인공지능이란?
학습데이터의 중요성

학습 주제

학습데이터란 무엇일까?

01 학습데이터의 중요성

02 학습데이터 구축의 과정

학습 목표

1. 학습데이터의 중요성에 대해서 알 수 있다.

2. 학습데이터의 구축과정에 대해서 알 수 있다.

학습데이터의 중요성

언뜻 생각해보면 복잡한 수식으로 나타나는 인공지능 모델은 어렵게 느껴지면서 자연스럽게 중요하다고 느껴질 것입니다. 이와 반대로 데이터는 그냥 있는 주어지는 것이라고 생각하기 때문에 중요하다고 느끼지 못할 것입니다. 하지만 실제 인공지능 프로젝트의 80%는 인공지능 모델 개발에 사용할 데이터를 수집하고 정제하고 라벨링하는 작업일정도로 데이터는 중요하게 다뤄집니다.

데이터 중요성 2Q

데이터의 중요성은 2가지의 Q에서 찾을 수 있습니다.

바로 질(Quality)과 양(Quantity)입니다.

Quality + Quantity

(품질) (양)

01 학습데이터 구축의 중요성

데이터의 Quality (품질)

데이터의 질(Quality)의 측면에서 보면. 데이터가 얼마나 모델(종속변수 Y)을 잘 설명할 수 있는지를 보아야 합니다.

예를 들어 고등학교 3학년 학생들의 시험 성적을 예측하는 인공지능 모델을 만든다고 생각해 봅시다. 키, 사교육비, 혈액형, 학습시간, 점심 식사량, 최근에 치른 시험 성적 등의 데이터로 모델을 학습시킬 수 있을 것입니다. 심지어 해운대에 상어가 출몰한 횟수도 사용할 수 있습니다.

데이터의 Quality (품질)

하지만 위의 데이터 중에서 키나 혈액형, 점심 식사량, 해운대에 상어가 출몰한 횟수는 시험 성적과는 관련성이 거의 없을 것입니다. 이처럼 종속변수를 잘 설명하지 못하는 필요 없는 데이터들은 전체 데이터셋의 질을 떨어뜨리고 모델을 불안정하게 만듭니다.

데이터의 Quantity (양)

데이터의 양(Quantity)의 측면에서 보면, 아무리 선정한 데이터들이 모델을 잘 설명한다고 하여도 데이터가 충분하지 않다면 현실을 잘 반영하지 못할 것입니다. 아까와 같이 고등학교 3학년 학생들의 시험 성적을 예측하는 인공지능 모델을 만든다고 생각해 봅시다.

데이터의 Quantity (양)

모델을 잘 설명할 수 있는 사교육비, 학습시간, 최근에 치른 시험 성적을 가지고

다음 주에 치를 중간고사 성적을 예측할 것입니다.

언뜻 보아도 해당 데이터들은 다음 주에 치를 시험 성적을 잘 설명해줄 것으로 보입니다.

그런데 전체 500명의 학생들 중에서 5명의 데이터만으로 모델을 학습시킨다면 제대로 된 모델이 나올까요?

아마 5명의 데이터는 현실을 제대로 반영하지 못할 것이고 학습시킨 모델은 제대로 예측하지 못할 것입니다.

목적에 맞는 데이터 + 적합한 양

결국 중요한 것은 인공지능 모델의 학습에 사용할 데이터가

첫째, 의도하는 목적에 맞아야 하며(Quality) 둘째, 현실을 잘 반영해야 합니다(Quantity).

아무리 만능일 것 같은 인공지능이라고 하여도 결국 학습된 데이터를 기반으로 작동하기 때문입니다.

