



deeplearning.ai

Mismatched training and dev/test data

Training and testing on different distributions

딥러닝은 training set을 위한 데이터가 많이 필요하다. 때로는 dev/test set과 다른 distribution에서 온 데이터를 트레이닝 셋에 사용해야 할 때도 있을텐데 이때 도움이 되는 best practice를 알아보자.

Cat app example

Data from webpages



우리가 트레이닝 데이터셋에 사용할 수 있는 데이터는 아래 두 종류이고 각각 20만 : 1만 개 이미지로 구성된다. 이때 딜레마는 distribution이 다른 왼쪽 데이터를 쓸 것이지, 1만개의 적은 데이터를 쓸 것인가?

care about this

Data from mobile app



$\rightarrow \approx 200,000$

$210,000$
shuffle

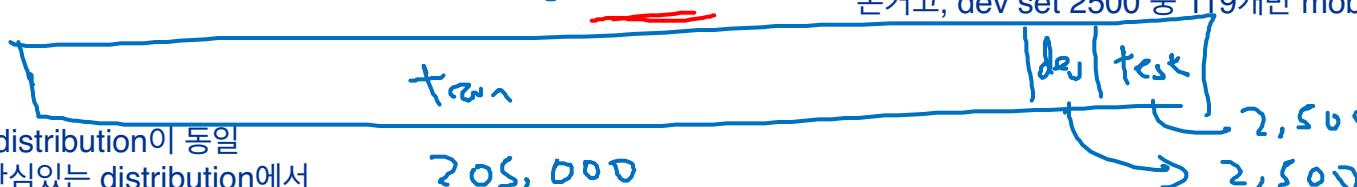
$\rightarrow \approx 10,000$

이렇게 하면 전체 21만 개 이미지 중에 20만 개가 web에서 온거고, dev set 2500 중 119개만 mobile에서 오게 된다.

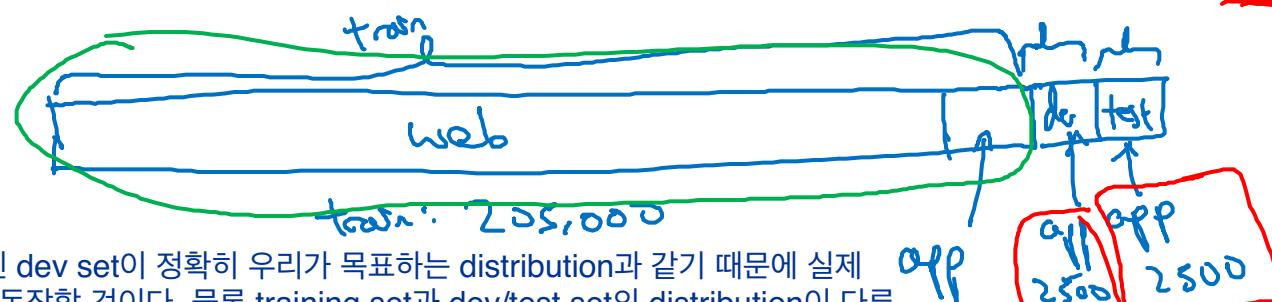
dev set을 세팅한다는건 목표로 하는 타겟이 어딘지 말해주는것이기 때문에 이렇게 되면 웹 이미지 distribution에 최적화하는데 많은 시간을 쓰게 된다.

X Option 1:

pros: training과 dev/test set의 distribution이 동일
cons: dev set의 대부분이 실제 관심있는 distribution에서 온 것이 아니게 된다.



O Option 2:



옵션 2의 경우엔 dev set이 정확히 우리가 목표하는 distribution과 같기 때문에 실제 상황에서 더 잘 동작할 것이다. 물론 training set과 dev/test set의 distribution이 다른 문제가 생기지만 이 방식이 장기적으로는 더 나은 성능을 보인다는 것이 밝혀졌다.



Andrew Ng

백미러야! 가까운 주유소 가는 길을 알려줘! 하면 네비게이션 정보 나오는 백미러 예제

Speech recognition example

Speech activated
rearview mirror



Training

일반 데이터

Purchased data $\downarrow \downarrow$
 x, y

Smart speaker control

Voice keyboard

...

500,000 utterances

Dev/test

일반 데이터와는 달리 주소같은게 많이 등장하므로 distribution이 다를 것이 자명하다.

Speech activated
rearview mirror

$\rightarrow 20,000$

train
500 K

10K SK SK D T

slot
510K

10K mirror

0 T
SK SK