Application example : Photo OCR

reason : a complex ML system
concept of ML pipeline
about computer vision
artificial data synthesis

Problem description and Pipeline

The Photo OCR problem

OCR : optical character recognition

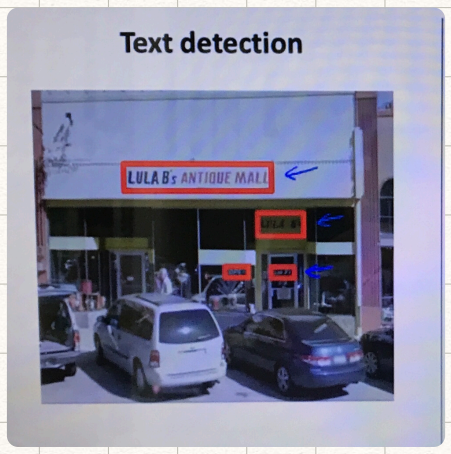
let PC 讀取到照片裡的文字

Photo OCR pipeline

0. Image
1. Text detection
2. Character segmentation (文字分割)
3. Character classification
4. 拼字校正

Sliding window (滑動窗) 就是常常看到的綠色框框

Text detection v.s. Pedestrain detection



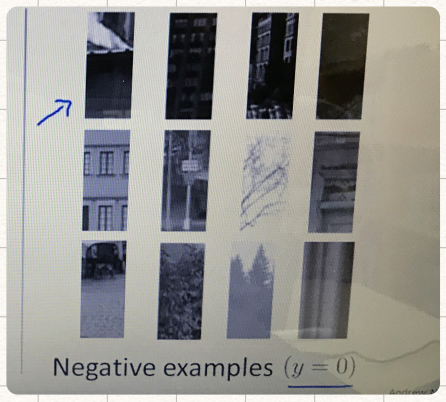
每個文字的長寬比不一樣(較難)

每個行人的長寬比理應相同(較簡單)

Supervised learning for pedestrian detection

先訓練分類器

x = pixels in 82x36 images patches



Sliding window detection

先以 82x36 的
滑動窗移動數個 step size，滑過整張圖片。
將讀到的圖像送入分類器

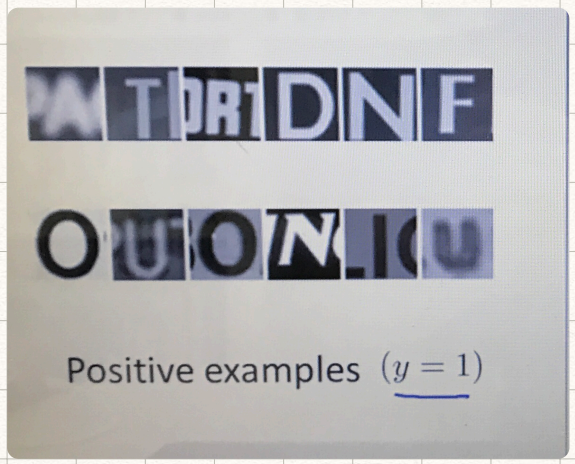
step size (stride parameter)

