Setting up your goal

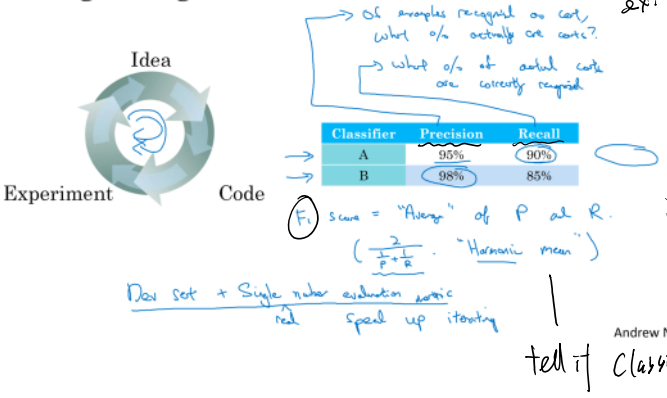
Tuesday, June 5, 2018 3:21 PM

singlenum

set up a number that can evaluate performance

Using a single number evaluation metric

ex: cat classifier



Single # evaluation metric.

problem: better precision, worse recall, or reverse.

例 $F_1 = \frac{2}{\frac{1}{P}+\frac{1}{R}}$

tell if Classifier A or B is better — speed up iterating.

Another example

Algorithm	US	China	India	Other
A	3%	7%	5%	9%
B	5%	6%	5%	10%
C	2%	3%	4%	5%
D	5%	8%	7%	2%
E	4%	5%	2%	4%
F	7%	11%	8%	12%

Andrew Ng

What if the algorithms have different performance (ranking) in different test sets (US, China, India, Other)?

Take average

satisficing

deeplearning.ai

Setting up your goal

Satisficing and optimizing metrics

Another cat classification example

Classifier	Accuracy	Running time
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

optimizing: Accuracy; satisficing: Running time

cost = accuracy − 0.5 × running time

maximize accuracy
subject to running time ≤ 100 ms.

N metrics: 1 optimizing
N-1 satisficing

voice
Wakewords / Trigger words
Alexa, OK Google,
Hey Siri, nihaobaidu
你好百度

accuracy.
#false positive

maximize accuracy.
s.t. ≤ 1 false positive
every 24 hours.

Andrew Ng

有不同的 criteria : Accuracy, run time
what to do ? Define cost function, then set constraint
如 — run time 不能超过 xx 时, 找 max accuracy.

traindevtest

deeplearning.ai

Setting up your goal

Train/dev/test distributions

Cat classification dev/test sets

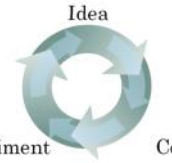
Regions:

- US
- UK
- Other Europe
- South America

Dev

- India
- China
- Other Asia
- Australia

Test



Andrew Ng

Dev and test sets should come from the same distribution

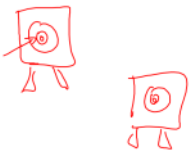
otherwise ML learn models that do not get good performance on the test set (WRONG TARGET)

True story (details changed)

A real world example

Optimizing on dev set on loan approvals for medium income zip codes

Tested on low income zip codes



Andrew Ng

Guideline

Same distribution

Choose a dev set and test set to reflect data you expect to get in the future and consider important to do well on.

Andrew Ng

Setting up your goal

Size of dev and test sets

rule of thumb.

Reasonable for smaller size of data.

Size of test set

→ Set your test set to be big enough to give high confidence in the overall performance of your system.

10,000
100,000

Train + Dev. No test set.

train	~~test~~ dev

can do train + ~~test~~ dev. with No test set

⇒ Not recommended

Take away: the 70/30 rule ~~NOT~~ no longer work

trend: assignment ~~[illegible]~~ more data to training set.

Andrew Ng

changeset

deeplearning.ai

Setting up your goal

When to change dev/test sets and metrics

If your evaluation matrix don't give you results that make sense

⇒ use a different matrix

Criteria

还是取决根据需求定义

Cat dataset examples

Metric + Dev : Prefer A
You/users : Prefer B.

→ Metric: classification error

Algorithm A: 3% error → + pornographic → 不过为啥把两个给混起来?

