PyTorch로 딥러닝 제대로 배우기

- 중급편 -

Part4-1. 데이터

강사: 김 동 희

목차

I. 데이터

1) 데이터
2) 인코딩
3) Table Data 인코딩
4) 이미지 데이터 인코딩
5) 음성 데이터 인코딩
6) 비디오 데이터 인코딩

II. Dataset & Data Loader

1) Loading Dataset
2) DataLoader

I. 데이터

1. 데이터

❑ 데이터

- 데이터베이스, 데이터 마이닝, 데이터 기반 의사결정, 데이터 사이언스, …

- 좋은 모델을 만들기 위해서는 좋은 데이터를 확보하는 것이 필수적!
  - 데이터의 **양** (크기가 크면 좋다)
  - 데이터의 **완결성** (비어있는 값이 없으면 좋다)
  - 데이터의 **신뢰도** (현실을 잘 계측한 데이터가 좋다)
  - 데이터의 **시기절적함** (timeliness, 필요할 때 수집하고 사용할 수 있어야 좋다)

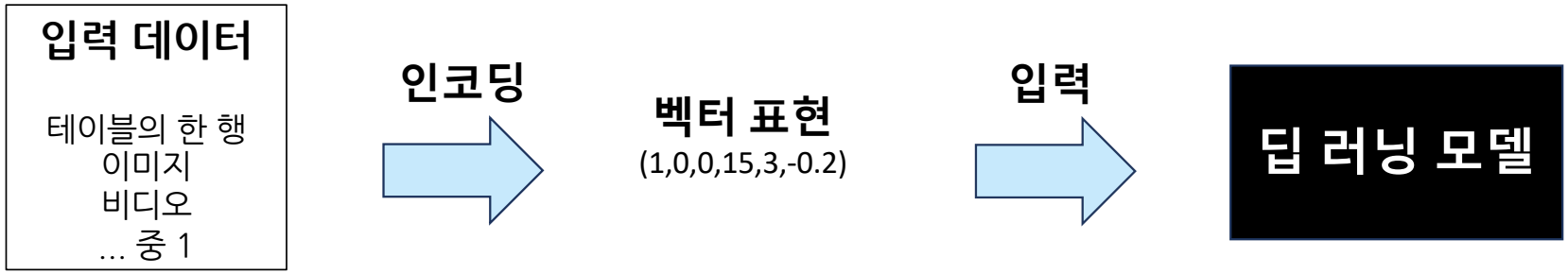
1. 데이터

❑ **데이터를 기술하는데 사용하는 단어**

- 데이터의 **특성**: 정형 데이터와 비정형 데이터
- 데이터의 **크기**: 데이터의 행 수, 차원 수, 크기
- 데이터의 **타입**: 범주형 변수와 연속형 변수
- 데이터의 **용도**: 학습 데이터, 검증 데이터, 평가 데이터

2. 인코딩

- 컴퓨터는 결국 수를 다루는 계산기이다. 테이블, 이미지, 비디오 등의 입력 데이터를 수치로 변환하는 과정을 **인코딩(encoding)** 작업이라고 한다.
  - 전처리 작업 중 하나
- 인코딩을 하고 나면, 입력 데이터는 정해진 개수의 차원으로 이루어진 **벡터(vector)**로 변환된다.



- 실제로는 각 유형의 데이터를 인코딩하고, 전처리할 때 다양한 기법을 적용한다.
  - 노이즈를 줄이고 정확도를 높이기 위함
- 어떤 입력 데이터든 결국 수치로 변환되어 딥 러닝 모델에 입력된다는 점을 기억한다.

3. Table Data 인코딩

입력 데이터

상장유무	상장
규모	중소기업
직원 수	153
매출	278억

- 남은 차원들은 연속형 차원이므로 수치를 그대로 사용하여 인코딩

- 단, 이 경우 값의 단위가 다르므로 각 차원의 최솟값을 0, 최댓값을 1로 두는 식의 정규화를 적용 가능
  · min-max normalizatoin

벡터 표현

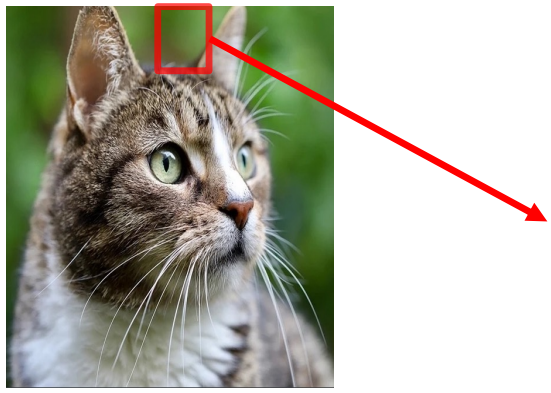
(0, 0, 0, 1, 153, 278)

4. 이미지 데이터 인코딩

❑ 이미지 인코딩

- 일정한 구조를 갖지 않는 데이터(unstructured data)

