Project Guide

학기	2023년도2학기	교과목	머신러닝	담당교수	이찬호
설계 목표	Hand-made MNIST dataset을 이용한 머신러닝 모델 최적화 및 분석				
운용방안	팀을 구성하여 설계 프로젝트를 진행				
설계 방법	○ 프로젝트 목표 Hand-made MNIST dataset과 original MNIST dataset을 이용하여 inference 결과를 비교하고 Hand-made dataset의 성능이 저하된 원인을 밝히고 Hand-made dataset을 포함한 학습데이터를 이용하여 인식 성능을 개선시킨 machine learning model을 학습시키고 최적화한다. ○ 프로젝트 수행 방법 1. 0 ~ 9 까지의 숫자와 +, -, x, /, = 을 손으로 쓰고 스캔하거나 사진을 찍어 각 숫자 및 기호 data instance를 MNIST instance와 같은 28x28 크기의 grey scale 이미지로 만들어 제출한다. 0 ~ 9의 숫자는 1인당 4장(총 400자, 숫자당 40자), 기호는 4장(총 400자, 기호당 80자)을 제출하며 배포하는 이미지 변환 코드를 이용하여 npz 파일로 변환하여 380x380 해상도의 이미지와 함께 제출. 이미지 변환 코드는 0 ~ 9의 숫자에 대한 것으로 연산 기호에 대해서는 1번 cell의 label 변경 필요. * 이미지 변환 코드 실행후 3번 cell 실행 결과에서 숫자 및 연산기호가 제대로 변환되었는지 확인. 하나의 숫자나 연산 기호라도 식별 불가능한 데이터가 포함될 경우 0점 처리 * 4장의 이미지는 각 팀원의 지인이나 가족이 작성하거나 필체를 달리하여 작성하여야 학습 데이터의 다양성을 확보할 수 있음. 2. 팀별로 숫자당 10개씩 100개의 hand-made test dataset을 구성한다. 이때, test dataset은 제출하는 학습 데이터와는 별도로 생성한다. 실습코드의 machine learning model과 original MNIST dataset을 이용하여 학습시킨 후 hand-made test dataset과 원래의 MNIST test dataset에 대해 각각 prediction후 score를 비교한다. 3. Hand-made dataset의 성능이 저하한 원인을 다양한 방법을 이용하여 분석한다. Dataset의 특성의 차이를 이미지 특성과 다양한 attribute 계산을 통해 분석하고 원인을 추론한다. 4. Original MNIST dataset과 강의시간(실습코드 포함)에 배운 machine learning 알고리즘을 이용하여 성능이 좋은 모델 후보군을 찾고 배포한 hand-made dataset을 추가하여(hand-made test dataset 제외) 선택한 machine learning model을 최적화하여 최종 모델을 선택한다. (Tensorflow나 Pytorch 등 NN 사용금지. 단, Scikit Learn의 NN 사용 가능) 5. 최종 model은 별도의 hand-made MNIST test dataset을 이용하여 평가함 (요구 조건) 1. 4인으로 팀을 구성 2. 프로젝트 수행계획을 세우고 업무마다 역할 분담: 프로젝트 수행계획서 제출 - 프로젝트 개요 및 목표, 수행과정에 따른 역할 분담 3. 프로젝트 수행일지 작성 4. 제출자료: 수행계획서, 보고서, 발표자료, 실행결과 포함한 노트북 파일. 모든 파일은				

