機器學習模型的可解釋性 (Explainable ML)

Create at 2022/06/25

- 機器學習模型的可解釋性 (Explainable ML)
  - Local Explanation
  - Global Explanation

- 上課資源：
  1. 機器學習模型的可解釋性 (Explainable ML) (上) – 為什麼類神經網路可以正確分辨寶可夢和數碼寶貝呢？ (https://www.youtube.com/watch?v=WQY85vaQfTI)
  2. 機器學習模型的可解釋性 (Explainable ML) (下) –機器心中的貓長什麼樣子？ (https://www.youtube.com/watch?v=0ayIPqbdHYQ)

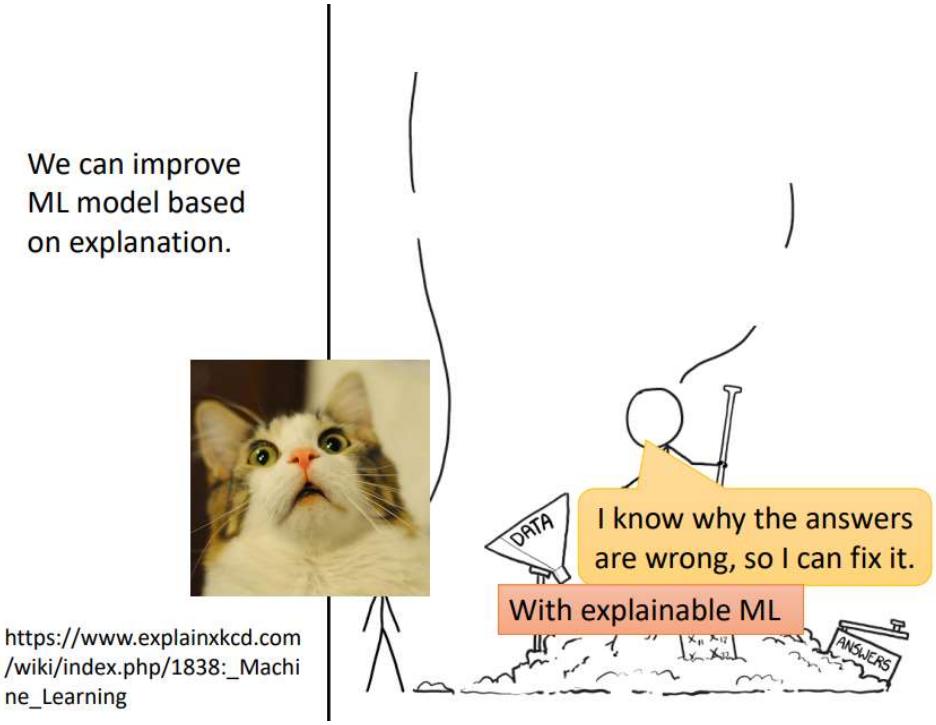
以前輸入一張圖片，得到一個答案，現在我們要機器得到答案的理由

- 為甚麼 Explainable 的 Machine Learning 是重要的議題？
  - 因為就算機器可以得到正確的答案，也不代表它一定非常聰明

Why we need Explainable ML?

- Loan issuers are required by law to explain their models.
- Medical diagnosis model is responsible for human life. Can it be a black box?
- If a model is used at the court, we must make sure the model behaves in a nondiscriminatory manner.
- If a self-driving car suddenly acts abnormally, we need to explain why.

- 在很多應用中 Explainable 的 Machine Learning 模型往往是必須的
- 機器必須給出得到答案的理由



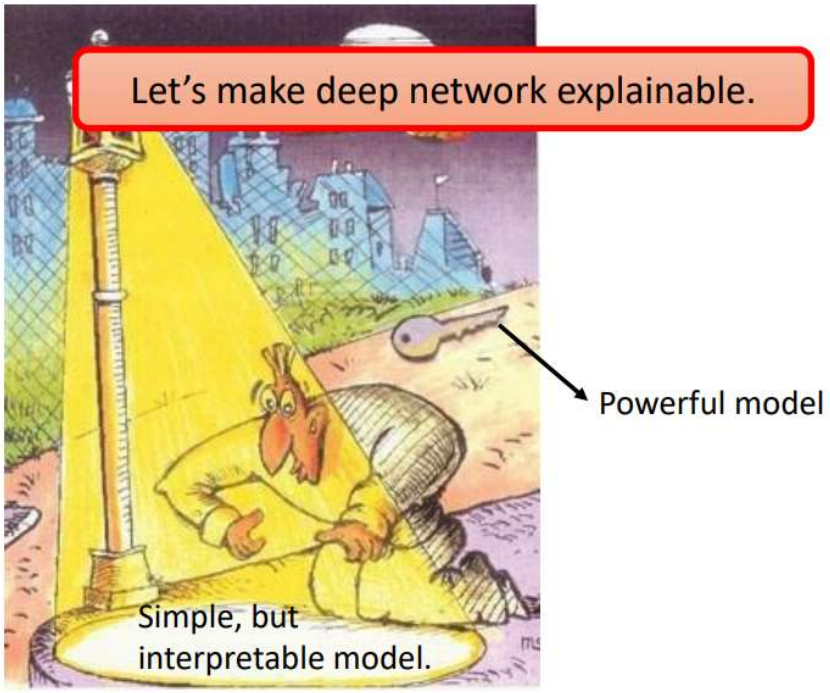
- 如果機器的模型具有解釋的能力的話，我們也許能憑藉解釋的結果去修正我們的模型

Interpretable v.s. Powerful

- Some models are intrinsically interpretable.
  - For example, linear model (from weights, you know the importance of features)
  - But not very powerful.
- Deep network is difficult to interpretable. Deep networks are black boxes … but powerful than a linear model.

We don't want to use a more powerful model because it is a black box.

This is "cut the feet to fit the shoes." (削足適履)



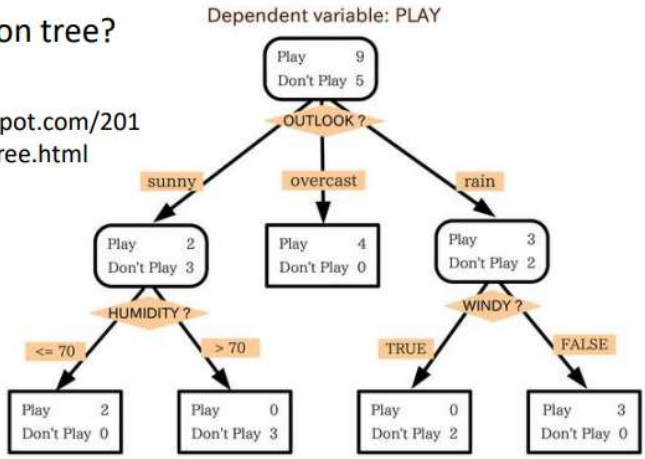
Source of image: https://kknews.cc/news/pnynzgp.html

- Explainable : 一個東西本來是黑箱，我們想辦法賦予它解釋的能力
- Interpretable : 一個東西本來不是黑箱，本來就知道它的內容

Interpretable v.s. Powerful

- Are there some models interpretable and powerful at the same time?
- How about decision tree?