✓ Algorithm B: 5% error

$$\text{Error}: \frac{1}{\sum_i w^{(i)}} \cancel{\frac{1}{m_{dev}}} \sum_{i=1}^{m_{dev}} w^{(i)} \, I\{ y_{pred}^{(i)} \neq y^{(i)} \}$$

y_{pred} ← predicted value (0/1)

$$\rightarrow w^{(i)} = \begin{cases} 1 & \text{if } x^{(i)} \text{ is non-porn} \\ 10 & \text{if } x^{(i)} \text{ is porn} \end{cases}$$

penalize 把porn misclassify 成 cat 的行为.

Andrew Ng

Orthogonalization for cat pictures: anti-porn

1. So far we've only discussed how to define a metric to evaluate classifiers. ← Place target
2. Worry separately about how to do well on this metric.

Aim/shoot at target

$$J = \frac{1}{\sum w^{(i)}} \sum_{i=1}^{m} w^{(i)} \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right)$$

Andrew Ng

1. Define metrics
2. Optimize on this matrix.

Another example

Algorithm A: 3% error
Algorithm B: 5% error ←

→ Dev/test

→ User images

If doing well on your metric + dev/test set does not correspond to doing well on your application, change your metric and/or dev/test set.

Andrew Ng

WRONG Target.
or. correct target, ill-defined in the evaluation metrics

Even you can't decide: set up quickly some evaluation matrix. try => adjust the metrics, => train model again

compare

Comparing to human-level performance

deeplearning.ai

Why human-level performance?

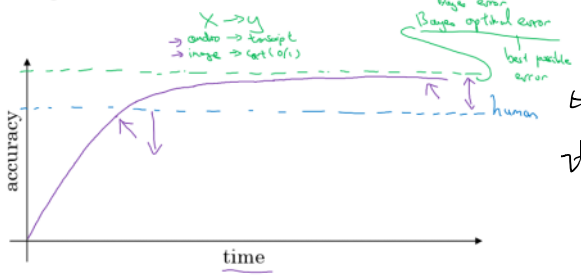
Comparing to human-level performance

$X \to y$
audio → transcript
image → cat (0/1)

Bayes error
Bayes optimal error

Bayes Optimal error: the Best possible error
e.g. Some audio is just too noisy; Some images are too blurry
progresses slow down after surpasses human performance

Comparing to human-level performance



Andrew Ng

Why compare to human-level performance

Humans are quite good at a lot of tasks. So long as ML is worse than humans, you can:

- Get labeled data from humans.
- Gain insight from manual error analysis: Why did a person get this right?
- Better analysis of bias/variance.

