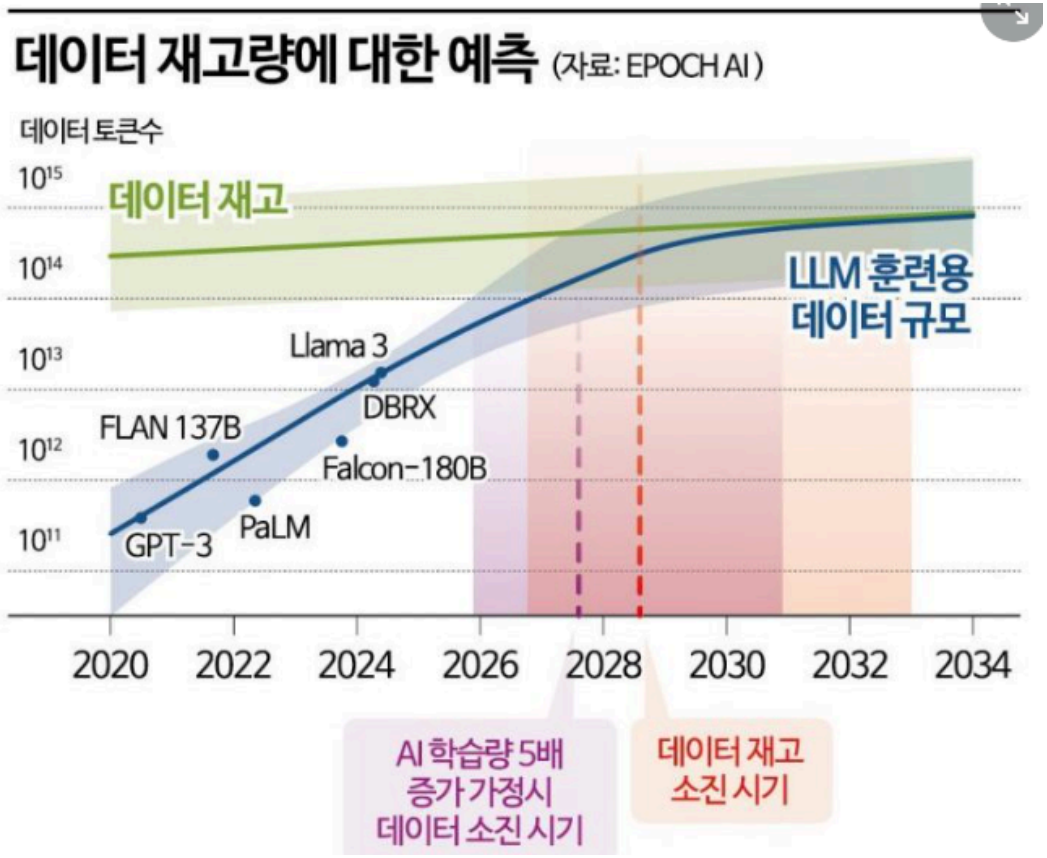
훈련데이터부족문제

2026년이면 그 AI학습 용 고품질 데이터(언어)가 고갈될 것.....

- 급속도로 성장하는 인공지능(AI) 모델이 데이터 부족 문제로 위협받고 있다는 지적이 나왔음

> - 현재 인터넷상에 존재하는 데이터만으로는 2년 내 AI 모델 성능을 높이는 데 한계를 맞을 것
> - 월스트리트저널(WSJ)은 1일(현지시간) 오픈AI의 'GPT-4'나 구글의 '제미나이'와 같은 기술이 잠재적인 데이터 부족에 직면해 있다고 보도
> - 이에 따르면 대형언어모델(LLM)의 규모가 커지면서 업계에서는 전례 없는 데이터 수요가 발생하고 있음
> - AI 기업은 향후 2년 이내에 고품질 텍스트 데이터에 대한 수요가 공급을 초과, 잠재적으로 AI 발전을 방해할 수 있다는 의견을 내놓고 있음



학습데이터 부족현상 해결법 - 전이학습(Transfer Learning)