Source of image:
https://mropengate.blogspot.com/2015/06/ai-ch13-2-decision-tree.html

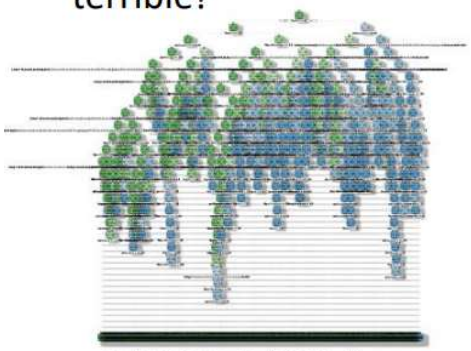


- 既 Interpretable 又 Powerful 的模型
  - Decision Tree 會不會是個好選擇呢？
    - 相較於 Linear model 是更強大的模型
    - 相較於 Deep Learning 是非常地 Interpretable

這堂課結束了 ... (?

Interpretable v.s. Powerful

- A tree can still be terrible!



https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

- We use a forest!

- Decision tree 也可能很複雜
- 其實真正用的技術叫做 Random Forest，是很多棵 Decision Tree 共同決定的結果

Goal of Explainable ML

- Completely know how an ML model works?
  - We do not completely know how brains work!
  - But we trust the decision of humans!

The Copy Machine Study (Ellen Langer, Harvard University)

"Excuse me, I have 5 pages. May I use the Xerox machine?" 60% accept

"Excuse me, I have 5 pages. May I use the Xerox machine, **because I'm in a rush**?" 94% accept

"Excuse me, I have 5 pages. May I use the Xerox machine, **because I have to make copies**?" 93% accept

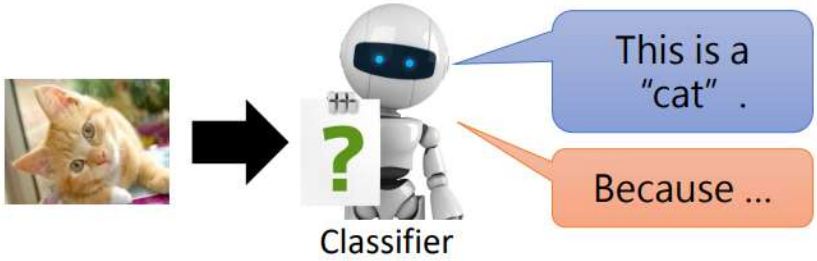
https://jamesclear.com/wp-content/uploads/2015/03/copy-machine-study-ellen-langer.pdf

Make people (your customers, your boss, yourself) comfortable.

(my two cents)

- 人能接受的 explanation 就是好的 explanation

Explainable ML



Local Explanation

Why do you think *this image* is a cat?

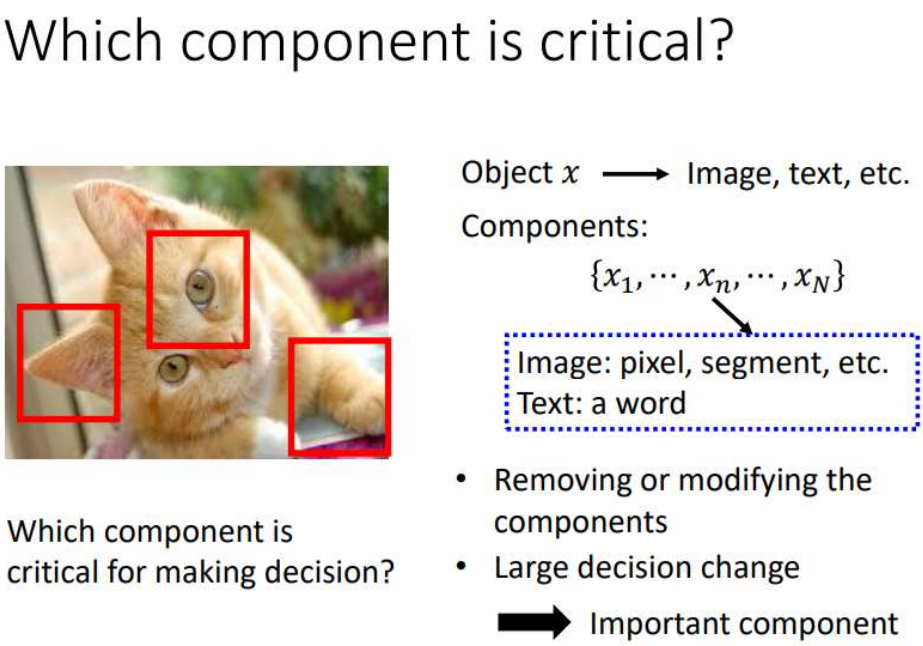
Global Explanation

What does a "cat" look like?

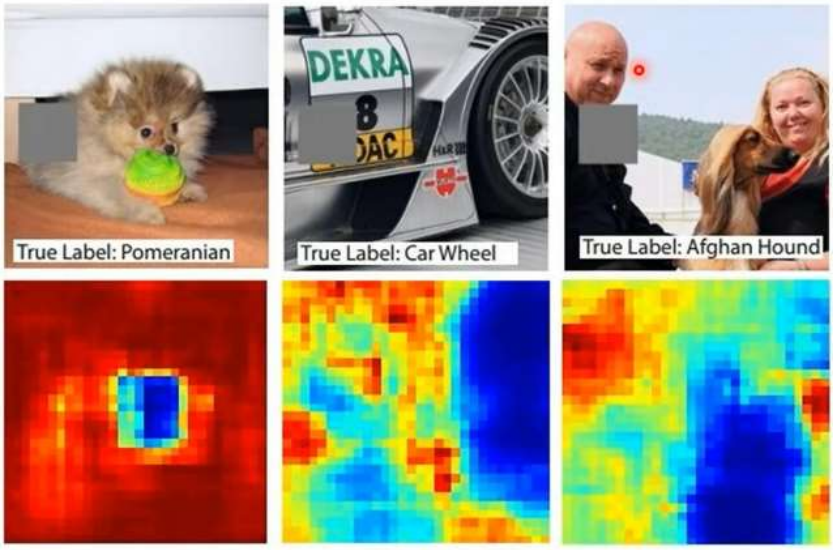
(not referred to a specific image)

- Explainable Machine Learning
  - Local Explanation
    - 假設我們有一個 image classifier，我們給它一張圖片讓它判斷是一隻貓
    - 根據某一張圖片來回答問題
    - 要問機器的問題是：為什麼你覺得這張圖片是一隻貓
  - Global Explanation
    - 沒有給 classifier 任何圖片
    - 不是對任何一張圖片進行分析
    - 要問機器的問題是：什麼樣的東西叫做一隻貓

