7조 발표 시작하겠습니다.

목차는 다음과 같습니다. 프로젝트 목표, 데이터 분석 및 전처리 과정, 모델을 선정하고 학습 시켰던 과정, 파이널 테스트, 순으로 설명드리겠습니다.

먼저, 프로젝트의 목표입니다. 저희는 handmade 데이터 셋의 성능이 저하된 원인을 분석하고 개선된 ml 모델을 학습하고 최적화 하는것에 더불어, 높은 점수보다 오버피팅을 막는, 최대한 내성이 강한 모델을 만들고자 하였습니다. 이를 위해 노이즈가 있는 데이터, 테두리가 있는 데이터, 시프트 된 데이터 등에 대한 데이터 전처리를 진행하였습니다.

데이터 분석 및 전처리 과정입니다.

첫번째로 15개의 클래스 이외의 데이터를 모두 제거하였습니다.

다음과 같이 분리 되어있던 데이터를 15개 클래스 분류기를 위한 데이터로 통합하였습니다. 이후 이외의 것을 제외하고 나니 다음과 같은 데이터 셋으로 정리되었습니다

또한 전처리 과정에서 잘못 라벨링 된 것은 육안으로 제거하였고 학습에 애매한 데이터또한 육안으로 판별 후 제거하였습니다.

이과정에서 transet 1910개 +testset 209개 합하여 약 2100개 삭제 하였습니다.

약 30000개의 데이터를 이용하여 Combined Dataset를 구성하였습니다

픽셀값(feature 값)의 차이는 파이프라인에 Normalizer( )를 추가 하여 해결하였습니다

노이즈가 있는 데이터는 노이즈를 추가하여 , trainset , testset 구축하였고 파이프라인에 denoising 함수 추가하여 해결하였습니다

shift된 데이터, 테두리가 남은 데이터는 파이프라인에 중앙화, 최대화 과정 추가하여 보완하고자 하였습니다.

비슷한 숫자 및 기호의 존재하여 삭제하였고 향후 Confusion Matrix 분석 할 예정입니다.

현재 28\*28 데이터만 있는데,  현재 가진 데이터를 확대, 축소해서 새로운 데이터를 만드는 데 한계가 있어 현재 방식으로 진행하였습니다.