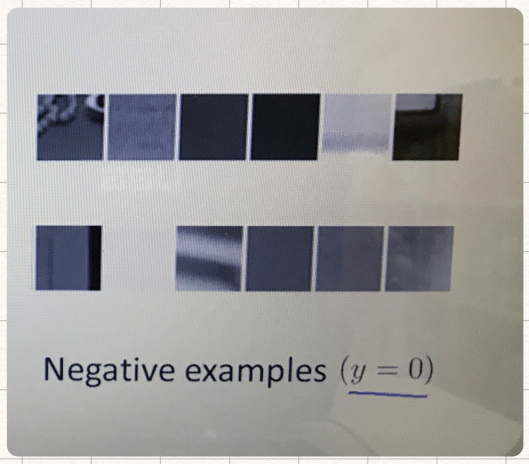
之後取更大的滑動窗，並將之壓縮至
82x36 的尺吋。

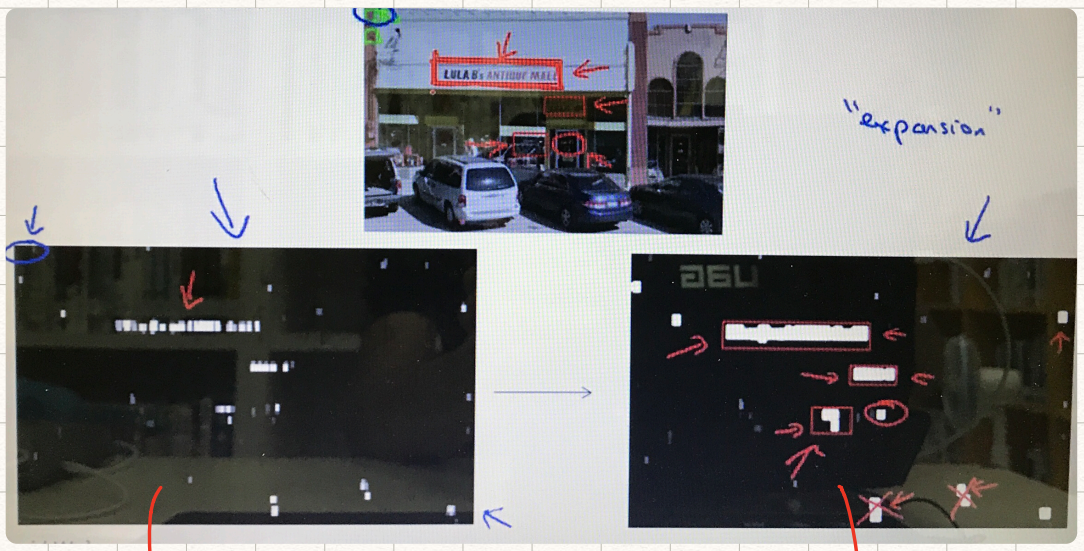
重複進行上述動作，便完成行人檢測

Text detection

同行人檢測，先訓練分類器





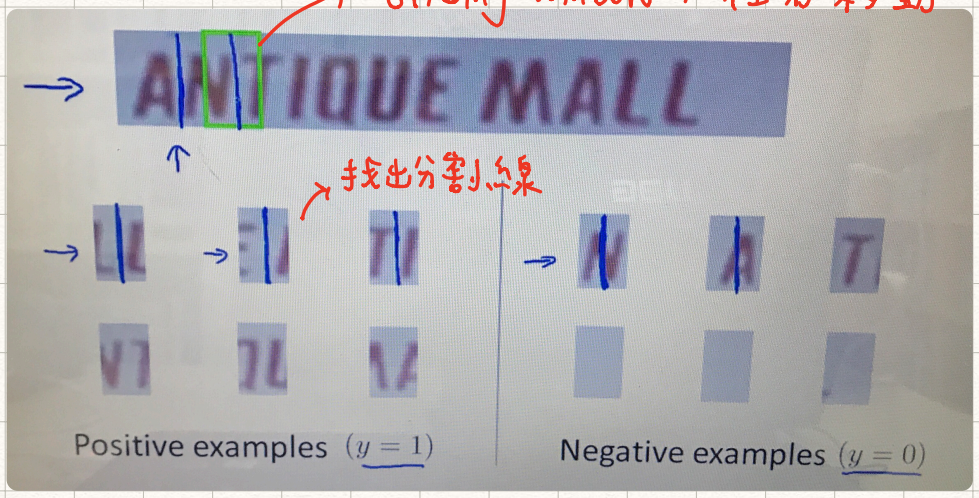


滑動窗滑出的結果

用擴展器擴展的結果

將左圖白色部分間的
像素點也變成白色。
爾後再淘汰高瘦的白框

1D sliding window for character segmentation



Getting Lots of Data and Artificial Data

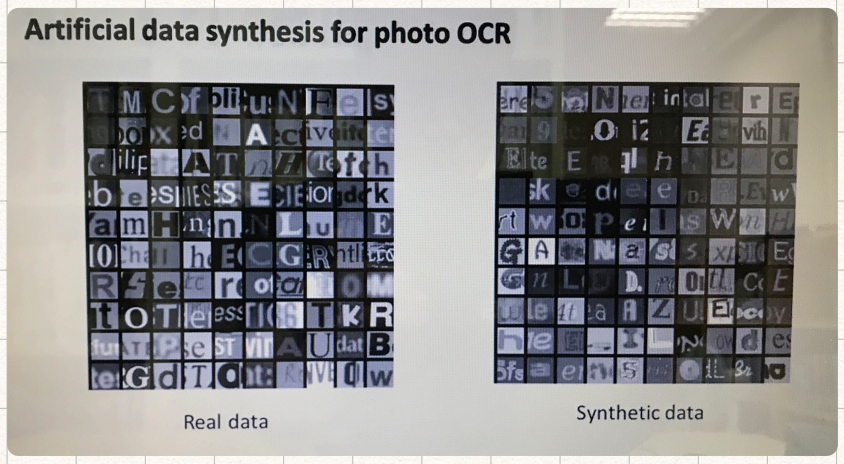
採用 low bias 演算法，並以大量 data 訓練之
⟶ 好的 ML system 🙂

此法用來創造大量 data，但並不適用所有 problem

two main variations:

1. 創造新的 data
2. 對已有的 data 進行放大

Artificial data synthesis for photo OCR



藉由此法，可創造近 ∞ 個變量

以不同的字體取代 Real data，再對其進行旋轉、模糊、裁切 ... 動作

Synthesizing data by introducing distortions

對原有資料進行不同程度的放大與扭曲

※ 注意，對 real data 的操作須合理
符合該 system 的特性

Distortion introduced should be repensentation of the type of noise / distortions in the test set.