Local Explanation

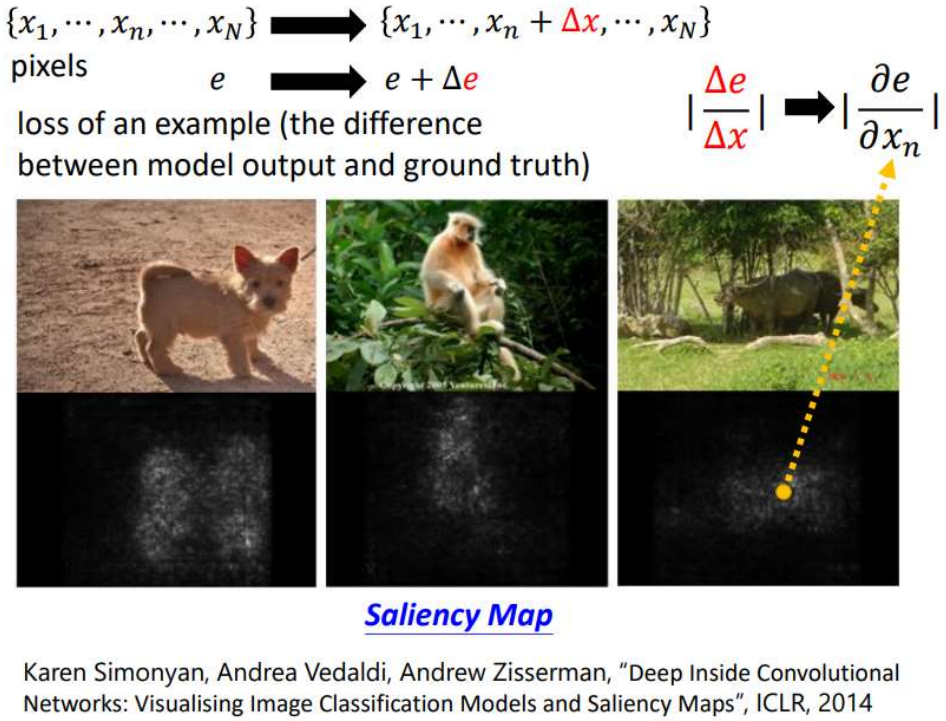


- 是什麼東西讓機器覺得是一隻貓
- 物件 x，有多個 components $\{x_1, x_2, \ldots, x_n, \ldots, x_N\}$
- 這些 components 哪個對於機器做出最終的判斷是最重要的呢
  - 把 components 拿出來做改造或刪除
  - 如果我們改造或刪除一個 components 之後，network 的輸出有巨大的改變，代表這個 components 很重要



Reference: Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014* (pp. 818-833)

- 在圖片的不同位置放上灰色的方塊
- 當方塊放在不同地方的時候，network 會 output 不同的結果
  - 下圖顏色代表輸出博美狗的機率，藍色是高、紅色是低



Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR, 2014

- 更進階的方法是計算 gradient
- $\{x_1, \ldots, x_n, \ldots, x_N\}$ 是圖片的 pixel
- 接著去計算圖片的 loss e，e 表示把圖片丟到模型裡面，模型輸出的結果與正確答案的差距 (Cross entropy)
- e 越大代表辨識的結果越差
- 怎麼知道某個 pixel 的重要性呢？
  - 把某個 pixel 加上 Δx
  - 再去看 loss 會有什麼變化 $e + \Delta e$
  - 如果 Δe 越大，代表那個 pixel 越重要
- 計算Δe 跟 Δx 的比值：$\left|\frac{\Delta e}{\Delta x}\right|$ 代表 x_n 的重要性
  - 比值越大，代表 x_n 越重要
- 把每一個 pixel 的比值都算出來，就會得到一個圖叫做 Saliency Map
  - 越偏白色，代表這個 pixel 越重要

Case Study: Pokémon v.s. Digimon

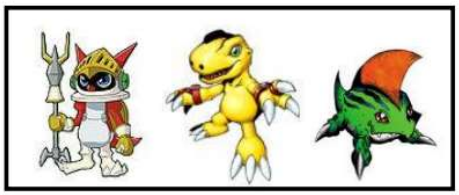
https://medium.com/@tyreeostevenson/teaching-a-computer-to-classify-anime-8c77bc89b881

Task

Pokémon images: https://www.Kaggle.com/kvpratama/pokemon-images-dataset/data

Digimon images: https://github.com/DeathReaper0965/Digimon-Generator-GAN

Pokémon



Digimon

Testing Images:

Experimental Results

```
model = Sequential()
model.add(Conv2D(32, (3, 3), padding='same', input_shape=(120,120,3)))
model.add(Activation('relu'))
model.add(Conv2D(32, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(64, (3, 3), padding='same'))
model.add(Activation('relu'))
model.add(Conv2D(64, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(256, (3, 3), padding='same'))
model.add(Activation('relu'))
model.add(Conv2D(256, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Flatten())
model.add(Dense(1024))
model.add(Activation('relu'))
model.add(Dense(2))
model.add(Activation('softmax'))
```

Training Accuracy: 98.9%

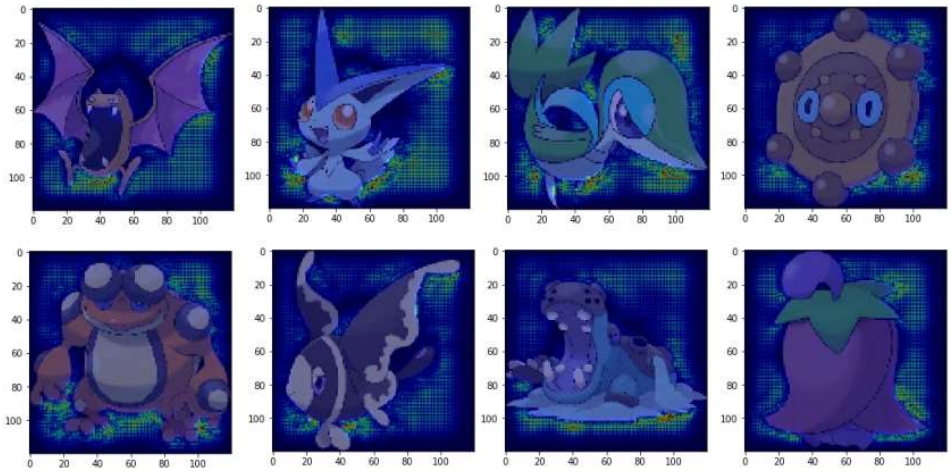
Testing Accuracy: 98.4%

Amazing!!!!!!

- 機器是憑藉什麼樣的規則，判斷寶可夢和數碼寶貝的差異呢？

Saliency Map

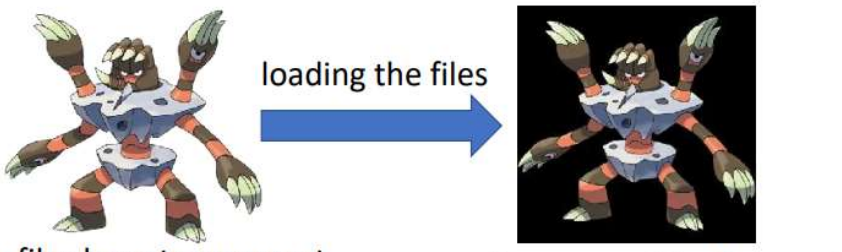
Saliency Map



- 發現亮亮的點都在四周，反而避開本體
- 為啥？

What Happened?

- All the images of Pokémon are PNG, while most images of Digimon are JPEG.



png files have transparent background

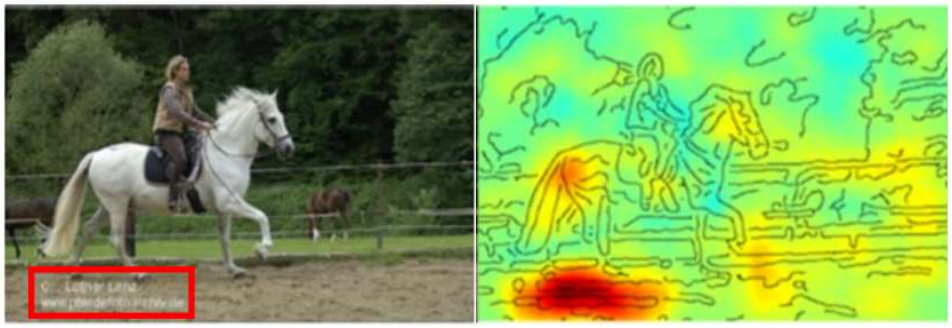
transparent background becomes black

Machine discriminates Pokémon and Digimon based on the background colors.