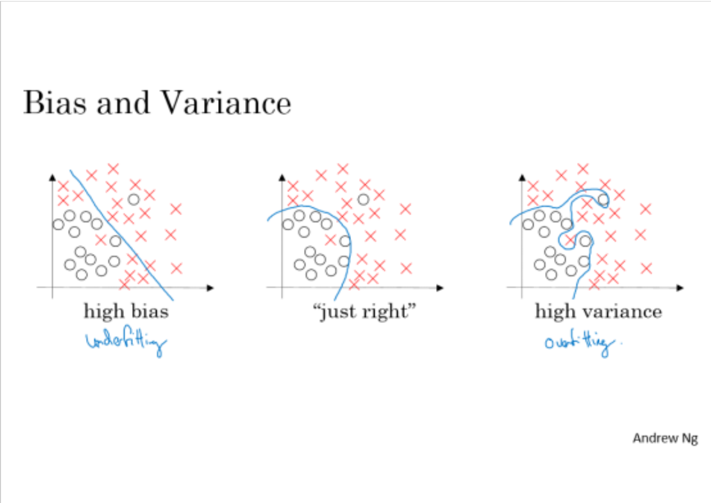
Andrew Ng

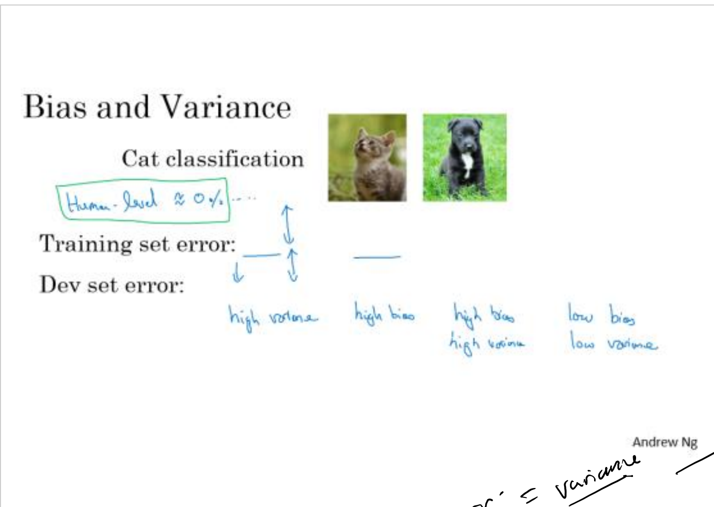
avoidablebias

deeplearning.ai

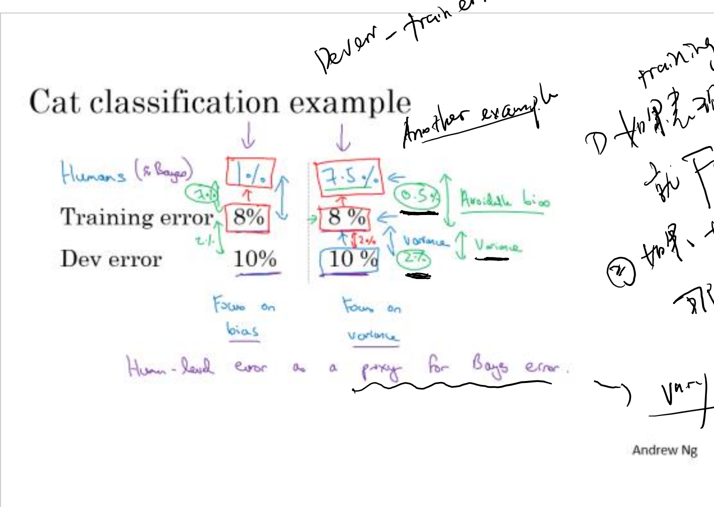
Comparing to human-level performance

Avoidable bias





Dev err - train err = variance → 这里的 variance 最好说明了...



training error
① 如果和 human 差距过大，
就 Focus on reducing bias.
② 如果 training error 接近 human error 了，
那就要 focus on ↓ variance (dev test error ↓
regularization, ...

→ vary across tasks

humanlevelperfom

deeplearning.ai

Comparing to human-level performance

Understanding human-level performance

Human-level error as a proxy for Bayes error

Medical image classification example:

Suppose:

(a) Typical human ..... 3 % error

(b) Typical doctor ..... 1 % error

(c) Experienced doctor ..... 0.7 % error

(d) Team of experienced doctors .. 0.5 % error

What is "human-level" error?

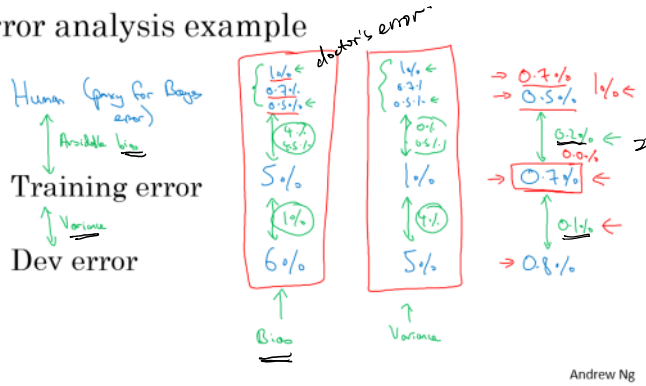
How to define

Andrew Ng

It is a matter of what you choose as the benchmark.

define human-level error as a proxy for Bayes error.

Error analysis example



Andrew Ng

What if all bad?

i.e. can't improve training error any more

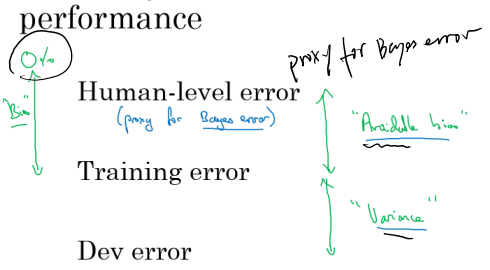
≠ variance ⇒ avoidable bias is a bigger problem

比較 training set performance 和 human level performance

test set performance

可能決定 Focus 在 bias 還是 variance reduction.