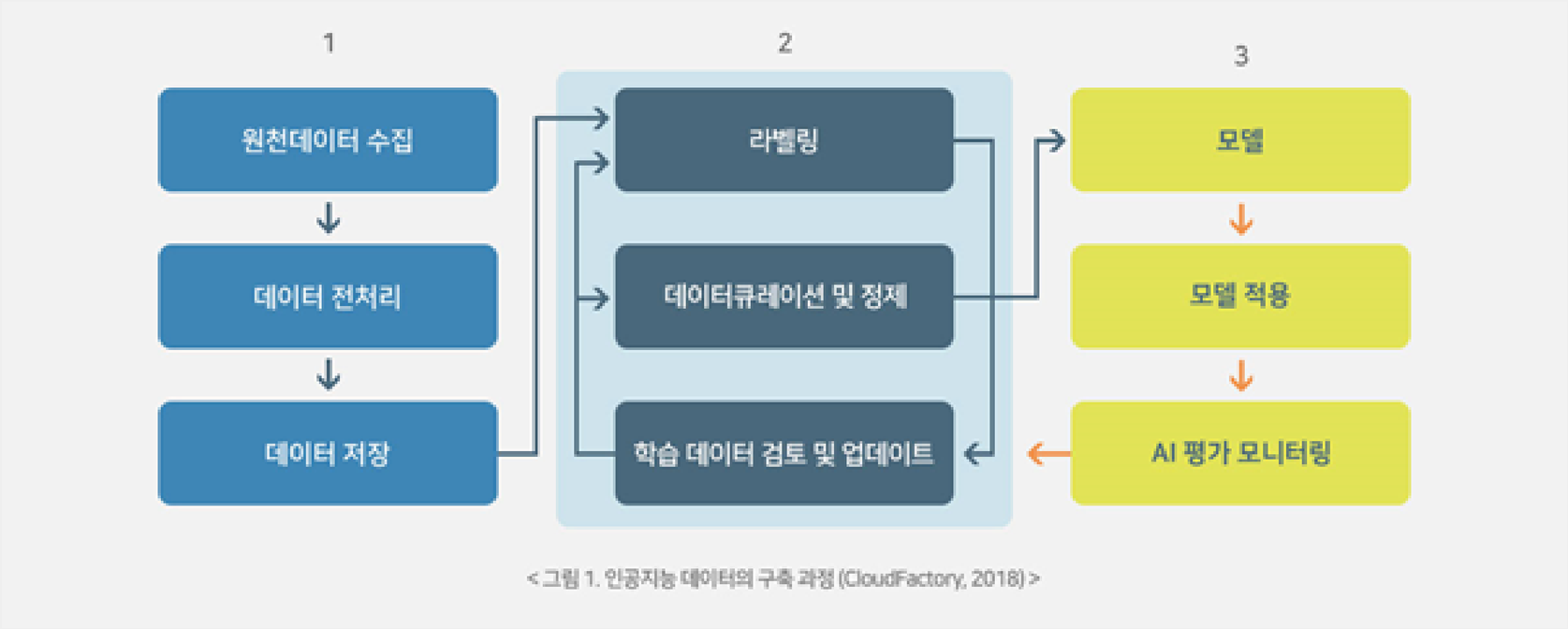
예를 들어, 한 번도 호랑이를 본 적이 없는 사람이 호랑이를 알아보지 못하는 것처럼

한 번도 호랑이 이미지를 학습하지 않은 인공지능 역시 호랑이를 알아보지 못할 것입니다.

이처럼 인공지능 모델은 우리가 가르쳐준 데이터를 기반으로 세상을 인식하기 때문에 데이터가 중요합니다.

02 학습데이터 구축과정

데이터가 중요하다는 것은 알았습니다. 하지만 아무리 그래도 대체 뭘 하기에 전체 인공지능 프로젝트에서 데이터 관련 작업이 80%를 차지하는 것일까요? 아래 그림을 따라가면서 알아보겠습니다.



< 그림 1. 인공지능 데이터의 구축 과정 (CloudFactory, 2018) >

02 학습데이터 구축과정

데이터 수집

인공지능 데이터 구축은 먼저 원시 데이터 수집부터 시작 됩니다. 목표로 한 데이터들을 수집하는 단계로 아날로그 (종이 문서) 데이터를 디지털화 하거나 수집된 적이 없는 데이터는 직접 데이터를 수집하고 기존에 다양한 곳에 흩어져 있는 데이터들(xls, txt등의 형식이나 로컬, USB, 클라우드 등 장소)을 한 곳에 모으는 단계입니다.

02 학습데이터 구축과정

데이터 전처리

다음은 한 곳에 모은 데이터들을 전처리하는 단계입니다. 전처리는 알고리즘에 넣을 수 있게 일정한 형식으로 정리하는 것을 말합니다. 예를 들어, A파일에는 남/여로 구분되어 있고 B파일에는 M/F, C파일에는 남자/여자로 되어 있다면 하나의 형식으로 통일하는 단계입니다.