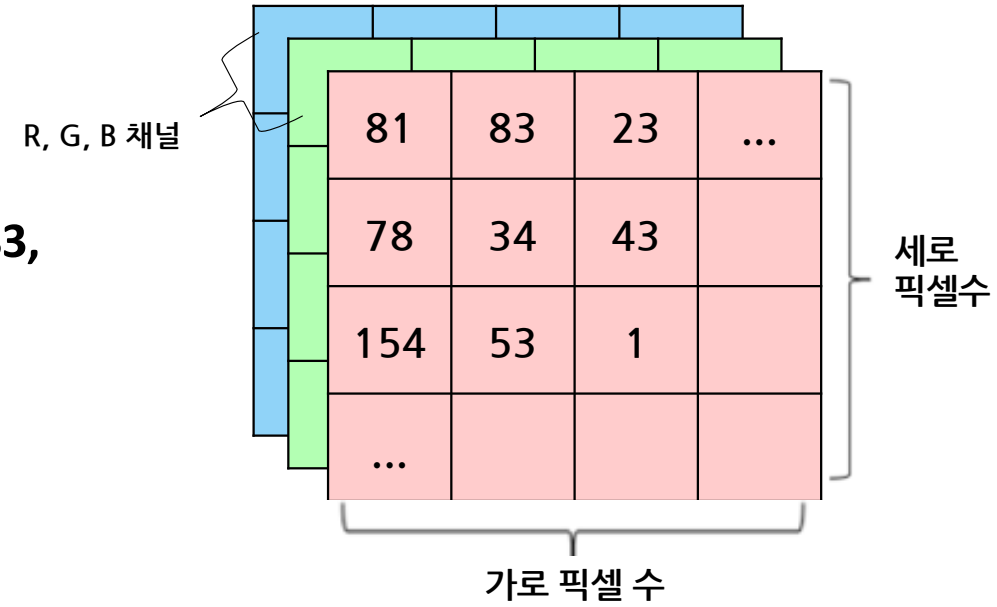
이미지 데이터



벡터 표현

(81, 109, 36, 83, 113, 45, ...)

텐서 표현



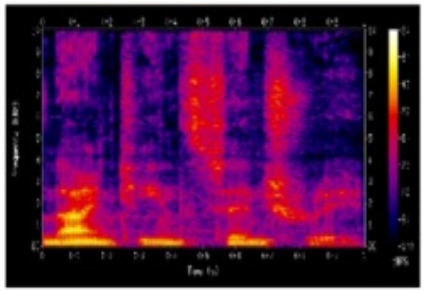
5. 음성 데이터 인코딩

❑ 음성 데이터 인코딩

- 일정한 구조를 갖지 않는 데이터(unstructured data)

음성 데이터

주파수 분석



텐서 표현

-30	-20	-10	...
20	0	10	
10	-20	-50	
...			

주파수 대역

재생 프레임 수

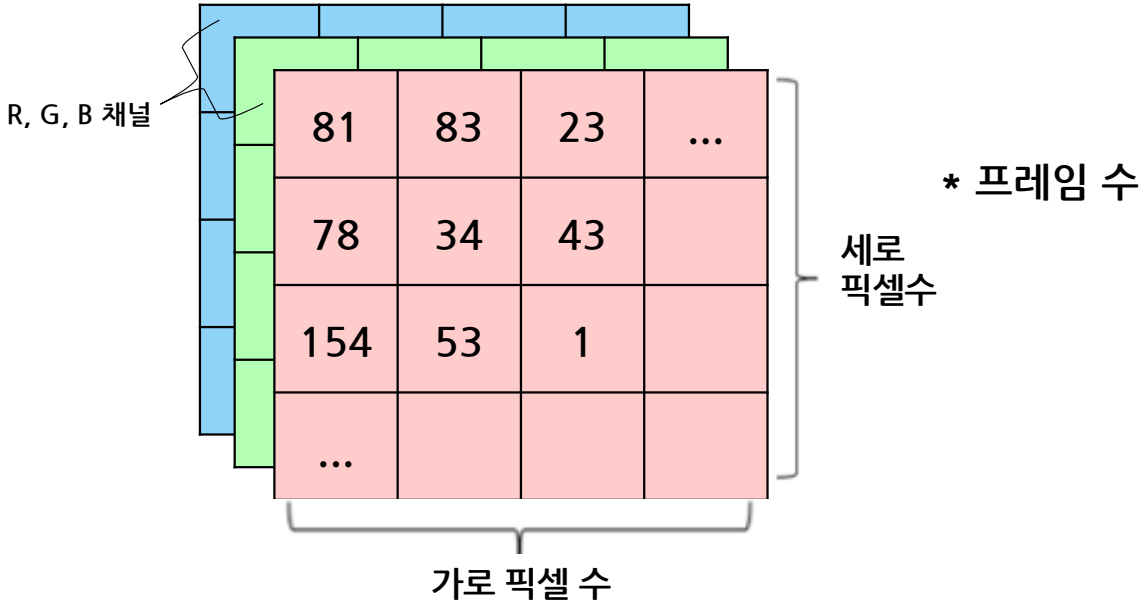
6. 비디오 데이터

❑ 비디오 데이터 인코딩

- 일정한 구조를 갖지 않는 데이터(unstructured data)

비디오 데이터

텐서 표현



감사합니다.