Summary of bias/variance with human-level performance



Instead of comparing { with 0% — bias
with human-performance — avoidable bias.

Human-level performance ⇒ Bayes Error ⇒
Decide whether to focus on { bias reduction / variance reduction

Noisy data: Having a better estimate of Bayes error helps knowing what the "avoidable error" is

Andrew Ng

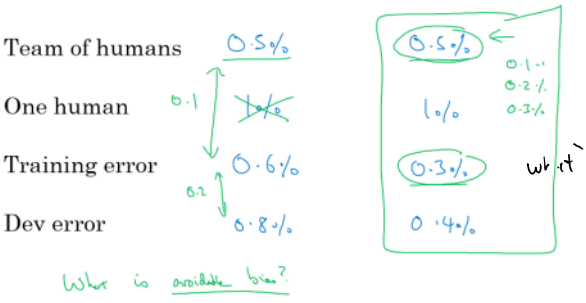
surpass human

Comparing to human-level performance

Surpassing human-level performance

deeplearning.ai

Surpassing human-level performance



What is avoidable bias?

what it means? ① over-fitting
② the Bayes error is actually lower.

⇓
That slows down the speed to make progress

Andrew Ng

Problems where ML significantly surpasses human-level performance

- Online advertising
- Product recommendations
- Logistics (predicting transit time)
- Loan approvals

Structured data
Not natural perception
lots of data

- Speech recognition
- Some image recognition
- Medical
  - ECG, Skin cancer, ...

↓

in some cases, can surpass human-level performance

Andrew Ng

ML 比 human 更好的 2 种情况

what these share in common

Structured data

e.g. what you bought.
your loan record.

不是 natural perception problems (?)

improve model

Comparing to human-level performance

deeplearning.ai

Improving your model performance

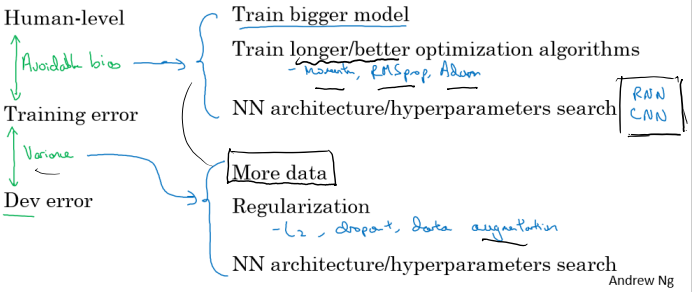
The two fundamental assumptions of supervised learning

low avoidable bias

1. You can fit the training set pretty well.
   ≈ Avoidable bias
2. The training set performance generalizes pretty well to the dev/test set.
   ≈ Variance not bad

Andrew Ng

Reducing (avoidable) bias and variance



Andrew Ng

error analysis

deeplearning.ai

Error Analysis

Carrying out error analysis

Look at dev examples to evaluate ideas

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

Error analysis:

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

Andrew Ng

手工看数据.

example: cat classifier 把狗认成猫.

Focus on "dog pictures".

Q: 值得 worth doing it — focusing on dogs

What to do? Examine if dogs account for a large proportion of mislabeled pictures.

Ceiling: how much improvement can focusing on the dog problem help?

This won't take long

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

猫科动物.

OK.. 所以手工检验花时间是重要的

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

filtering

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

Andrew Ng

work on them seperately

Create a table for this

cleagning up

Error Analysis

Cleaning up Incorrectly labeled data

deeplearning.ai

Incorrectly labeled examples

mislabeled

x							
y	1	0	1	1	0	1	1

Training set

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

less robust to Systematic errors

eg. label all white dogs as cats

need not worry much

Andrew Ng

How about incorrectly labeled data in dev or test sets.

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%	2%
Errors due incorrect labels	0.6%	0.6%
Errors due to other causes	9.4%	1.4%

not a large problem for now.

model 的 error 相对还是大于 mislable %.

Overall dev set error	10%	2%
Errors due incorrect labels	0.6%	0.6%
Errors due to other causes	9.4%	1.4%
		2.1% 1.9%

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

Andrew Ng

can fix this 9.4%

not a large problem for now.
model 的 error 目前还远大于 mislable %.
所以主要精力还是model. 不是 correct labeling

Correcting incorrect dev/test set examples

- 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. (98.1) (2%)
- Train and dev/test data may now come from slightly different distributions.