- 發現寶可夢都是 png 檔，數碼寶貝都是 jpeg 檔
- png 檔讀進來之後背景都是黑的，所以機器只要看背景就知道圖片是寶可夢還是數碼寶貝

More Examples ...

- PASCAL VOC 2007 data set

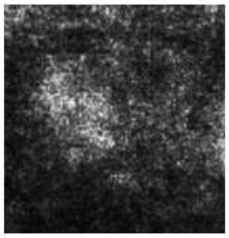


This slide is from: GCPR 2017 Tutorial — W. Samek & K.-R. Müller

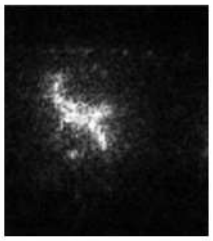
- 機器沒有學到馬的本體是什麼樣子
- 所以 Explainable Machine Learning 是很重要的

Limitation: Noisy Gradient

Gazelle
(瞪羚)



Typical



SmoothGrad

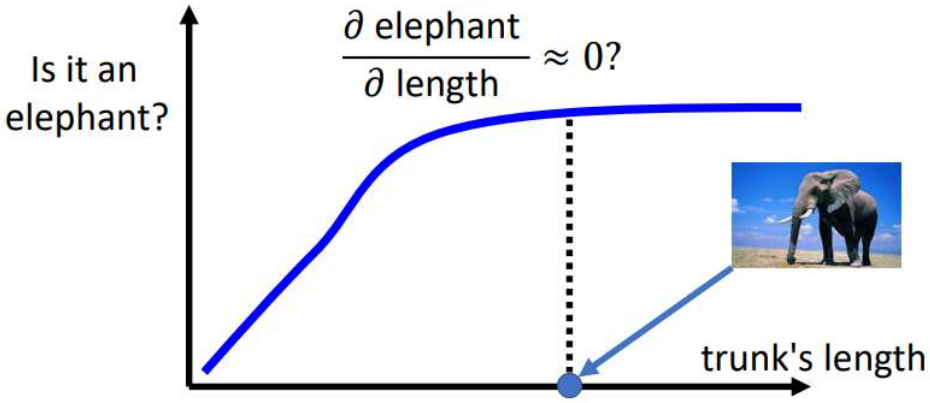
SmoothGrad: Randomly add noises to the input image, get saliency maps of the noisy images, and average them.

https://arxiv.org/abs/1706.03825

- 原圖是指瞪羚
- 期待機器做 Saliency map 時，會把它主要的精力集中在瞪羚身上
  - 但可能在瞪羚之外的地方也會有一些雜訊
- SmoothGrad 方法，會讓 Saliency map 上面的雜訊比較少
  - 方法是：在圖片上加上各種不同的雜訊
  - 加 100 種雜訊就有 100 種 saliency map
  - 平均下來，就得到 SmoothGrad 的結果

Limitation: Gradient Saturation

Gradient cannot always reflect importance



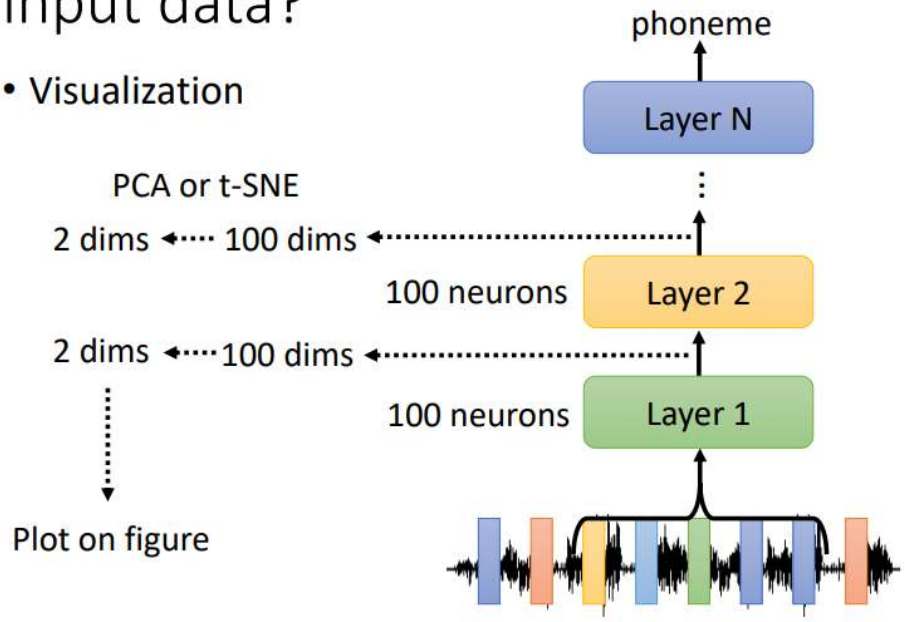
Alternative: Integrated gradient (IG)

https://arxiv.org/abs/1611.02639

- 光看 gradient 並不能完全反映一個 component 的重要性
- 橫軸 : 大象鼻子的長度
- 縱軸 : 是大象的可能性
- 另一個方法叫做 Integrated Gradient (IG)

How a network processes the input data?

- Visualization

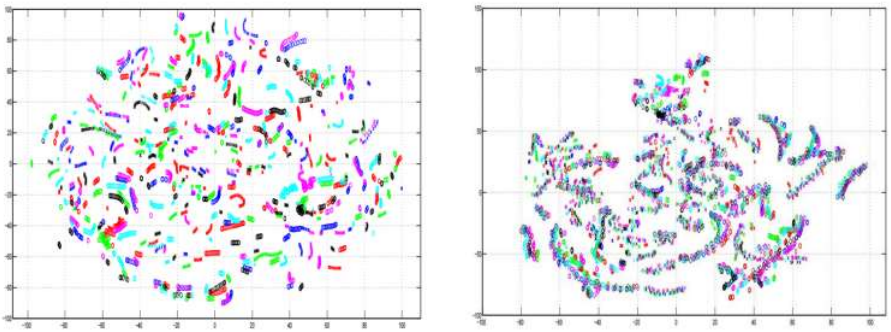


- 當我們給 network 一個輸入的時候，network 是怎麼去處理這個輸入，然後得到最終答案的？

How a network processes the input data?

- Visualization
  Colors: speakers

A. Mohamed, G. Hinton, and G. Penn, "Understanding how Deep Belief Networks Perform Acoustic Modelling," in ICASSP, 2012.



