Assignment 3

G202448018 최은경

- Process

Marabou를 사용하여 외부 신경망 모델(MLP, FashionMNIST)과 데이터셋에 대해 property 기반 검증을 진행하였다.

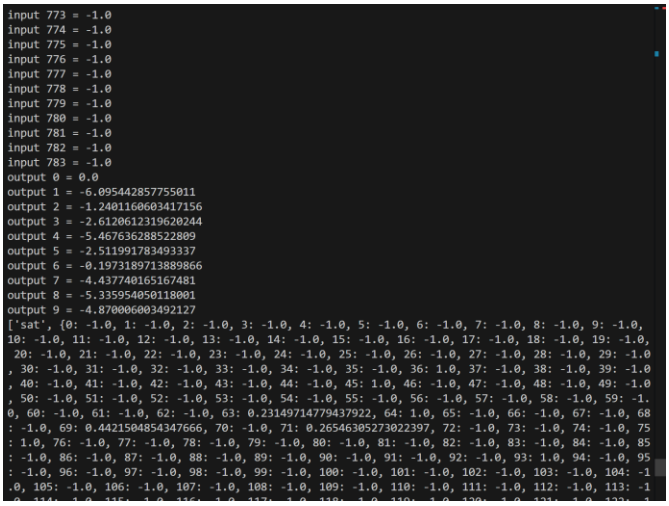
모델 및 데이터셋은 다음과 같다.

- 2-layer MLP, 28x28 입력, 32 hidden, 10 출력
- FashionMNIST 데이터셋
- PyTorch로 학습 (Test Accuracy: 약 85.7%)
- ONNX로 변환 후 Marabou에 입력 (참고한 링크는 다음과 같다)
  - https://neuralnetworkverification.github.io/Marabou/Examples/0_NNetExample.html

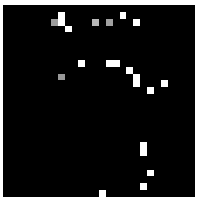
Marabou 실험 과정은 다음과 같다.

- 입력 변수 : 28×28=784차원 벡터 (inputVars), 각 값의 범위는 -1.0 ~ 1.0
- Property 설정 : Marabou에 입력의 bound를 -1.0 ~ 1.0로 제한
- 목적: Marabou SMT solver로 FashionMNIST 2-layer MLP(ONNX 변환 모델)에 대해, output[0] (0번 클래스의 logit)이 [0, 10] 범위를 만족시키는 입력(이미지)을 찾기
- Solver 실행: network.solve() 수행, SAT/UNSAT 여부 및 solution vector 반환

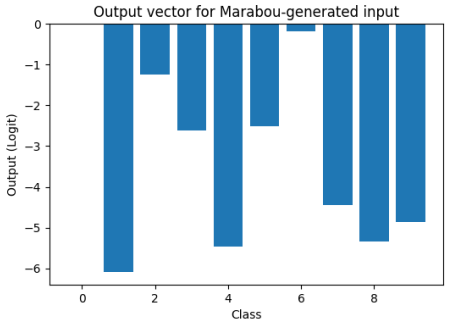
결과 및 해석은 다음과 같다.

- Marabou 반환값: SAT 반환. 즉, 주어진 property를 만족하는 입력이 존재함
- 반환 input/output 예시화면:
  - 

- Marabou가 찾은 입력 벡터(784차원): 설정해놓은 제약(property)을 만족하는 모델의 입력. 예를 들어, output 0의 logit이 0.0이 나오도록 하는 입력을 찾으라는 property를 줬다면, 실제로 그런 입력이 존재함을 보여주는 것으로 파악
- input 벡터의 특성: 대다수 값이 -1.0 (최소 bound), 일부 값이 1.0 또는 0~1 사이의 값으로 보임.
- 실제 이미지로 복원하면 사람 눈에는 전혀 의미 없는 sparse한 패턴 (실험 이미지 첨부)



- output 해석: output[0]=0.0, 나머지는 -6~-1 사이의 음수(logit 값). 해당 입력이 주어졌을 때 모델은 output 0에 대한 logit을 0.0, 나머지 클래스는 모두 음수로 예측. 즉, 해당 입력에서는 0번 클래스를 가장 높은 logit으로 분류한다는 의미



결론은 다음과 같다.

Marabou를 이용해 property 기반 검증에 성공했다. 모델의 property (예: 특정 logit이 가장 높도록 하는 입력) 만족 입력이 존재함을 확인하였으며 해당 입력은 실제 데이터셋에 없는, 인위적인 패턴의 입력임을 관찰했다.