Andrew Ng

if only fix those getting wrong - overconfidence on performance.

this is ok — there are tricks to deal with this

Training set is too large.
→ Sometimes you don't want to manually correct advice

Advice: Manual data analysis is important in ML practice
Sit down and look at the data.
good use of time — help to decide what direction to prioritize

build

deeplearning.ai

Error Analysis

Build your first system quickly, then iterate

lots of things to do...

Speech recognition example

- Noisy background
  - Café noise
  - Car noise
- Accent
- Far fro

Guideline:
Build your first

- Set up dev/test set and metric
- Build initial system quickly
- Use Bias/Variance

analysis helps to examine

- Car noise
- Accent
- Far fro
- Young
- Stutter
- …

Guideline:
Build your first system quickly, then iterate

- Build initial system quickly
- Use Bias/Variance analysis & Error analysis to prioritize next steps.

Andrew Ng

error analysis helps to examine where you got wrong

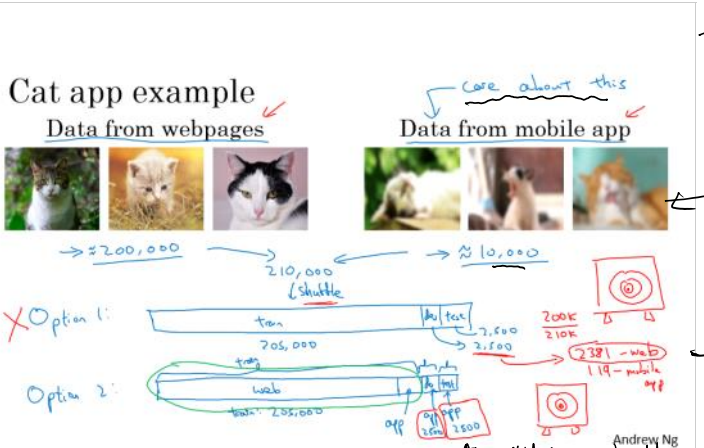
But if there's a literature telling you where to start, it helps to start with complex systems.

- Don't OVERTHINK when starting
- Build something that works (quick & dirty), then fix it using Bias/variance analysis & error analysis

diff distribution

Mismatched training and dev/test data

Training and testing on different distributions

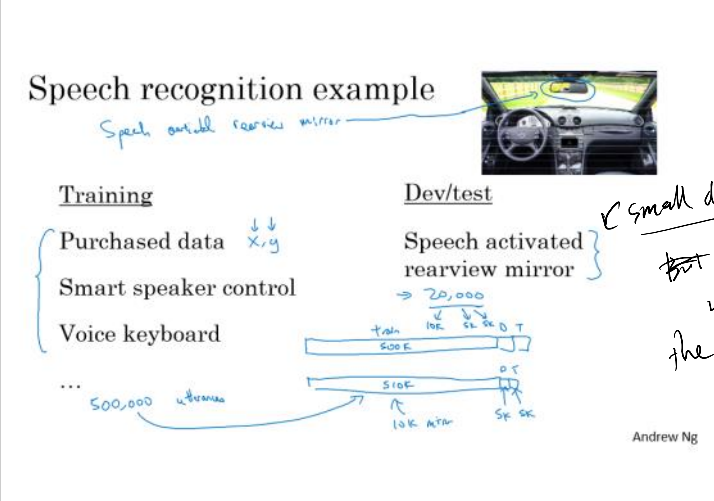


train model and test it on data from a different distribution.

← if use this only — small dataset.

— WRONG TARGET. optimize for a different dist of data from what you care about

☆ All test should be from the app (what you ultimately care about)



< small dataset

~~But~~ since this is what you want, the dev/test should consist of this only.

