

Cogsys Winter Code Hack: 2회차

End-to-end pipeline with collaboration



Daehyun Cho

Cognitive System Lab

A.I. Dept, Korea Univ.

전체적인 계획

1. 대회 참여하기 & GitHub Repository 개설하기D
2. Pytorch DataLoader 구축하기 & Code Review하기
3. Pytorch Model 개발하기 & 제출해보기
4. 여러 가지 실험을 하기 위해 세팅하기 (w/ Hydra) & wandb logging하기
5. Pytorch-Lightning으로 갈아타보기
6. Accuracy 100% 만들기

이 과정을 조금 불친절하게 진행할 예정

지향

- ✓ 알잘딱깔센
- ✓ Collaborations
- ✓ 조금이라도 모르면 구글링 (선무당이 사람잡음)

지양

- ✗ Copy & Paste: 단순한 syntax정도는 괜찮은데, 전체 파이프라인을 베껴오는 건 취지를 아주 많이 벗어남
- ✗ 나의 온 시간을 투자하기: 짬짬히... 틈틈히 진행하는 것을 권장 (교수님한테 혼나기 싫음)

TODO List

- ✓ Download data to local repository (DIY)
 - ✓ Put data file/directory in .gitignore (prevent data files being uploaded to github) (DIY)
 - ✓ Configure python environments with Conda
 - ✓ Exploratory Data Analysis
 - ✓ Plot data
 - ✓ Apply Augmentations and plot
 - ✓ Create torch Dataset & DataLoader
-
- ✓ Commit as more often as you can during your work
 - ✓ Push your work to remote repository





Before we start

- 예시코드를 넣어봤습니다.
- Step-by-step으로 commit을 넣긴했는데, 놓치는 부분이 있을 때 확인해주시면 좋을 듯 합니다.

Commits on Jan 12, 2023		
Add augmentations and its visualizations 1pha committed now	92b9d4b	<>
Add multiple plots 1pha committed 6 minutes ago	44e7595	<>
Add digit plots for sanity check 1pha committed 15 minutes ago	387c374	<>
Make torch datasets & dataloader 1pha committed 19 minutes ago	0d89858	<>
Add typehints to help readability 1pha committed 20 minutes ago	d962575	<>
Make digit vector to matrix 1pha committed 26 minutes ago	ca9c2cd	<>
Add digit plotting code 1pha committed 30 minutes ago	62e7611	<>
Add train.csv reading code 1pha committed 36 minutes ago	9201c4f	<>
Add requirements.txt 1pha committed 48 minutes ago	e59d2fc	<>
Add .DS_Store to .gitignore. Ignore desktop.ini for windows users 1pha committed 48 minutes ago	083e1e4	<>
Edit README.md with links and title 1pha committed 1 hour ago	e0e5400	<>
Add .csv file in .gitignore 1pha committed 1 hour ago	3d9f2d9	<>



- ☒ 저를 Collaborator로 추가해주세요 (1pha 라고 검색하면 모리 사진 나옴)
- 지난 주에 진행한 것 처럼 branch를 따서 code review를 받으실 수 있습니다.
- 혹시 강하게 크고 싶으시면 Branch protection rule에 main branch를 Approve 필수로 걸어두시면 제가 승인할 때만 넘어갈 수 있도록 도와드립니다



Add a collaborator to digit-recognizer



Select a collaborator above

repository can view it. this repository. Only you can



.gitignore

Kaggle

- .gitignore는 GitHub가 Tracking하지 않을 파일들을 등록합니다.
-  여기에 우리의 데이터인 .csv 파일을 무시하도록 파일에 추가해봅시다.
-  또한 windows는 desktop.ini 파일을 무시하도록, mac유저는 .DS_Store 파일을 무시하도록 설정해주세요.



