deeplearning.ai

Mismatched training and dev/test data

Training and testing on different distributions

Cat app example

General Approach, just get data from Anywhere & shove it into traning

Data from webpages

≈ 200,000 Images

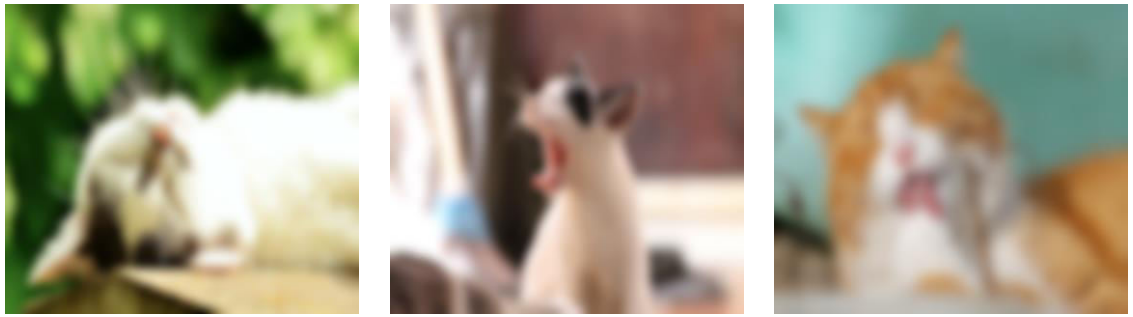
Dist of training is diff than dist of test

① option 1: combine all Images (210,000)

Disadv: dev/Test set has too many Images that are not of the "true" test set dist, ie, mobile uploads

out of 2.5K, $\frac{200K}{210K} \times 2.5K = 2381$ will be high quality non real test Images

Data from mobile app



≈ 10,000 Images

(You care about this, doesnt matter if you're trained on high quality Images, if you cant perform in real world, your App fails)

② option 2:

- Train = 200K web + 5K app = 205K
- Atleast this way you're optimizing for the correct dev set Images
- Disadv: Train dist is diff from dev/test

Go w/ this split

Andrew Ng

Speech recognition example

Alexa located in the rear view mirror of your car

Training

Purchased data

Smart speaker control (Alexa)

Voice keyboard

- Say you collect 500K samples from all training example sources

...

Dev/test

Speech activated rearview mirror

20K samples

∴ the split could look like

500K + 10K	5K	5K
Train	dev	Test

The problem is that none of the trained examples will be specific to the context of the car → Alexa, drive me to the nearest gas station / Alexa, open the trunk of the car

⇒ train is diff from dev/test

Andrew Ng