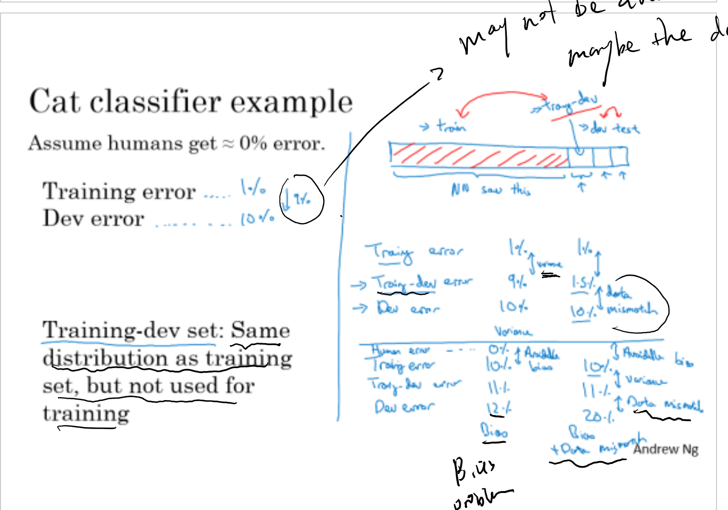
Q: should you always use all the data you have? Not really

mismatch

may not be a variance problem

maybe the dev set is just larger

Data mismatch

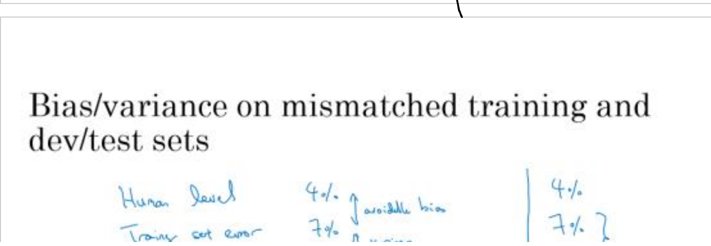


Bias problem

这像是个不错的打表办法.

利用一些和本任务弱相关的数据

检验模型的表现.



Bias/variance on mismatched training and dev/test sets

Human level	4%	4%
Training set error	7%	7%
Training-dev set error	10%	10%
→ Dev error	12%	6%
→ Test error	12%	6%

avoidable bias; variance; data mismatch; degree of overfitting to dev set.

Andrew Ng

More general formulation

	General speech recognition	Rearview mirror speech data
Human level	"Human level" 4%	6%
Error on examples trained on	"Training error" 7%	6%
Error on examples not trained on	"Training-dev error" 10%	"Dev/Test error" 6%

avoidable bias; variance; data mismatch

Andrew Ng

Avoidable Bias: Human-level performance ↕ training set performance

Variance: training set performance ↕ train-dev set performance

Data mismatch: train-dev set performance ↕ dev set performance

degree of overfitting to the dev set: dev set performance ↕ test set

if overfit ⇒ find a larger dev set.

What if you do better in dev/test set than on training set? e.g.

one possibility: dev/test set is easier (clearer photo)

More general formulation

General speech recognition

Human level 4%

Training error 7%

Training-dev set error 10%

Dev/test set error 6%

data mismatch

Rearview mirror speech recognition.

6%

6%

avoidable bias

variance

1. Human level.
2. Errors of examples trained on
3. Errors of example not trained on

How to deal with data mismatch? up next. — not much

addressmismatch

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

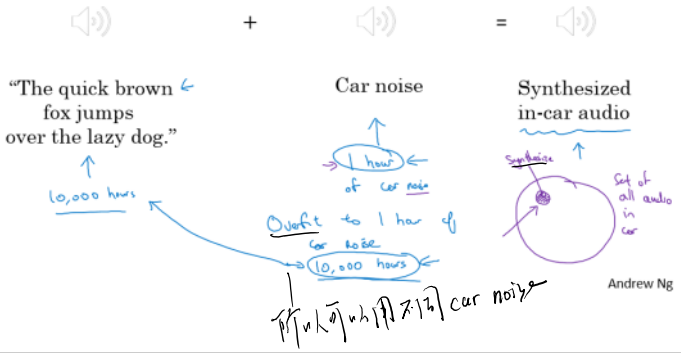
Key: manual error analysis

"How your dev/test sets are different from your training set?"

eg ~~too~~ too noisy

Artificial data synthesis

SYNTHESIS



加入同一种 car noise

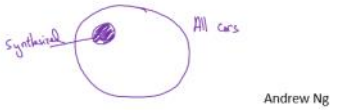
会导致 overfit to certain car noise.