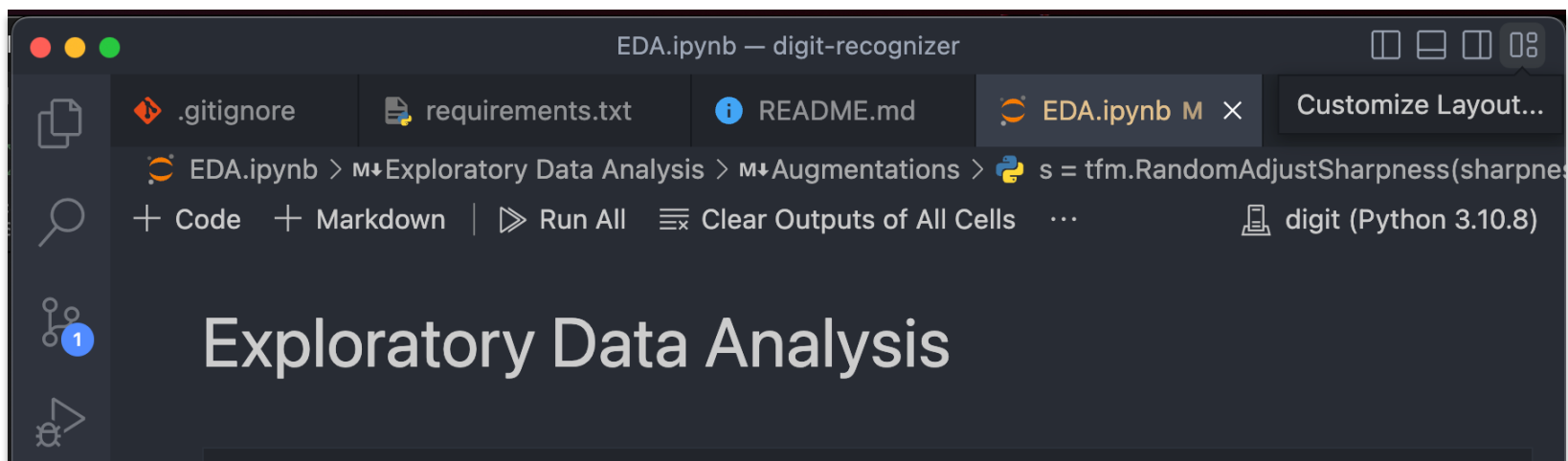
Configure python environments with Conda

- ✓ Python 가상환경이 뭔지 대충 찾아봅시다. +conda
- ✓ Python version 3.10이면서 이름이 digit인 conda environment를 만들어보세요
- ✓ 거기에 **PyTorch, scikit-learn, numpy, pandas, jupyter, ipykernel, matplotlib, seaborn, torchvision**를 설치해보세요 (Hint: Pypi)
- ✓ 방금 설치한 라이브러리들을 requirements.txt 파일을 만들어서 정리해보세요 (important for reproducibility)
 - Requirements.txt에는 라이브러리 버전과 함께 넣어주세요
- ✓ 새로운 가상환경 하나를 더 만들어서 방금 만든 requirements.txt에 있는 라이브러리들을 한큐에 설치해보세요
pip install -r requirements.txt 로 가능합니다.
시도해봤으면 가상환경 다시 지워요.
- ✓ 만들어낸 가상환경이 어디 있는지 한 번 찾아보세요
Hint: Anaconda가 어디에 설치되어있는지 찾아보세요 (/opt~)






Configure python environments

- 코드를 작업하기 전에, jupyter notebook을 활용해보시다.
- Jupyter notebook은 interactive하게 코드들을 바로 실행하고 결과를 시각화해서 볼 수 있는 유용한 툴입니다.
- Vscode에서는 새 파일을 만들고 확장자를 .ipynb로 만들어주면 됩니다.
- ☒ EDA.ipynb 파일을 만들어서 작업해줍시다.
- 여기서 우상단에 **kernel**을 고를 때 digit을 골라줍니다. 만약 conda에 jupyter, ipykernel이 설치되어있지 않으면 kernel을 고를 때 나타나지 않을 수 있습니다.



