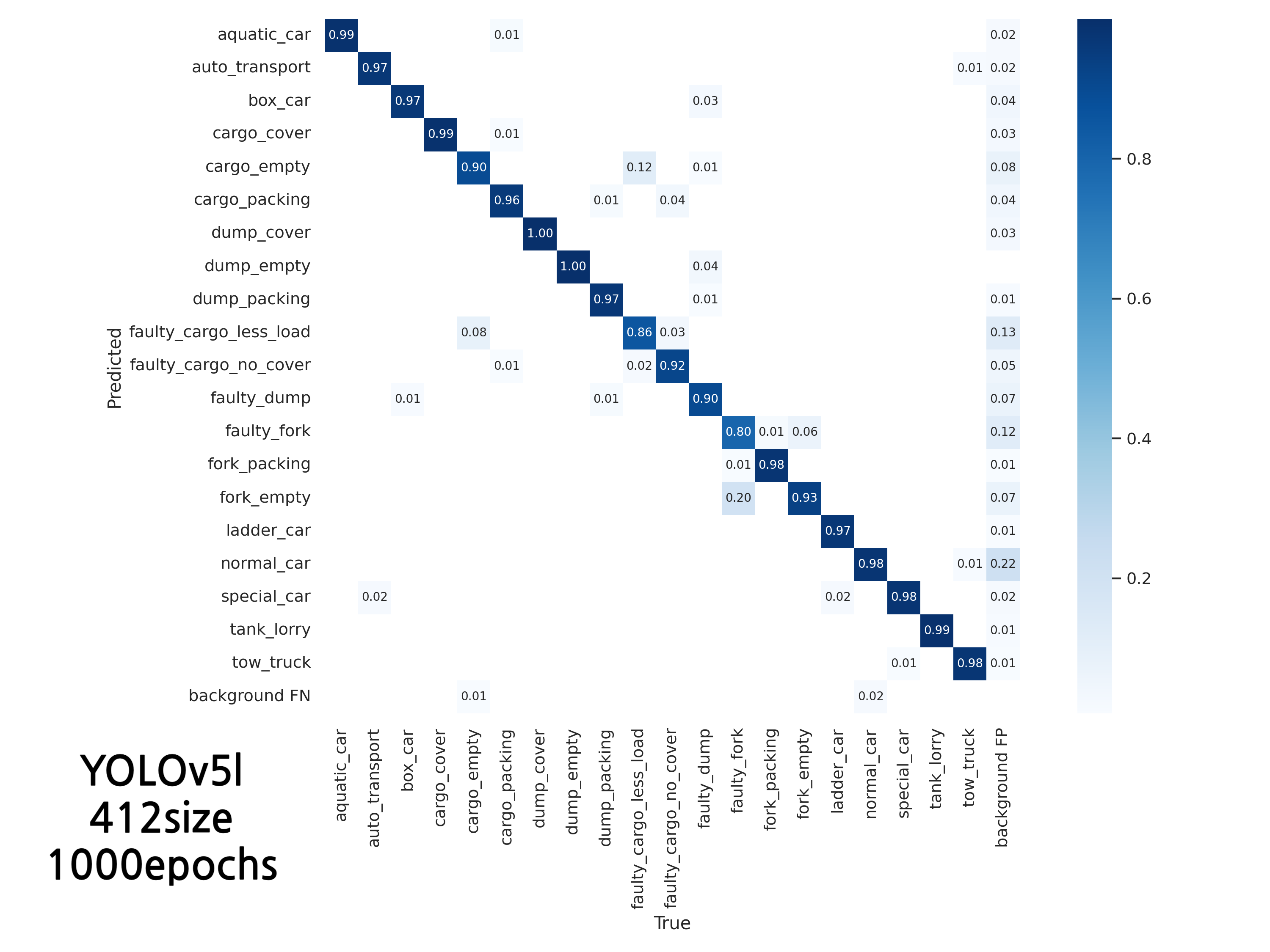


Figure 1. 김민중 연구원의 YOLO 모델 confusion matrix

* Epoch 20만개로 학습한 모델의 Average\_precision이 0.9보다 안 좋게 나온 라벨(12, 14)이 김민중 연구원의 confusion matrix에서 성능이 안 좋게 나온 라벨(9, 12)과 비슷함.
* 또한 다른 라벨에 비해 상대적으로 Average\_precision이 잘 나오는 라벨과 안 나오는 라벨이 같은 것을 확인할 수 있다.
* **관계형 테이블 임베딩** :
  + 캐글에서 카테고리가 다르지만 융합이 가능한 테이블 데이터들을 찾아 데이터 셋에 추가

Figure . 융합 가능한 데이터셋과 그 이유

* + - 총 4쌍, 8개의 데이터셋을 추가
  + 추가적으로 5개의 카테고리(Food, Travel, Science and Technology, Arts and Entertainment, Game)에 대해 각각 3개씩의 데이터를 추가
  + 융합 가능한 데이터셋 + 카테고리별 데이터셋을 추가하여 총 23개의 데이터 추가하여 총 70개의 데이터 셋을 수집
* **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
  + 위의 임베딩 방식에 대해 관련 논문을 읽고 구체적인 아이디어 얻기
  + 임베딩 벡터가 가까운 융합 가능한 데이터 셋이 탐색되기 위해 가까워지길 기대,
  + 딥러닝 모델에서 융합 가능한 것을 넣으면 임베딩벡터가 가까워야 되고 융합 가능하지 않으면 임베딩벡터가 멀어야
  + A / B 는 90%정도로 융합가능 B / C 는 60%정도로 융합가능. 이런식의 정답을 만들어서 학습데이터로 사용, 딥러닝 모델을 학습하여 hidden layer의 파라미터를 임베딩 벡터로 가져오기
  + 입력을 위한 임베딩 벡터(1차 임베딩 벡터) – 기존 윤종찬 연구원의 임베딩 방식으로 생성
  + 융합을 위한 임베딩 벡터(2차 임베딩 벡터) – 딥러닝 모델의 hidden layer에서 가져오기

* **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