- 학습 데이터가 부족한 분야의 모델 구축을 위해 데이터가 풍부한 분야에서 훈련된 모델을 재사용하는 머신러닝 학습 기법

> - 머신러닝 모델 구축 시에 품질에 영향을 주는 큰 요소가 학습데이터의 양과 질임
> - 편향되지 않고 이상치, 결측치가 없는 풍부한 데이터를 바탕으로 한 학습모델은 좋은 성능을 보여줄 가능성이 높음
> - 하지만 모든 분야에서 학습 데이터가 풍부한 것은 아님
> - 학습 데이터의 부족은 곧 완전하지 않은 모델(Incomplete Model)로 이어질 수 밖에 없음 -> 이러한 제약을 극복하기 위해 학습 방법 자체에 대한 연구, 즉 메타-학습에 관한 연구도 활발히 이뤄지고 있음

- 학습에 사용할만한 양질의 데이터가 충분히 확보되지 못한다면 다른 분야의 풍부한 데이터를 바탕으로 한 좋은 성능의 모델에서 일부 계층을 재활용하여 모델을 구축하는 방법을 고려해 볼 수 있음 -> 이러한 학습 기법이 바로 전이학습(Transfer Learnning)

클래스 불균형 문제

불균형 데이터란?

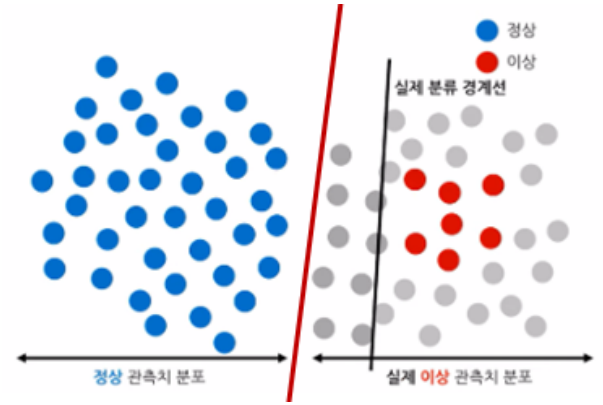
- 불균형 데이터란 정상 범주의 관측치 수와 이상 범주의 관측치 수가 현저히 차이나는 데이터

> - 예) 암 발생 환자가 암에 걸리지 않은 사람보다 현저히 적고, 신용카드 사기 거래인 경우가 정상 거래인 경우보다 현저히 적은 경우

문제점

- 정상을 정확히 분류하는 것과 이상을 정확히 분류하는 것 중 일반적으로 이상을 정확히 분류하는 것이 더 중요
- 불균형한 데이터 세트는 이상 데이터를 정확히 찾아내지 못할 수 있다는 문제점이 존재

