deeplearning.ai

Mismatched training and dev/test data

Addressing data mismatch

Addressing data mismatch

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

E.g. noisy – car noise   street numbers

→ • Make training data more similar; or collect more data similar to dev/test sets

E.g. Simulate noisy in-car data

Andrew Ng

Artificial data synthesis

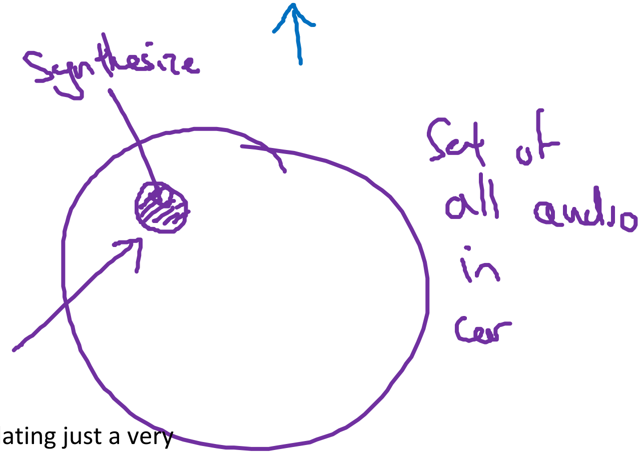
"The quick brown fox jumps over the lazy dog." + Car noise = Synthesized in-car audio

10,000 hours

1 hour of car noise

Overfit to 1 hour of car noise

10,000 hours



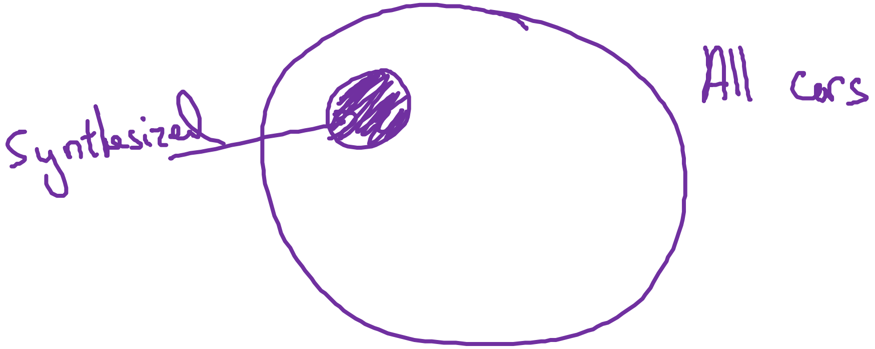
you might be simulating just a very small sunset of this space

Andrew Ng

Artificial data synthesis

Car recognition:

≈ 20 cars



But if you're using artificial data synthesis, just be cautious and bear in mind whether or not you might be accidentally simulating data only from a tiny subset of the space of all possible examples

Andrew Ng