Look at dev examples to evaluate ideas

90% accuracy
→ 10% error

Should you try to make your cat classifier do better on dogs?

Error analysis: → 5-10 min

- Get ~100 mislabeled dev set examples. "ceiling"
- Count up how many are dogs.

→ 5% → 50%

5/100 50/100

10% → 9.5% Goes down by 5% error

10% → 5%

Andrew Ng

Evaluate multiple ideas in parallel

Ideas for cat detection:

- Fix pictures of dogs being recognized as cats
- Fix great cats (lions, panthers, etc..) being misrecognized
- Improve performance on blurry images

Image	Dog	Great Cats	Blurry	Instagram	Comments
1	✓			✓	Pitbull
2			✓	✓	
3		✓	✓		Rainy day at zoo
⋮	⋮	⋮	⋮		
% of total	8%	43%	61%	12%	

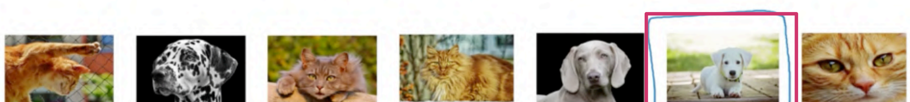
At most improve by 8% or 12%

Misclassifications

Andrew Ng

Incorrectly labeled examples

e.g.



x							
y	1	0	1	1	0	1	1

Training set

DL algorithms are quite robust to random errors in the training set.

Systematic errors

Andrew Ng

Error analysis

Image	Dog	Great Cat	Blurry	Incorrectly labeled	Comments
...					
98				✓	Labeler missed cat in background
99		✓			
100				✓	Drawing of a cat; Not a real cat.
% of total	8%	43%	61%	6%	

Overall dev set error 10%

Errors due incorrect labels .... 0.6% ← $\frac{.06}{1} = 6\%$

Errors due to other causes .... 9.4% ←

2%

0.6% $\frac{.06}{.02} = 30\%$ of labels

1.4%

A 2.1% B 1.9%

Goal of dev set is to help you select between two classifiers A & B.

Andrew Ng

Correcting incorrect dev/test set examples

Manually

- Apply same process to your dev and test sets to make sure they continue to come from the same distribution
- Consider examining examples your algorithm got right as well as ones it got wrong.
  98% right, 2% wrong
- Train and dev/test data may now come from slightly different distributions.

Andrew Ng

Speech recognition example

- Noisy background
  - Café noise
  - Car noise
- Accent...
- Far fro...
- Young...
- Stutte...
- ...

Guideline:

Build your first system quickly, then iterate

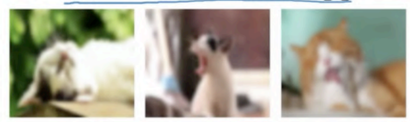
- Set up dev/test set and metric
- Build initial system quickly
- Use Bias/Variance analysis & Error analysis to prioritize next steps.

Andrew Ng

Cat app example

care about this

Data from webpages | Data from mobile app



→ ≈ 200,000 → 210,000 ← → ≈ 10,000

↓ shuffle   No!

X Option 1: train | dev | test

205,000   2,500   2,500

200K / 210K

2381 – web
119 – mobile app

Option 2: web (train) | dev | test

train: 205,000   app 5000   app 2500   app 2500

Andrew Ng

Speech recognition example

Speech activated rearview mirror

Training	Dev/test
Purchased data ↓x, y↓	Speech activated rearview mirror
Smart speaker control	→ 20,000
Voice keyboard	
…	

500,000 utterances

train 500K | D | T

10K   5K   5K

510K | D | T

10K mirror   5K   5K

Andrew Ng

If your dev data cam
from the same
distribution as your
training set, you wo
say that here you ha
large variance probl
that your algorithm'
not generalizing wel
the training set whic
doing well on to the
set, which it's sudde
doing much worse c
But in the setting wl
your training data ai
your dev data come
from a different
distribution, you car
longer safely draw t
conclusion. In partic
maybe it's doing jus
on the dev set, it's ju
that the training set
really easy because
was high res, very c
images, and maybe
dev set is just much
harder.

So maybe there isn'
variance problem ar
just reflects that the
set contains images
are much more diffi
classify accurately.

Cat classifier example

Assume humans get ≈ 0% error.

Training error 1%
Dev error .... 10%   ↓ 9%

Training-dev set: Same distribution as training set, but not used for training

→ train   training-dev   → dev   test

NN saw this

Training error	1%	↕ variance	1%	
→ Training-dev error	9%		1.5%	↕ data mismatch
→ Dev error	10%		10%	
	Variance			

Human error	0%	↕ Avoidable bias	0%	↕ Avoidable bias
Training error	10%		10%	↕ Variance
Training-dev error	11%		11%	↕ Data mismatch
Dev error	12%		20%	
	Bias		Bias + Data mismatch	

you can conclude from
that when you went
raining data to training
ata the error really
ıp a lot. And only the
nce between the
g data and the
g-dev data is that
eural network got to
e first part of this. It
ained explicitly on
ut it wasn't trained
itly on the training-dev
So this tells you that
ave a variance
m.

