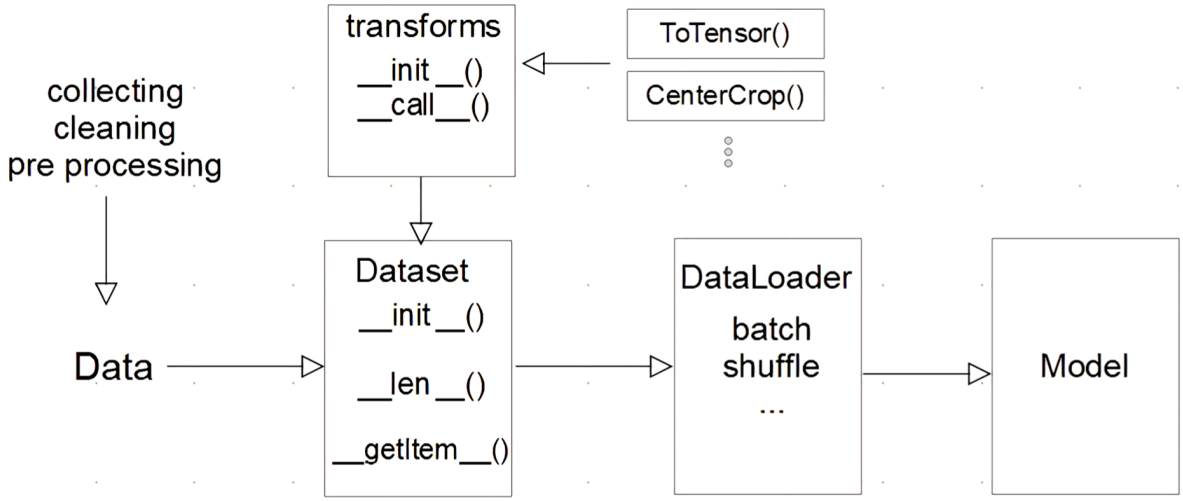
Dataset & Dataloader



- **Datasete Class**

  데이터 입력 형태 정의하는 클래스

  데이터 입력하는 방식 표준화

  img, text, audio 따라 다른 입력 정의

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

class CustomDataset(Dataset):
  # 초기 데이터 생성 방법 지정
  def __init__(self, text, labels):
    self.labels = labels
    self.data = text

  # 데이터 전체 길이
  def __len__(self):
    return len(self.labels)

  # index 값 주었을 때, 반환되는 데이터 형태 (X, y)
  def __getitem__(self, idx):
    label = self.labels[idx]
    text = self.data[idx]
    sample = {"Text": text, "Class": label}
    return sample
```

  - **생성 시, 유의점**

    데이터 형태에 따라 각 함수 다르게 정의

    모든 것을 데이터 생성 시점에 처리 필요 X

    => img의 Tensor 변화는 학습에 필요한 시점에 변환

데이터 셋에 대한 표준화된 처리방법 제공 필요

최근 HuggingFace 등 표준화된 라이브러리 사용

- **DataLoader Class**

  data의 Batch 생성해주는 클래스

  학습직전 (GPU feed 전) 데이터의 변환 책임

  Tensor로 변환, Batch 처리

  병렬적 데이터 전처리 코드 고민 필요