Input Acoustic Feature (MFCC)

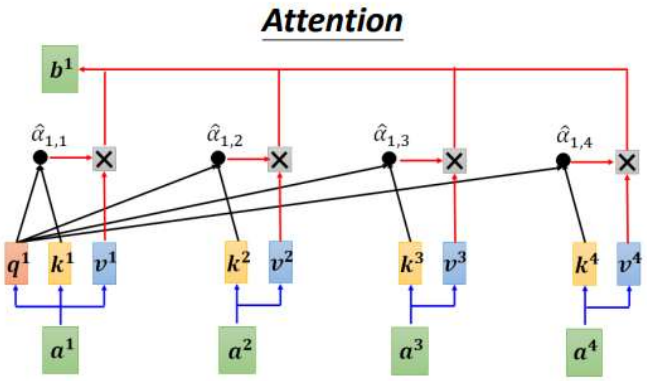
8-th Hidden Layer

- 把模型的 input Acoustic Feature (MFCC)，拿出來降到二維，畫在二維的平面上
- 圖上每一個點代表一小段聲音訊號
- 每一個顏色代表了某一個 Speaker
- 從 Acoustic Feature 發現不同的人說同樣的句子在圖上看不出來
- 同一個人他說的話就是比較相近
- 就算不同的人說同樣的句子，也不會被 align 在一起

- 把第八層的 network 拿出來看
- 圖片變成一條一條的
- 每一條就代表同樣內容的某一個句子
- 經過八層 network 之後，機器知道說這些話是同樣的內容

How a network processes the input data?

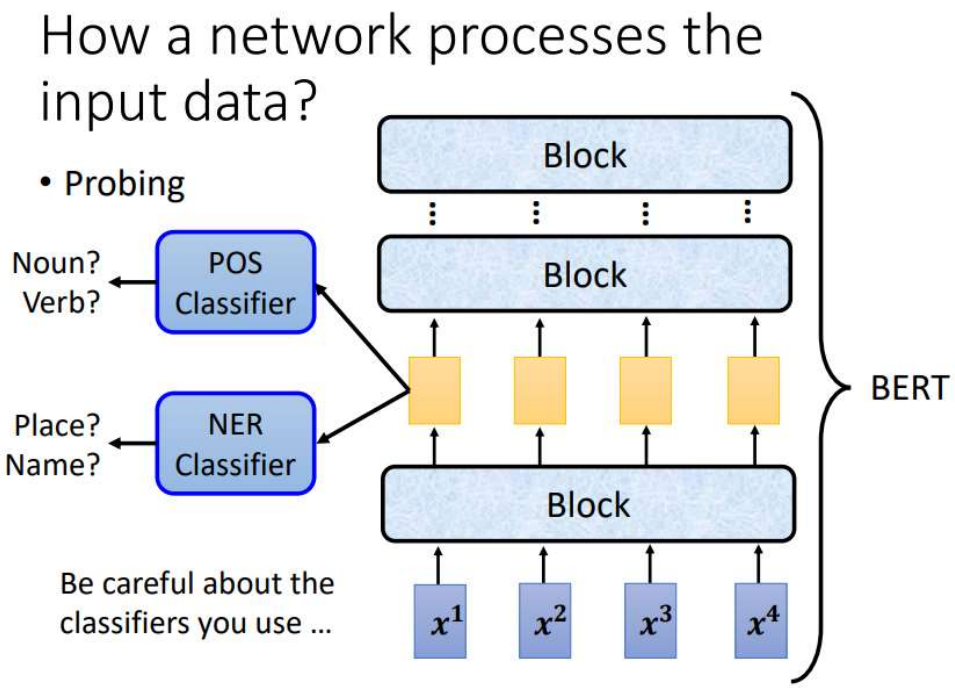
- Visualization



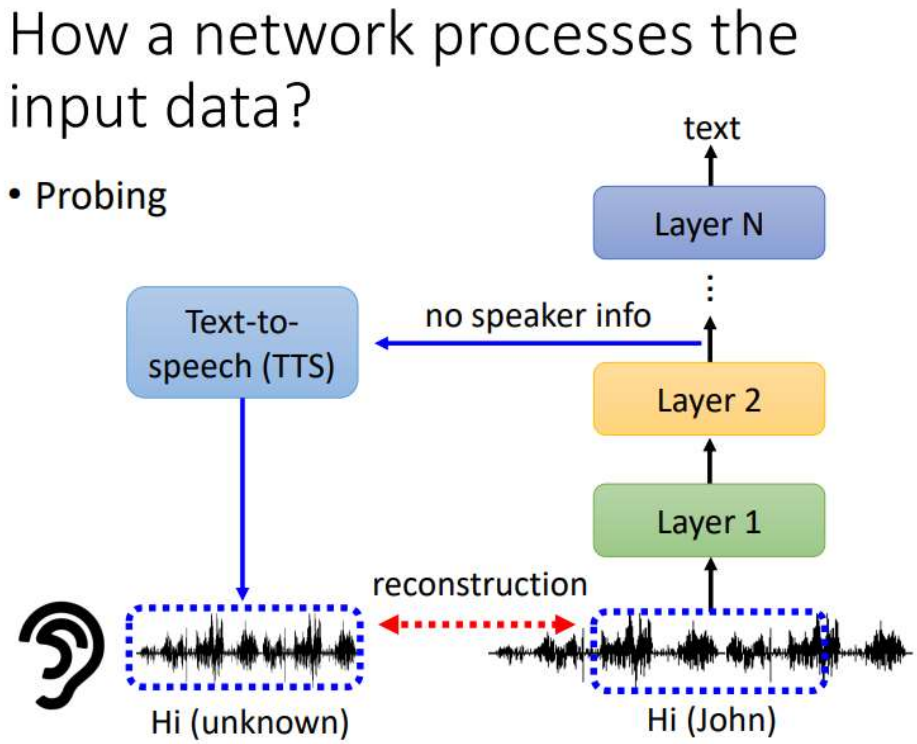
Attention is not Explanation
https://arxiv.org/abs/1902.10186

Attention is not not Explanation
https://arxiv.org/abs/1908.04626

- 人眼觀察



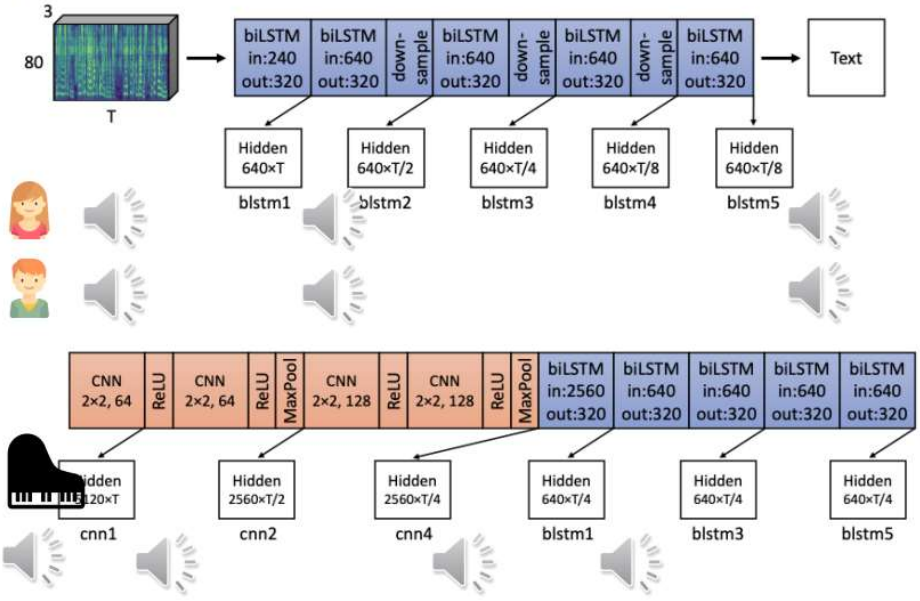
- Probing
- 可以訓練一個 probing (其實就是分類器)
- 分類器是要根據一個 feature 向量，決定現在這個詞彙它的 POS Classifier
  - 把 BERT 的 embedding 丟到 POS 的 Classifier 裡面，決定這些 embedding 是來自於哪一個詞性的詞彙
    - 如果 POS Classifier 的正確率高，就代表這些 embedding 裡面有很多詞性的資訊
- 或是 learn 一個 NER Classifier (Name Entity Recognition Classifier)，看 feature 之後決定看到的詞彙屬於人名還是地名或是不是專有名詞
- 用 probing model 的時候要小心，不要太快下結論
- 有時候只是因為 classifier 沒有 train 好，導致 classifier 的正確率沒有辦法當評斷的依據