Usually does not help to add purely random / meaningless noise to your data.

Synthesizing data by introducing distortions:

Speech recognition

對語音加入背景聲音，模擬差的電話通信

對原先乾淨的 data 加入躁音

Discussion on getting more data

1. Make sure you have a low bias classifier before expending the effort. (Plot learning curves) E.g. keep increasing the number of features/ number of hidden units in neural network until you have a low bias classifier.

2. "How much work would it be to get 10x as much data as we currently have?"

- Artificial data synthesis
- Collect / label it yourself
- "Crowd source"

  (E.g. Amazon Mechanical Turk)

  ↑

  人工標記數據集

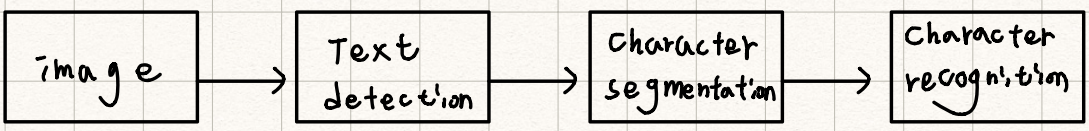
通常，收集資料並標記所花的時間少到令人驚訝

Ceiling Analysis: What part of the pipeline to work on next

(上限分析)

最寶貴的是時間

Estimating the errors due to each component (ceiling analysis)



What part of the pipeline should you spend the most time trying to improve?

Component	Accuracy
Overall system	72 %
Text detection	89 %
Character segmentation	90 %
Character recognition	100 %

將流水線各模塊逐至100%正確後，測試系統效能為何

↓ 17 %

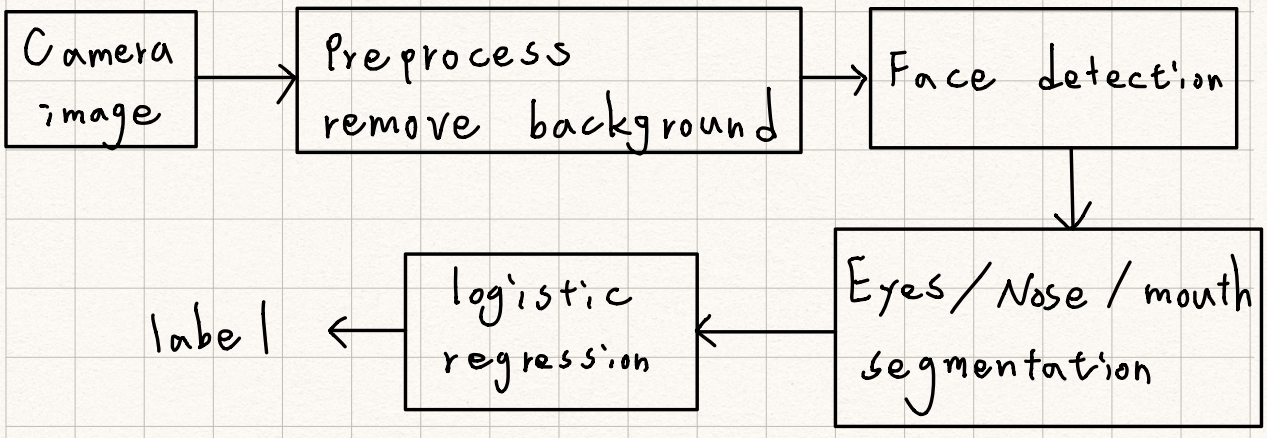
↓ 1 %

↓ 10 %

並且觀察增進的效能%

Another ceiling analysis example

Face recognition from images



Component	Accuracy		
Overall system	85 %		
Preprocess	85.1 %	↓	0.1 %
Face dection	91 %	↓	5.9 %
Eyes segmentation	95 %	↓	4 %
Nose segmentation	96 %	↓	1 %
Mouth segmentation	97 %	↓	1 %
Logistic regression	100 %	↓	3 %

花最多時間在改進空間最大的模塊!

ML Summary

- supervised learning

  Linear regression, logistic regression
  neural networks, SVMs

- Unsupervised learning

  K-means. PCA. Anomaly detection

- Special applications / special topics

  Recommender systems, large scale mL

- Advice on building a mL system

  Bias / variance, regularization, deciding what to work on next: evaluation of learning algoriths. learning curves, error analysis, ceiling analysis