deeplearning.ai

Mismatched training and dev/test data

Addressing data mismatch

You've done your "Error Analysis" & see most of it is because of data mismatch

ie, dist(Train) ≠ dist(dev/Test)

So what can we do?

Addressing data mismatch

- Carry out manual error analysis to try to understand difference between training and dev/test sets

eg car Rearview mirror Assistant
= one way data differs → car samples are noisy [car noises in speech samples]
↳ traffic
↳ inside/Engine noise
↳ Honk
compared to training data

- Make training data more similar; or collect more data similar to dev/test sets & put this in training set

Artificial data synthesis

+

=

"The quick brown fox jumps over the lazy dog."

Clean Audio

Car noise

Synthesized in-car audio

Caveat

↳ You have 10K hours of speech (clean, noiseless speech)

↳ You have 1 hour of car noise

⇒ You generate all synthesised car Audio speech using this 1 hour car noise, Repeated over 10K hours

⇒ overfitting on that 1 hour of car noise

Andrew Ng

Artificial data synthesis

Car recognition: For Self Driving car, You need to Recognize nearby cars



- Synthesize car Images instead of taking 1M pictures of cars on the street, why not?