- probing 不一定要是 classifier
- 這邊的語音模型不是吃一段文字，是吃 network output 的 embedding 作為輸入，然後試圖去輸出一段聲音訊號
  - 希望 TTS 模型可以去重現 network 的輸入

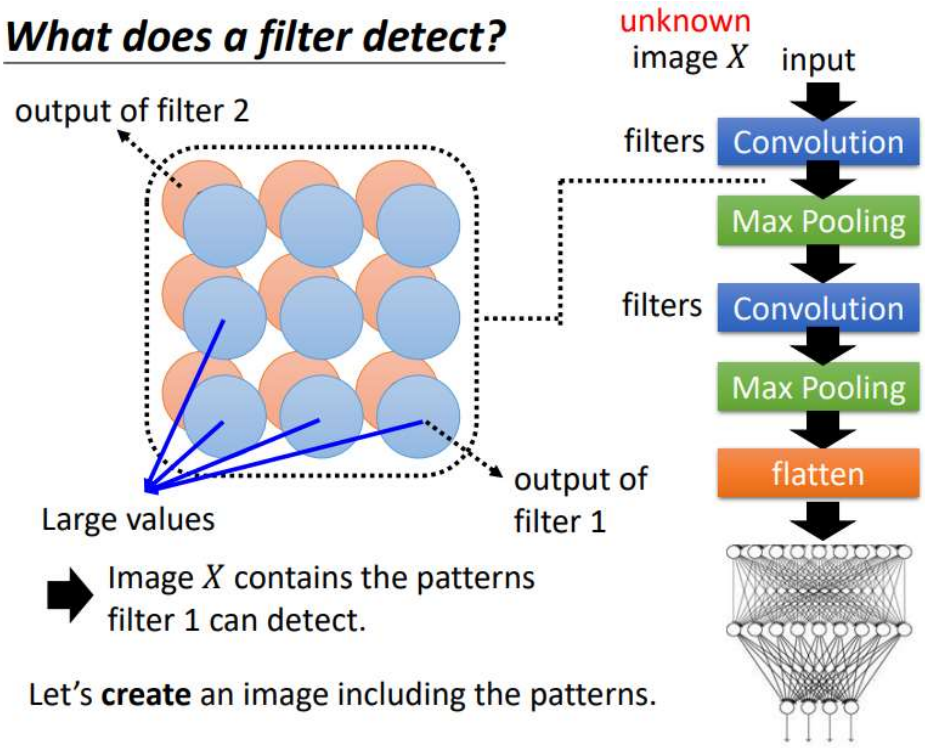
What does a network layer hear? Analyzing hidden representations of end-to-end ASR through speech synthesis

https://arxiv.org/abs/1911.01102
https://youtu.be/6gtn7H-pWr8

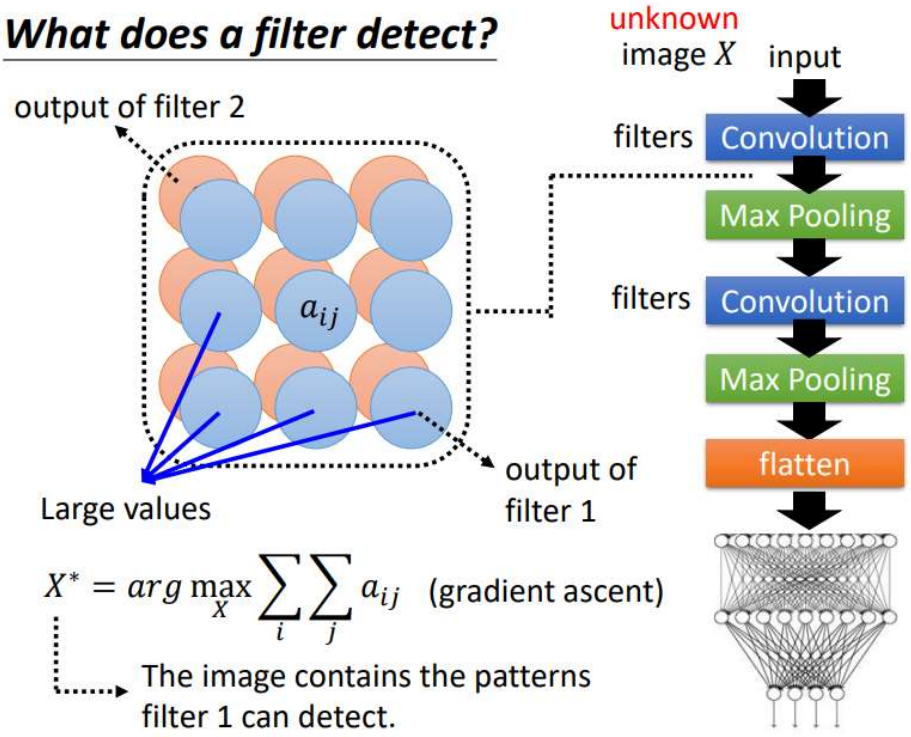


Global Explanation

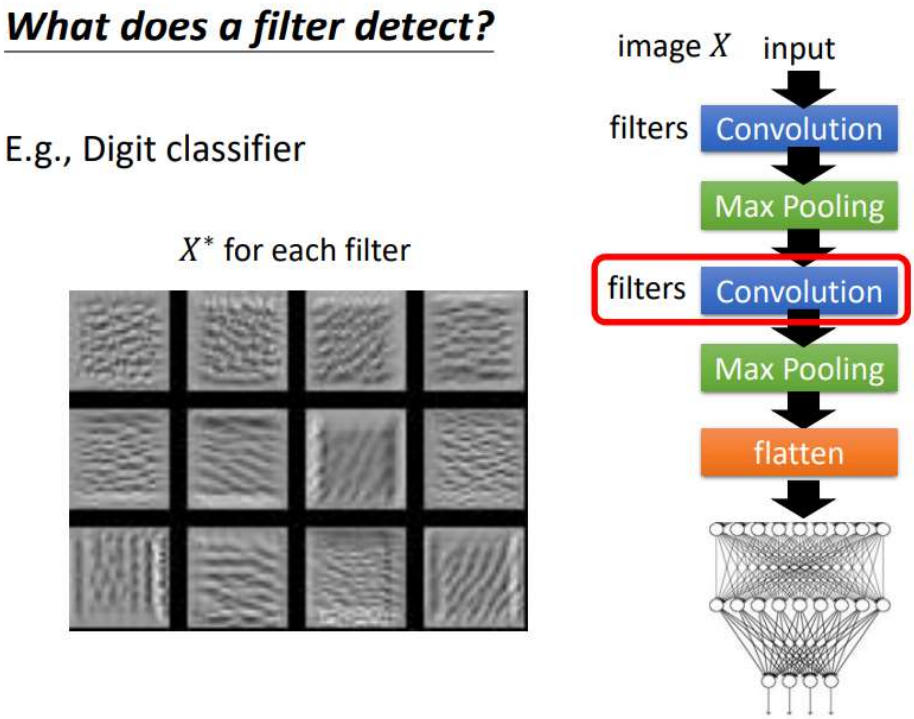
把訓練好的模型拿出來，根據模型裡面的參數去檢查，說對這個 network 而言，一隻貓長什麼樣子