02 학습데이터 구축과정

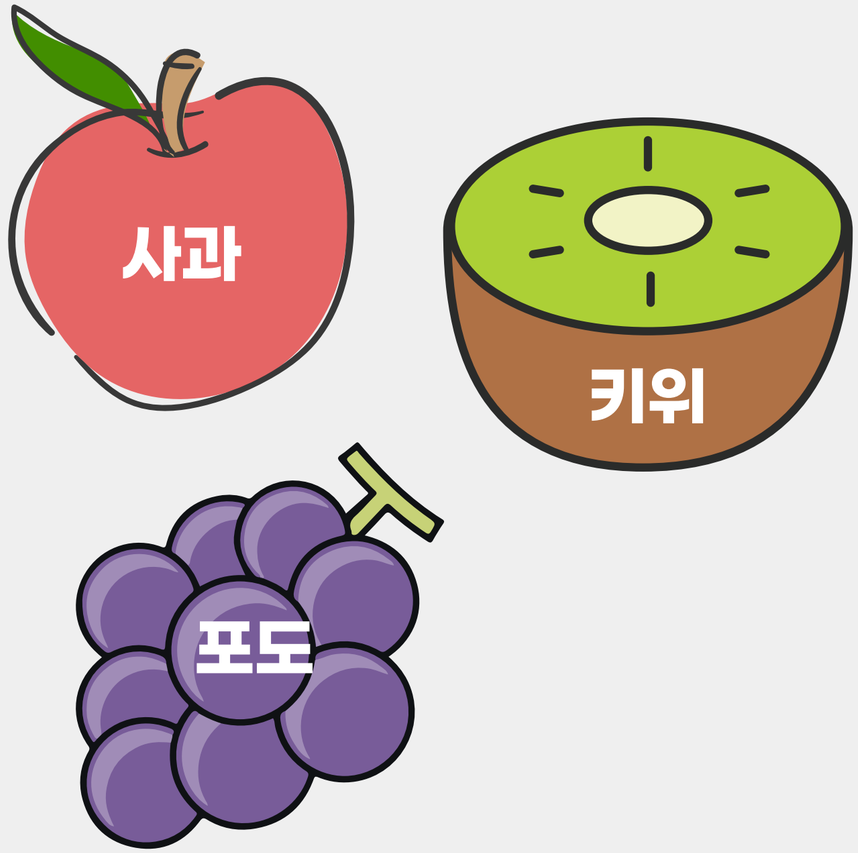
데이터 저장

마지막으로 전처리까지 완료된 데이터를 데이터베이스에 저장합니다. 여기까지는 기본적으로 데이터베이스에 저장하기 위한 단계입니다.

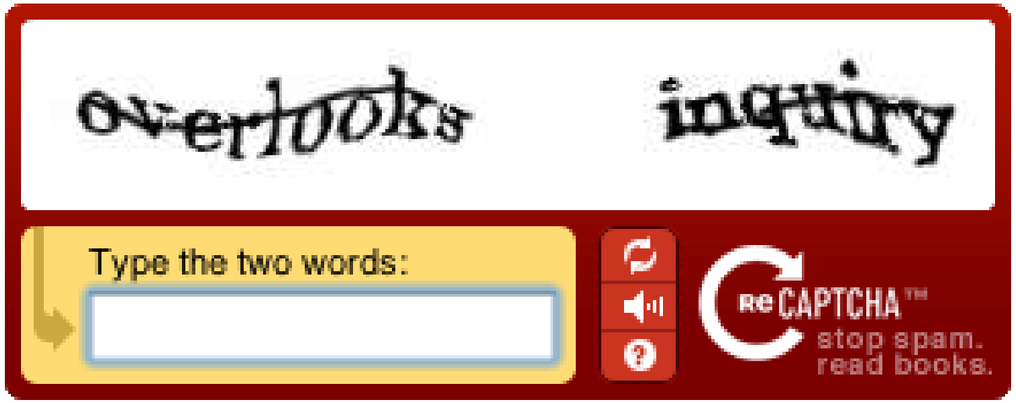
02 학습데이터 구축과정

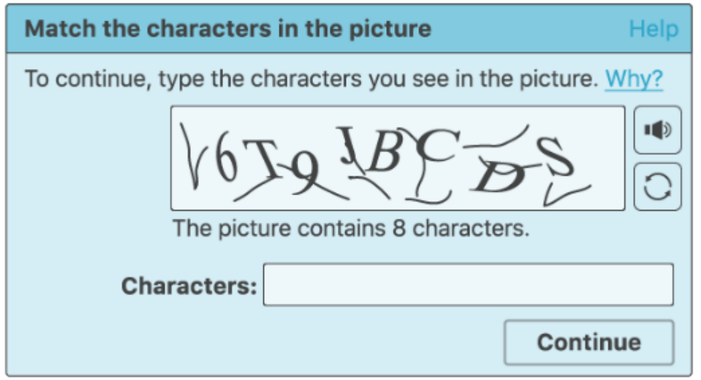
데이터 라벨링

라벨링부터는 인공지능 모델을 만들기 위해 데이터를 다시 가공합니다. 라벨링은 데이터의 정답을 입력하는 단계입니다. 지도학습과 같이 컴퓨터에게 정답을 알려줘서 학습하는 인공지능 모델은 '○ + ○ + ○ = 정답'에서 정답을 인간이 수작업으로 모두 달아주어야 합니다.



02 학습데이터 구축과정





우리는 회원가입을 할 때 이런 화면을 보게 되는데요. 캡차(CAPTCHA)라는 프로그램은

컴퓨터와 인간을 구별하여 자동 가입 등 부정 사용을 막기 위한 기술이랍니다.

우리가 컴퓨터는 알아보기 힘든 손글씨에 답을 달아주면서 인공지능 학습을 위한 데이터가 만들어지게 되는데요.

즉, 우리가 하는 3초의 작업이 데이터에 라벨링을 하는 것입니다. 이처럼 라벨을 추가하고 구조화하는 것은

인공지능에 필수적이며 고품질 데이터를 만드는 데 핵심 과정이라고 할 수 있답니다.