ıse the training-dev
vas measured on data
omes from the same
ution as your training
you know that even
h your neural network
vell in a training set,
t not generalizing well
a in the training-dev
ich comes from the
distribution, but it's
t generalizing well to
rom the same
ution that it hadn't
efore.

his example we have
a variance problem.

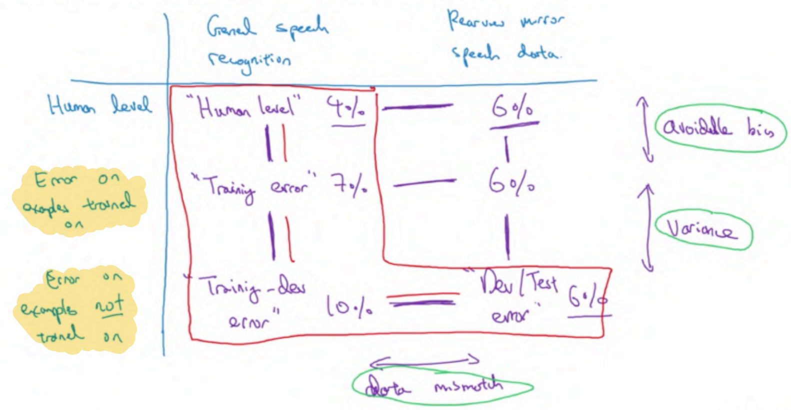
Andrew Ng

Bias/variance on mismatched training and dev/test sets

Human level	4%	avoidable bias	4%
Training set error	7%	variance	7%
Training-dev set error	10%	data mismatch	10%
→ Dev error	12%	degree of overfitting to dev set.	6%
→ Test error	12%		6%

More general formulation

Rearview mirror



Addressing data mismatch

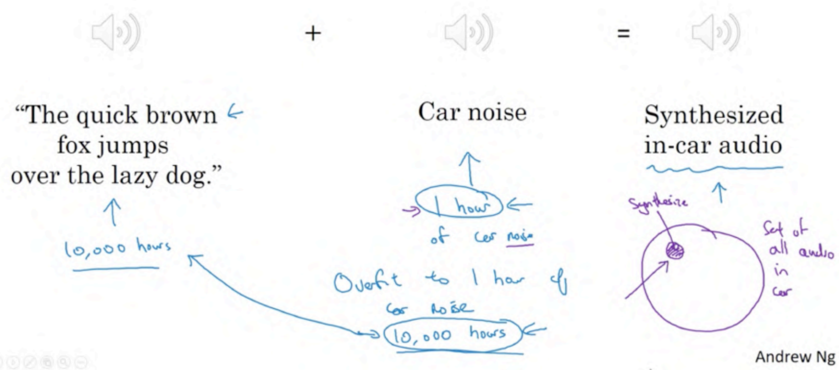
- → Carry out manual error analysis to try to understand difference between training and dev/test sets

  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

Artificial data synthesis

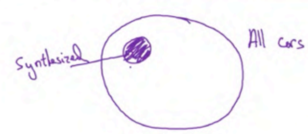


Andrew Ng

Artificial data synthesis

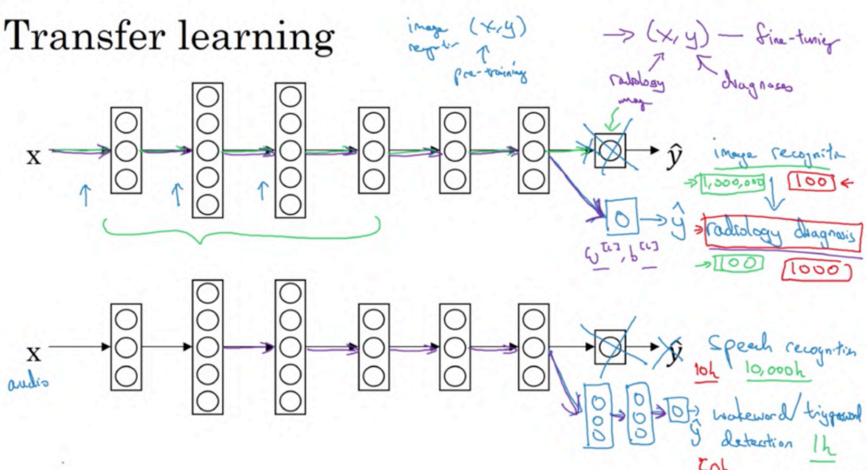
Car recognition:

≈ 20 cars



Andrew Ng

Transfer learning



Andrew Ng

When transfer learning makes sense

Transfer from $A \rightarrow B$

- Task A and B have the same input x.
- You have a lot more data for Task A than Task B.
- Low level features from A could be helpful for learning B.