Artificial data synthesis

Car recognition:



≈20 cars



Andrew Ng

recognize car

用 car video game 的截图.

门槛：竞赛 game 里面车的种类有限

Take away: ① Error analysis

② Artificial synthesis. — caution: overfit.

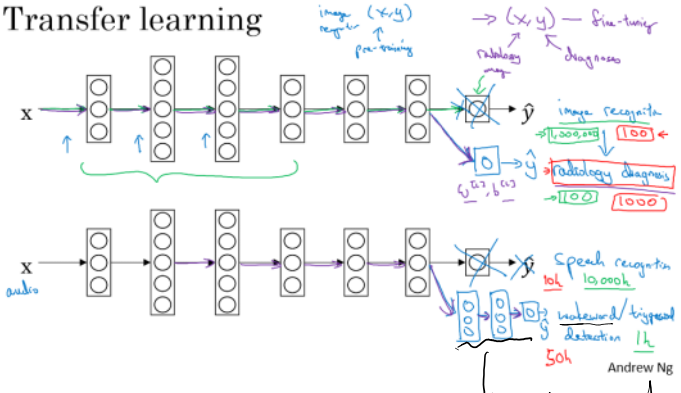
transfer learning

Learning from multiple tasks

Transfer learning

deeplearning.ai

Transfer learning



Andrew Ng

create several new layers.

transfer learned model to other task.

把最后一层重新 train

pre-training：用步们很多 data train 出来的部分

fine-tuning.

If you want to collect more data.

When transfer learning makes sense

Trade from A → 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

learn lower-level features

If you want to collect more data. Data of the task of interest B more important.

multitask

Learning from multiple tasks

Multi-task learning

deeplearning.ai

Simplified autonomous driving example

Detect multiple objects

$x^{(i)}$

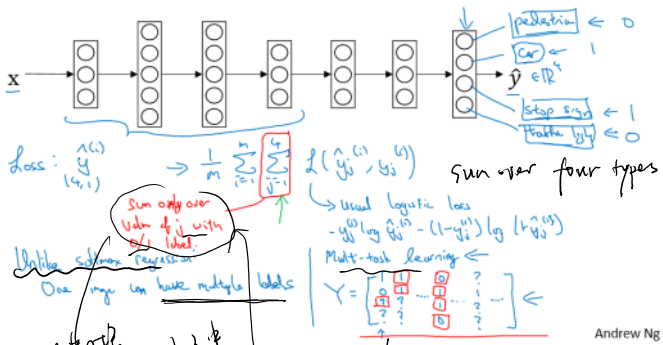
$y^{(i)}$ (4,1)

Pedestrians 0
Cars 1
Stop signs 1
Traffic lights 0

Outcome $= \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ y^{(1)} & y^{(2)} & y^{(3)} & \cdots & y^{(m)} \\ 1 & 1 & 1 & & 1 \end{bmatrix}$ (4, m)

Andrew Ng

Neural network architecture

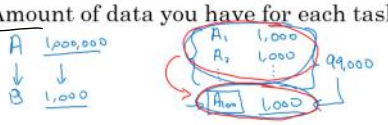


Sum over four types of object wanted

Building an NN that solves 4 problems could have done it by training 4 networks. But if some common lower-level features for 4 task, it will be helpful to train a multi-task learning

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.



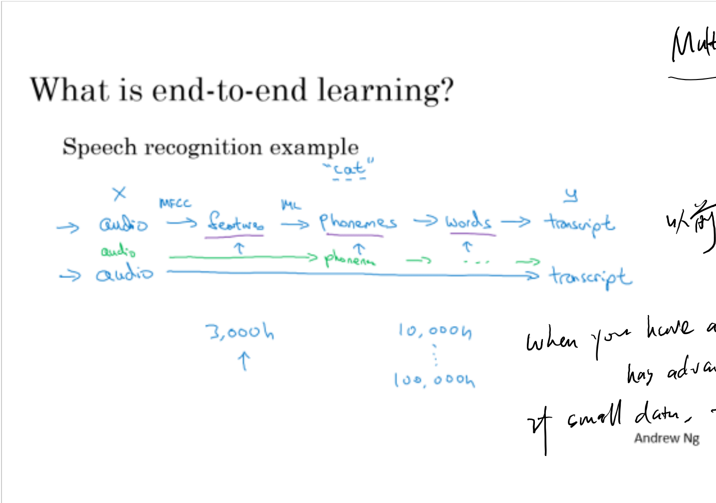
- Can train a big enough neural network to do well on all the tasks.

