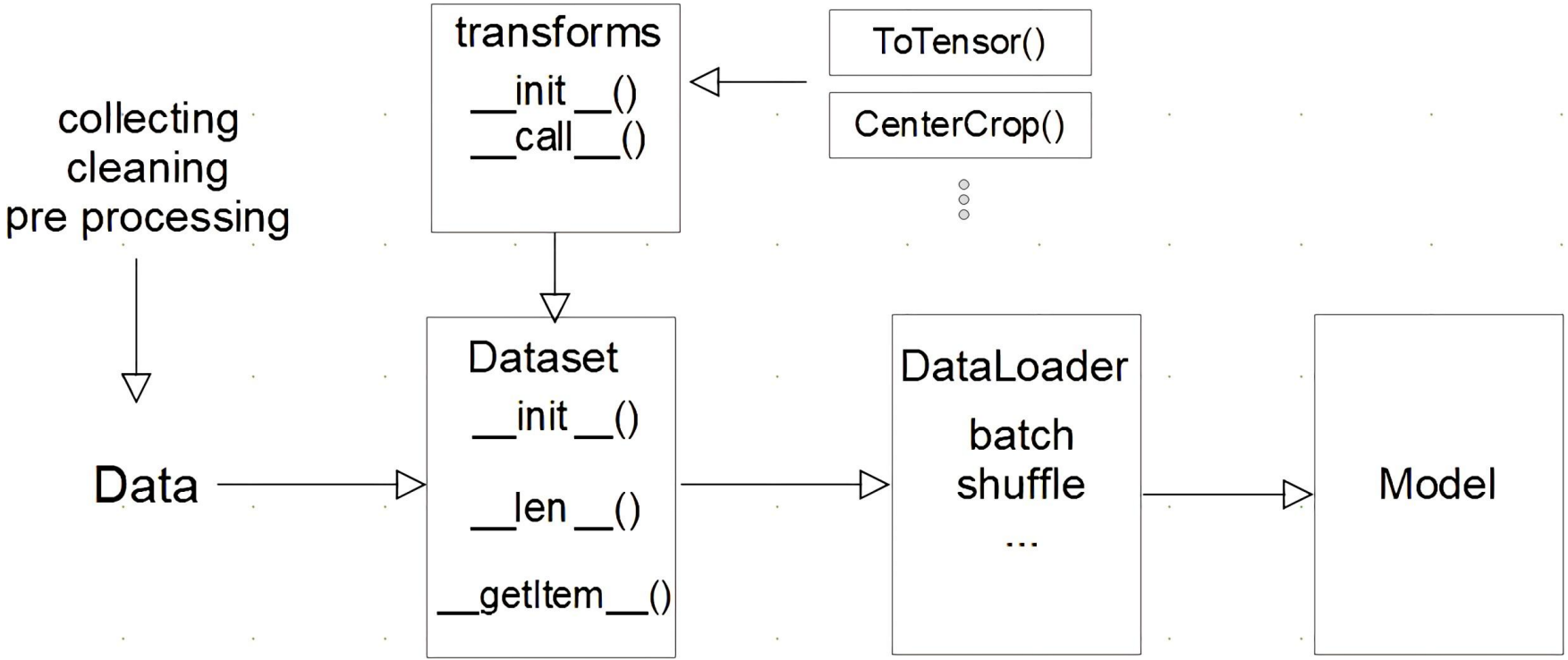
PyTorch datasets & dataloaders

TEAMLAB director

최성철

모델에 데이터를 먹이는 방법



Dataset 클래스

- 데이터 입력 형태를 정의하는 클래스
- 데이터를 입력하는 방식의 표준화
- **Image, Text, Audio** 등에 따른 다른 입력정의

Dataset 클래스


```
import torch
from torch.utils.data import Dataset

class CustomDataset(Dataset):
    def __init__(self, text, labels):
        self.labels = labels
        self.data = text

    def __len__(self):
        return len(self.labels)

    def __getitem__(self, idx):
        label = self.labels[idx]
        text = self.data[idx]
        sample = {"Text": text, "Class": label}
        return sample
```

초기 데이터 생성 방법을 지정

데이터의 전체 길이

index 값을 주었을 때 반환되는 데이터의 형태 (X, y)

Dataset 클래스 생성시 유의점

- 데이터 형태에 따라 각 함수를 다르게 정의함
- 모든 것을 데이터 생성 시점에 처리할 필요는 없음

  : **image의 Tensor** 변화는 학습에 필요한 시점에 변환
- 데이터 셋에 대한 표준화된 처리방법 제공 필요

  → 후속 연구자 또는 동료에게는 빛과 같은 존재
- 최근에는 **HuggingFace**등 표준화된 라이브러리 사용

DataLoader 클래스

- Data의 Batch를 생성해주는 클래스
- 학습직전(GPU feed전) 데이터의 변환을 책임
- Tensor로 변환 + Batch 처리가 메인 업무
- 병렬적인 데이터 전처리 코드의 고민 필요

DataLoader 클래스

Dataset 생성

```
text = ['Happy', 'Amazing', 'Sad', 'Unhappy', 'Glum']
labels = ['Positive', 'Positive', 'Negative', 'Negative', 'Negative']
MyDataset = CustomDataset(text, labels)

MyDataLoader = DataLoader(MyDataset, batch_size=2, shuffle=True)
next(iter(MyDataLoader))
# {‘Text’: ['Glum', ‘Sad'], 'Class': ['Negative', 'Negative’]}
```

DataLoader Generator

```
MyDataLoader = DataLoader(MyDataset, batch_size = 2, shuffle = True)
for dataset in MyDataLoader:
    print(dataset)
# {‘Text’: ['Glum', 'Unhappy'], 'Class': ['Negative', 'Negative’]}
# {‘Text’: [‘Sad', ‘Amazing'], 'Class': ['Negative', ‘Positive’]}
# {‘Text’: [‘happy'], 'Class': [‘Positive']}
```

DataLoader 클래스


```
DataLoader(dataset, batch_size=1, shuffle=False, sampler=None,
           batch_sampler=None, num_workers=0, collate_fn=None,
           pin_memory=False, drop_last=False, timeout=0,
           worker_init_fn=None, *, prefetch_factor=2,
           persistent_workers=False)
```

https://subinium.github.io/pytorch-dataloader/

Casestudy

- 데이터 다운로드 부터 **loader**까지 직접 구현해보기
- **NotMNIST** 데이터의 다운로드 자동화 도전

http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html

End of Document
Thank You.