> - 왜냐하면 보통 이상 데이터가 target값이 되는 경우가 많기 때문임



- ✓ 회색 원은 아직 관측되지 않은 모르는 데이터
- ✓ 경계선 왼쪽의 회색 원들은 실제로는 이상 데이터이기 때문에 정상 데이터로 오분류됨

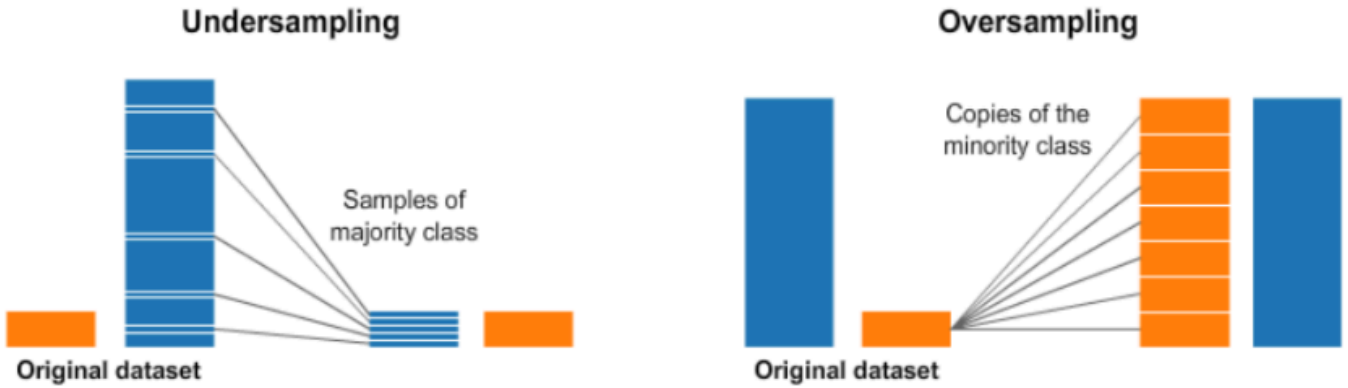
데이터를 조정해서 불균형 데이터를 해결하는 샘플링 기법들

1) 언더 샘플링

- 다수 범주의 데이터를 소수 범주의 데이터 수에 맞게 줄이는 샘플링 방식
  - 불균형한 데이터 셋에서 높은 비율을 차지하던 클래스의 데이터 수를 줄임으로써 데이터 불균형을 해소하는 것
  - 이 방법은 학습에 사용되는 전체 데이터 수를 급격하게 감소시켜 오히려 성능이 떨어질 수 있음

2) 오버 샘플링

- 소수 범주의 데이터를 다수 범주의 데이터 수에 맞게 늘리는 샘플링 방식
  - 낮은 비율 클래스의 데이터 수를 늘림으로써 데이터 불균형을 해소하는 것
  - 언더 샘플링보다 훨씬 좋은 해결책이 될 수 있을것 같은데, 문제는 "어떻게" 없던 데이터를 생성하느냐 임



```
#https://www.geeksforgeeks.org/python-random-sample-function/
#Simple implementation of sample() function.
# the use of sample() function .
# import random
from random import sample

# Prints list of random items of given length
list1 = [1, 2, 3, 4, 5]
print(sample(list1,3))
```

```
[2, 5, 1]
```

```
#https://rfriend.tistory.com/494
#[Python pandas] resample() 메소드로 시계열 데이터를 10분 단위 구간별로 집계/요약 하기
#먼저 '년-월-일 시간:분:초'로 이루어진 time-stamp를 index로 가지고,
#가격(price)와 수량(amount)의 두 개의 칼럼을 가지는 간단한 시계열 데이터를 만들어보겠습니다.
#pandas의 date_range(from, to, freq) 함수를 해서 '2분 간격(freq='2min')의 date range 데이터를 만들었습니다.
#이 중에서 20개 행만 선택해서 예를 들어보겠습니다.
import pandas as pd
import numpy as np
# generate time series index
range = pd.date_range('2019-12-19', '2019-12-20', freq='2min')
df = pd.DataFrame(index = range)[:20]
# add 'price' columm using random number
np.random.seed(seed=1004) # for reproducibility
```

```
df['price'] = np.random.randint(low=10, high=100, size=20)
# add 'amount' column unsing random number
df['amount'] = np.random.randint(low=1, high=5, size=20)
print('Shape of df DataFrame:', df.shape)
df
```

```
Shape of df DataFrame: (20, 2)
```

	price	amount
2019-12-19 00:00:00	12	4
2019-12-19 00:02:00	21	2
2019-12-19 00:04:00	41	1
2019-12-19 00:06:00	79	4
2019-12-19 00:08:00	61	2
2019-12-19 00:10:00	81	1
2019-12-19 00:12:00	24	3
2019-12-19 00:14:00	62	1
2019-12-19 00:16:00	76	3
2019-12-19 00:18:00	63	1
2019-12-19 00:20:00	95	2
2019-12-19 00:22:00	82	1
2019-12-19 00:24:00	82	3
2019-12-19 00:26:00	70	1
2019-12-19 00:28:00	30	4
2019-12-19 00:30:00	33	1
2019-12-19 00:32:00	22	2
2019-12-19 00:34:00	77	3
2019-12-19 00:36:00	58	3
2019-12-19 00:38:00	96	3

<<<참조자료 사이트>>>

1.AI 훈련용 빅데이터 2026년 고갈…문제점과 대책은
2."학습 데이터 부족" AI개발에 닥친 난관… 차세대 모델 개발 지연