Andrew Ng

endtoend

End-to-end deep learning

What is end-to-end deep learning

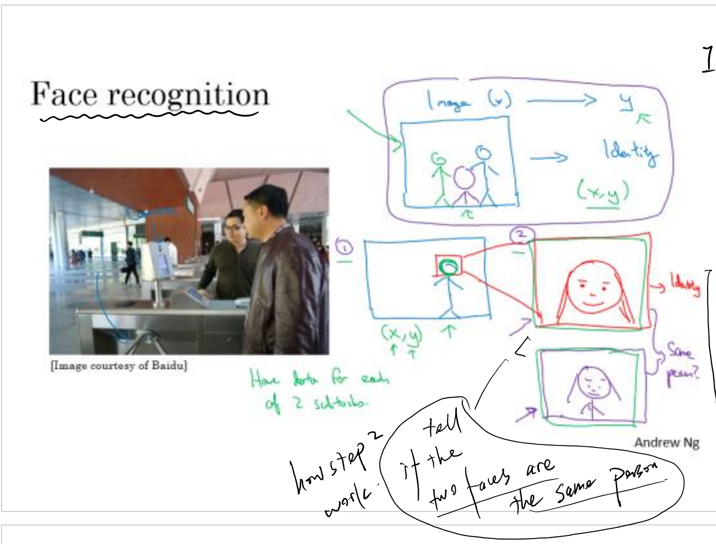


Multiple stages => one stage

以前break成不同stages的任务, 用DL一步完成

When you have a large dataset, the end-to-end approach has advantage.

If small data, traditional approach is better.



In this example, the two step approach works better?

① Two tasks are both simpler (much simpler). (更简单)

② lots of data for each ~~step~~ of the steps respectively (数据)

But much for the end-to-end tasks

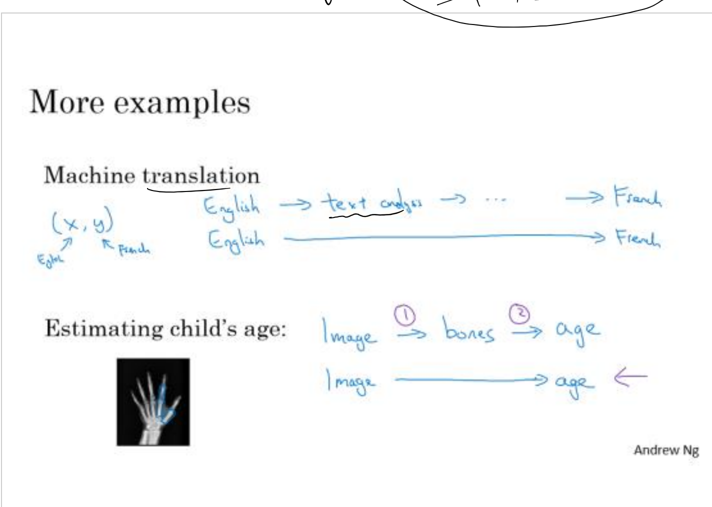
multi step

① Detect where the person's face is

② Crop image, center the person's face

③ Recognize the face

Two tasks — this



End-to-end DL.

works, but not a panacea.

endtoend2

deeplearning.ai

End-to-end deep learning

Whether to use end-to-end learning

End to end learning

Just let the data speak.

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

no need to reflect human pre-conception.
Some of these pre conception may be ~~err~~ wrong.

Cons:

- May need large amount of data $x \rightarrow y$ (x, y)
- Excludes potentially useful hand-designed components

Need LOTS of data

Data. Hand-design.

An algorithm has two sources of knowledge
[data → when small human knowledge is more important.
human design.

inject hand-designed knowledge

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?

$x \rightarrow y$

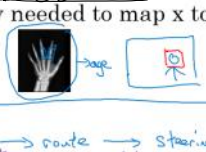


image → cars → route → steering
radar, lidar → pedestrians → (motion planning) → (control)

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

image → steering

Andrew Ng

for simple task → easily using end-to-end DL with small datasets

Example: Autonomous driving

for this task, end-to-end DL does not work well.