	pdf와 ipynb로 제출 (팀당 1개씩). ○ 보고서 및 발표자료 포함사항 1. 프로젝트 개요 및 목표 2. 수행계획 및 역할 분담 (수정한 계획 및 수행일지): 팀원간 합의한 개인별 기여도 기록. 팀원 모두 동일한 기여도를 부여하는 것은 인정하지 않으며, 모두 동일한 기여도를 부여할 경우 사유를 작성하고. 수행일지의 기록과 일치해야함. 팀장이 있을 경우 팀장의 기여도는 다른 팀원보더 높아야 함. 3. 수행과정 가. MNIST dataset 추가 및 분석: 새로운 data instance를 추가하기 전과 후의 dataset 특성 비교 (Get the data/Discover and visualize the data) - npz 파일 또는 이미지 파일을 이용하여 데이터를 읽어 기존 dataset에 추가 - Dataset 분류: 기존 data와 새로운 data를 training/validation/test dataset으로 어떻게 분배할지를 결정. **총 3개의 dataset 준비: original, hand-made, combined** (Prepare the data) 나. Hand-made dataset과 original dataset을 이용하여 inference 성능 및 원인 분석: 필요시 cleaning data 작업. '마'에서 최적화의 효과를 높이는 가장 중요한 단계. 라. 학습에 사용할 모델 선택: combined dataset을 이용하여 다수의 후보를 대상으로 가장 성능이 좋은 최종 모델 선택. 근거와 학습 계획 제시. (Select and train a model) **노트북 파일에 이러한 과정이 나타나야 함.** 마. 모델 최적화 및 분석: - combined dataset을 이용한 학습 및 model 최적화 - Model과 training hyperparameter의 최적화를 통해 최대 성능을 획득. - 최적화 과정 제시 및 결과 분석. 학습시간, 예측시간(inference time), 정확도 측면에서 분석. Original dataset과 combined dataset으로 각각 학습한 모델에 대해 orignal test dataset과 hand-made test dataset에 대한 성능 비교 분석. - Epoch에 따른 learning curve 제시. - **노트북 파일에 이러한 과정이 나타나야 함.** (Fine tune the model) 4. 결론
제출결과물 및 일정	1. 팀구성 : 10/16(월) - 10/22(일) 23:59 2. 수행계획서: 10월 29일(일) 23:59 3. Hand-made dataset: 11월5일(일) 23:59 4. 중간 발표자료: 11월 19일(일) 23:59 5. 최종보고서, 최종 발표자료, 실행결과 포함한 노트북 파일 (pdf, pptx, ipynb 파일을 zip으로 묶어 제출. 팀당 1개씩): 12월3일(일) 23:59 6. 발표: 12월4일(월) 대면 강의시간+ 온라인
평가방법	1. 수행계획서(팀별): 보고서 내용(1) 2. 데이터셋(개인별): 보고서 내용(1) 보고서 형식(1) 3. 중간보고서(팀별): 보고서 내용(1), 보고서 형식(1) 4. 최종보고서(팀별): 보고서 내용(5), 보고서 형식(3), 결과의 검증 (5), 발표 (2), 팀웍(2), 팀원간 기여도 평가(개별평가, 3)

	- 보고서 내용 (내용의 충실도 및 문제 인식과 해결 과정의 질적 우수성) - 보고서 양식 (보고서 양식 준수 및 문제 인식과 해결 과정 포함 여부) - 결과의 검증 (문제 해결 방법의 검증 과정 및 근거) - 팀워크 (역할분담의 적절성) - 기여도 (개인별 점수) - 발표 (프로젝트 내용을 이해하고 각 팀원이 본인의 역할에 따라 발표 여부, 발표시간 준수) * 문제의 분석과 solution 도출 과정이 나타나야 하며 (채점시 이 부분을 중점적으로 봅니다) 팀원별 역할과 기여가 나타나야 합니다. 아이디어 및 분석 결과에 기여자 및 기여자를 수행계획서 및 보고서에 표시하고 팀원별 기여도를 총합이 100%가 되도록 자체적으로 평가하여 작성.

[첨부 1] 프로젝트 수행일지

팀원	팀원1	팀원2	팀원3	팀원4	팀원5
역할					
모임일자	발표내용				

[첨부 2] 이미지 예제

원본 스캔 이미지 (2642 x 3592)

해상도 변경후 이미지 (380x380)