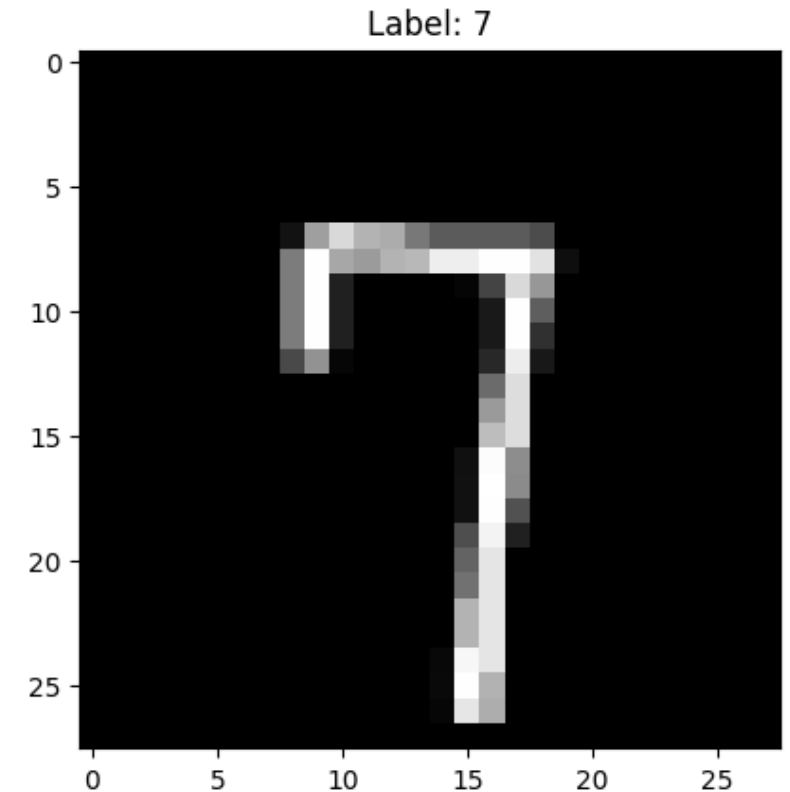


- 우리의 데이터는 이미지이지만 이미지가 아닙니다. .csv라는 comma-separated values라는 파일에 저장되어 있어요.
- 애를 열고 가지고 놀려면 pandas를 써야합니다.
-  Pandas 아까 설치했으니까 이제 데이터를 열어보세요
 -  How to open csv files with pandas?
- 잘보면 28×28개(=784)의 값을 한 줄로 불러냈습니다.
-  애를 numpy로 변환해서 이미지로 그려봅시다.

	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	pixel9	...	pixel774	pixel775	pixel776	pixel777	pixel778	pixel779	pixel780	pixel781	
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
...	
41995	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
41996	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
41997	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
41998	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
41999	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
42000 rows x 784 columns																				



- 이미지 데이터니까 그려봐야합니다
- ☒ Matplotlib에서 imshow 함수를 이용해서 그려보세요
- ☒ 기왕 그리는 거 라벨도 같이 title에 넣어보세요
- ☒ 그림 그리는 코드를 함수화해보세요.
 - 중간 중간 함수화 해놓는 습관을 길러야합니다.
 - 함수화를 해놓으면 몇 가지 일처리들을 하나로 묶어놓을 수 있습니다.
 - 이게 왜 중요하냐면 복잡하다고 생각되는 일을 함수 하나로 정의해두면, 더더더 복잡한 작업들을 쉽게 정의하고 떠올릴 수 있게 됩니다.





- [illegible]



To PyTorch

- ✓ DataLoader에 넣어서 batch 단위로 뽑힐 수 있게 해보세요
- 잘보면 반환된 값의 첫 번째 shape가 batch_size랑 일치하는 것을 알 수 있습니다.
- ✓ Batch 안에는 여러 개의 데이터가 들어있는데, 모두 잘 들어왔는지 확인하는 코드를 짜보세요 + 함수화해보세요

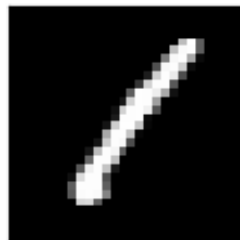
```
ds = DigitDataset()
dl = DataLoader(dataset=ds, batch_size=16)

X, y = next(iter(dl))

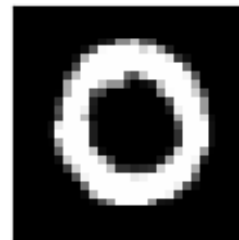
X.shape, y.shape
✓ 0.9s

(torch.Size([16, 28, 28]), torch.Size([16]))
```