- 現在想要知道，對 filter 1 而言，想要看的 pattern 長什麼樣子
- 創造一張圖片，這張圖片它包含有 filter 1 要 detect 的 pattern
- 藉由看這張圖片裡面的內容，就可以知道 filter 1 它負責 detect 什麼樣的東西



- 怎麼找這張圖片？
  - 假設左圖是 filter 1 的 feature map
  - 裡面的每一個 element 叫做 a_{ij}
  - 現在要找一張圖片 X，這張圖片不是 database 裡面的圖片，而是我們把這個 X 當作一個 unknown variable，當作我們要訓練的那個參數
  - 丟到 filter 通過 convolution layer 以後，輸出 feature map 之後，filter 1 對應的 feature map 裡面的值 a_{ij} 越大越好
  - 所以我們要找一個 X，讓 a_{ij} 的總和越大越好，用 X^* 表示
  - 然後去觀察 X^* 有什麼樣的特徵，可以 Maximize 這個 filter map 的 value
  - 就是 filter 1 在 detect 什麼樣的 pattern

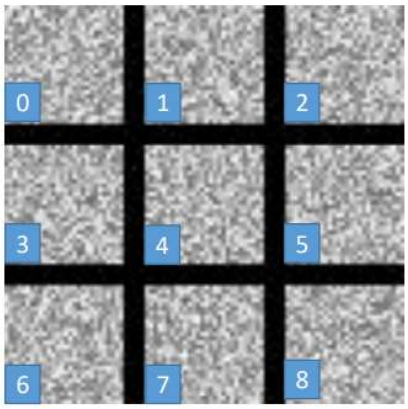


- 用 mnist train 出一個 classifier
- 這個 classifier 給它一張圖片，會判斷這張圖片是 0~9 哪一個數字
- 訓練好這個 classifier 之後，把它的第二層 convolution layer 裡面的 filter 拿出來
- 找出每一個 filter 對應的 X^*
- 所以左圖的每一張圖片都是一個 X^*
- 每一個圖片都是某一個 filter 想要 detect 的 pattern

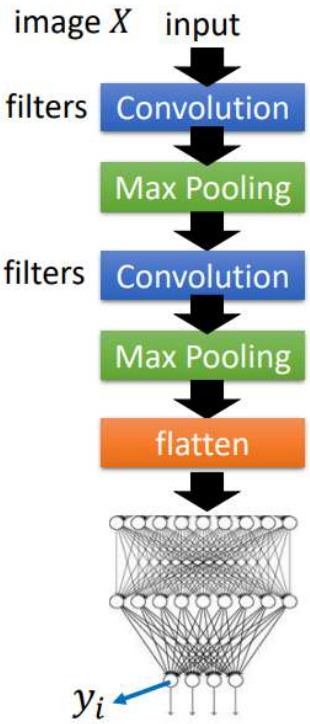
What does a digit look like for CNN?

E.g., Digit classifier

$X^* = arg \max_X y_i$ Can we see digits?



Surprise? Consider adversarial attack!

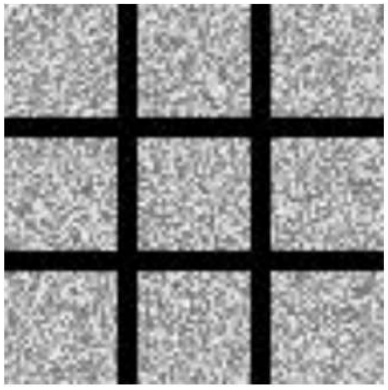


- 假設我們不是看某一個 filter，而是去看最終的 image classifier 的 output
- 想辦法找一個 X，X 可以讓某一個類別的分數越高越好
- 現在是做手寫辨識，所以 y 總共會有 10 個值分別對應到 0~9
- 例如選數字 1 出來，希望找一張圖片丟到 classifier 之後，數字 1 的分數越高越好
- 對機器來說它不需要看到很像某個數字的圖片，才知道是某個數字，有時候給亂七八糟的雜訊，就可以讓機器看到各式各樣的物件

What does a digit look like for CNN?

Find the image that maximizes class probability

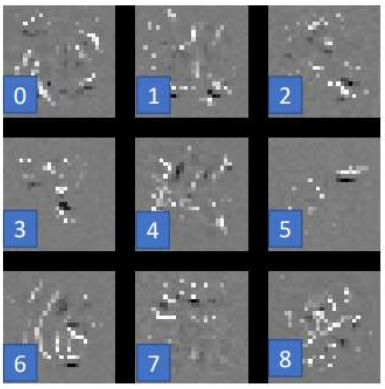
$X^* = arg \max_X y_i$



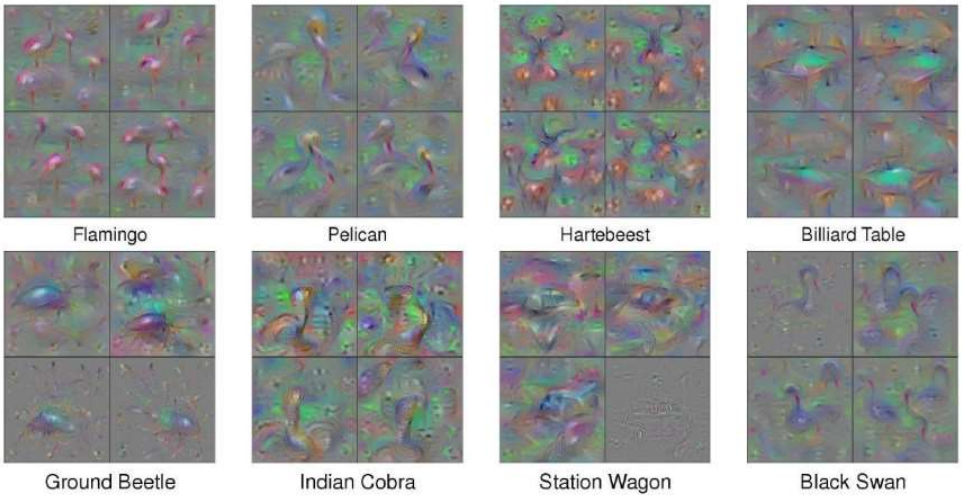
The image should looks like a digit.