Andrew Ng

Simplified autonomous driving example

$x^{(i)}$

	$y^{(i)}$
Pedestrians	0
Cars	1
Stop signs	1
Traffic lights	0
⋮	⋮

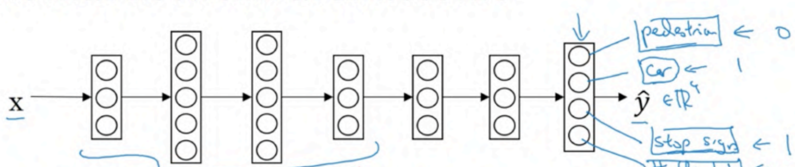
$(4,1)$

$$Y = \begin{bmatrix} | & | & | & & | \\ y^{(1)} & y^{(2)} & y^{(3)}, & \cdots, & y^{(m)} \\ | & | & | & & | \end{bmatrix}$$

$(4, m)$

Andrew Ng

Neural network architecture



Loss: $\underset{(4,1)}{\hat{y}^{(i)}}$ $\rightarrow \frac{1}{m} \sum_{i=1}^{m} \boxed{\sum_{j=1}^{4}} \mathcal{L}(\hat{y}_j^{(i)}, y_j^{(i)})$

Sum only over value of j with 0/1 label.

↳ Usual logistic loss

$-y_j^{(i)} \log \hat{y}_j^{(i)} - (1-y_j^{(i)}) \log (1-\hat{y}_j^{(i)})$

Unlike softmax regression:
One image can have multiple labels

Multi-task learning ⇐

$$Y = \begin{bmatrix} 1 & \boxed{1} & & \boxed{0} & & ? & \\ 0 & \boxed{1} & \cdots & \boxed{1} & \cdots & 1 & \cdots \\ \boxed{?} & ? & & \boxed{1} & & ? & \\ ? & ? & & \boxed{0} & & ? & \end{bmatrix}$$

Andrew Ng

When multi-task learning makes sense

- Training on a set of tasks that could benefit from having shared lower-level features.
- Usually: Amount of data you have for each task is quite similar.

A 1,000,000 → B 1,000

A_1 1,000
A_2 1,000
⋮
A_{100} 1,000

99,000

- Can train a big enough neural network to do well on all the tasks.

Andrew Ng

What is end-to-end learning?

Speech recognition example

"cat"

x → audio →(MFCC) features →(ML) Phonemes → words → transcript y

audio → phonemes → ... → transcript

audio → transcript

mapping x to y with no intermediate steps

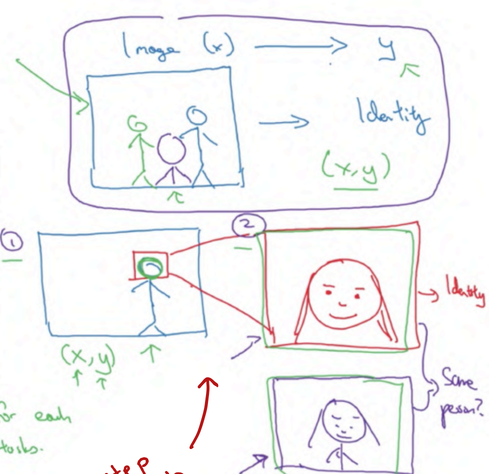
3,000h

10,000h
⋮
100,000h

Andrew Ng

Face recognition

[Image courtesy of Baidu]



Have data for each of 2 subtasks.

Multi step approach

Andrew Ng

More examples

Machine translation

(x, y)

Eglish ↗ ↖ French

English → text analysis → ... → French

English ————————————→ French

Estimating child's age:

Image →① bones →② age

Image ——————→ age

Andrew Ng

Pros and cons of end-to-end deep learning

Pros:

- Let the data speak

$x \rightarrow y$

→ "phonemes"

c a t

- Less hand-designing of components needed

Cons:

$x - - \quad - - \rightarrow y$

input end ↓ output end ↓

$x \rightarrow y$ (x, y)

- May need large amount of data
- Excludes potentially useful hand-designed components

Data. Hand-design.

Andrew Ng

Applying end-to-end deep learning

Key question: Do you have sufficient data to learn a function of the complexity needed to map x to y?

$x \rightarrow y$

→ age

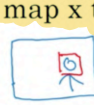


image →DL→ cars ————→ route → steering

radar lidar →DL→ pedestrians — motion planning — control

- Use DL to learn individual components
- Carefully choose $X \rightarrow Y$ depending what tasks you can get data for.

image ———————————→ steering

Andrew Ng