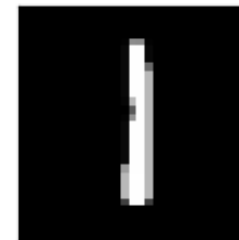
Label: 1



Label: 0



Label: 1



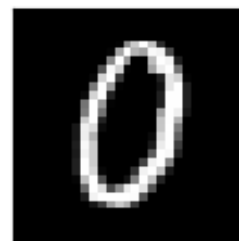
Label: 4



Label: 0



Label: 0



Label: 7



Label: 3

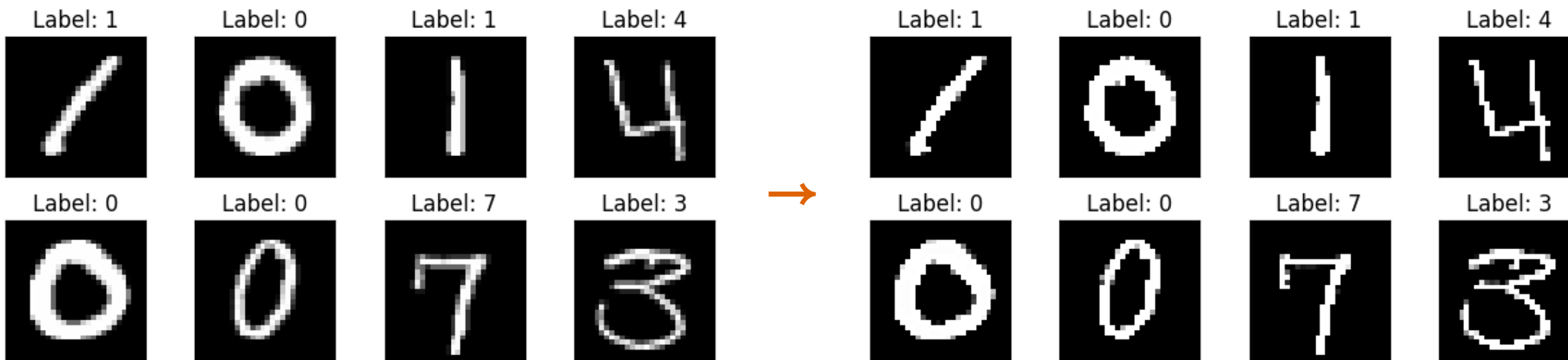




Configure python environments with Conda

- ☒ torchvision 라이브러리를 통해 이미지들을 변환해보세요.
 - 나중에 Data augmentation에 활용될 것입니다.
- ☒ Augmentation에서 사용되는 인자들을 바꿔보면서 진행해보세요.
 - Augmentation에 정답은 없기 때문에 내가 사용하는 데이터에 적합한 변화를 찾아야합니다.
- ☒ 또한 현재 int64로 0 - 255 사이의 값이 들어있는데, 딥러닝 모델은 작은 값에서 더 좋은 학습을 해낼 수 있기 때문에, `torch.transforms.Normalize`로 적절한 학습을 수행할 수 있도록 변환해주세요

e.g. [RandomAdjustSharpness](#)





Configure python environments with Conda

- ☒ 적절한 augmentation 조합을 찾았으면 [torchvision.transforms.Compose](https://pytorch.org/docs/stable/transforms.html#torchvision.transforms.Compose)를 통해 하나의 transform으로 변환할 수 있도록 구성해주세요.
- ☒ 잘 진행되었으면 augmentation을 dataloader 안에 넣어주세요 (__getitem__을 진행했을 때 augmentation된 결과와 라벨이 반환되도록)
- ☒ 열심히 만든 파일들을 script화 해주세요 (.py 파일 안에 넣어주세요)
- ☒ 결과가 잘 나오는지 EDA.ipynb에서 확인해주세요. (Sanity check, always!)

```
> .ipynb_checkpoints
v data
  sample_submission.csv
  test.csv
  train.csv
v digitrec
  > __pycache__
  dataloader.py
  .gitignore
  EDA.ipynb
  README.md
  requirements.txt
```

Thank you 🙏

Daehyun Cho

1phantasmas@korea.ac.kr