$X^* = arg \max_X y_i + R(X)$

$R(X) = -\sum_{i,j} |X_{ij}|$ How likely X is a digit



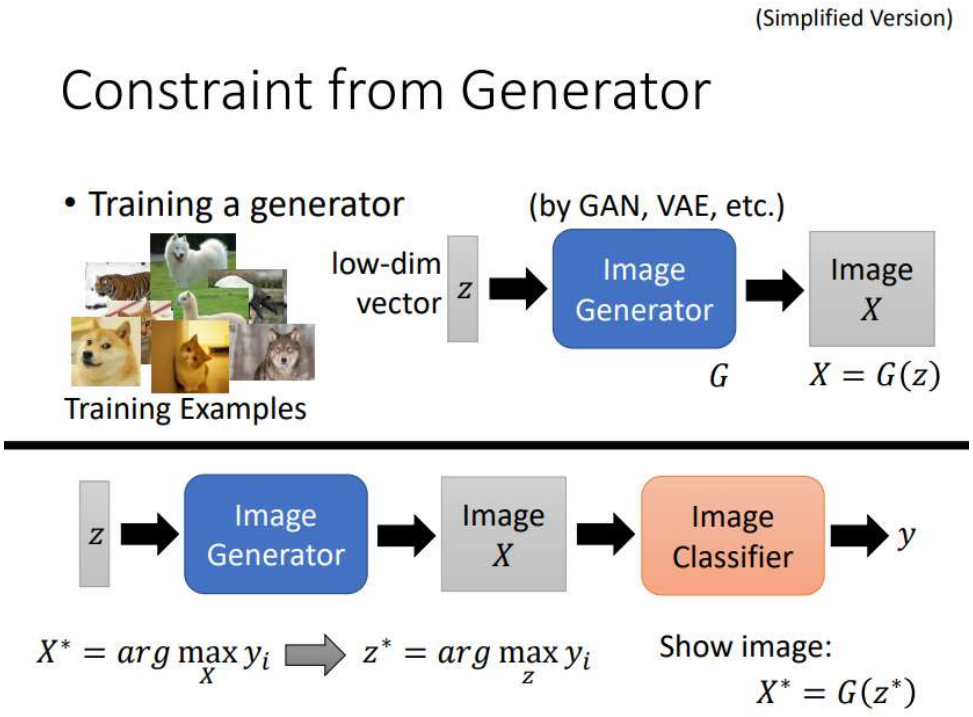
- 假設我們希望今天看到的比較像是人想像的數字
- 在解 optimization 問題的時候，要加上更多的限制
- 我們把我們要的限制加到 optimization 的過程裡面
- 現在要找一個 X，同時讓 y_i 和 $R(X)$ 的分數都越大越好
- $R(X)$ 是要拿來衡量 X 有多像是一個數字



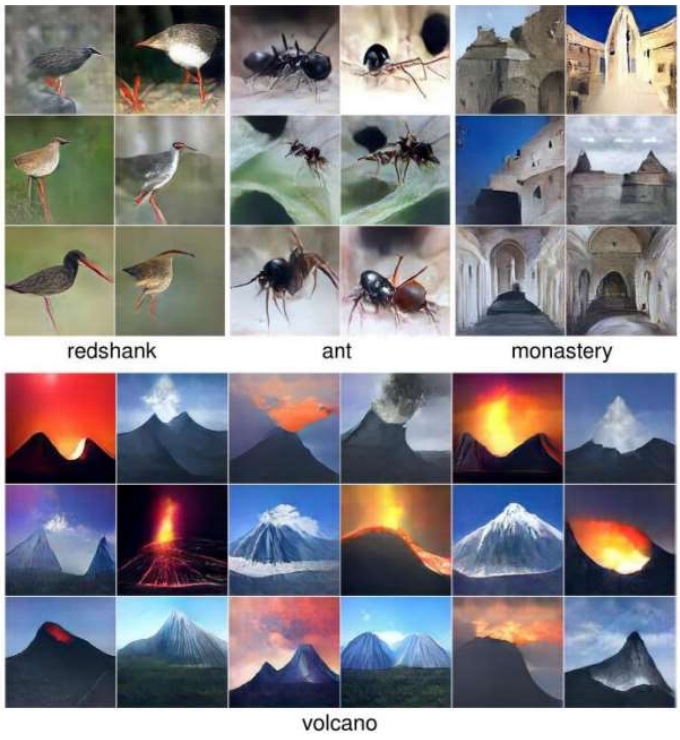
With several regularization terms, and hyperparameter tuning .....

https://arxiv.org/abs/1506.06579

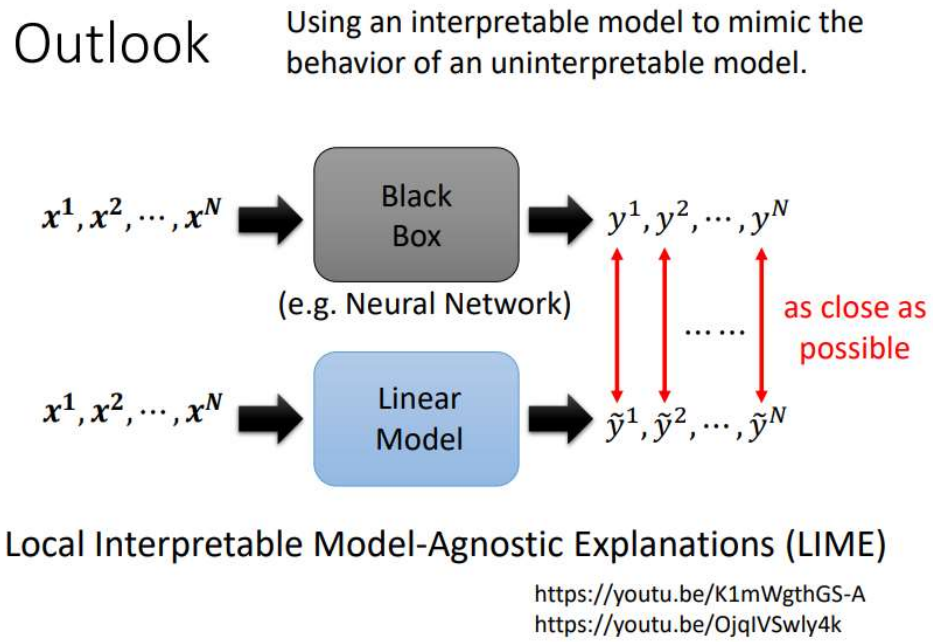
- 要加一大堆的限制才有可能做到
- 不容易啊



- 如果想要看到一張很清晰的圖片的話
- 現在有一個技術是使用 Generator
- 訓練一個 image generator
  - 輸入 low-dim vector z (從 gaussian distribution 裡面 sample 出來的低維度向量)
  - 丟到 image generator 之後
  - 輸出就是一張圖片 X
- 怎麼拿 image generator 來幫助我們反推一個 image classifier 裡面，它所想像的某一種類別長什麼樣子？
  - 把 image generator 跟 image classifier 接在一起
  - image generator 輸入 z 輸出一張圖片 X
  - image classifier 輸入 X 輸出分類結果 y
  - 要找一個 z 通過 image generator 產生 X，再把 X 丟到 image clasifier 產生 y 之後，希望 y 對應的某個類別它的分數越大越好



https://arxiv.org/abs/1612.00005



- Explainable 的 Machine Learning 還有很多的技術
- 舉例來說
  - 可以用一個比較簡單的模型，想辦法去模仿複雜的模型的行為
  - 有一個 neural network 因為它是一個 black box
    - 輸入東西進去，輸出結果出來，但是我們不知道它中間決策的過程
    - 因為 neural network 本身非常的複雜
- 拿一個簡單的模型出來，比較能夠分析的模型 (Linear model、Interpretable)
- 去訓練這個 Linear model 去模仿 Neural network 的行為
- 如果可以成功模仿黑盒子的行為，再去分析 linear model 做的事情，也許就可以得到黑盒子在做的事情
- 但是 neural network 可以做到的事情，linear model 不一定能做到

tags: 2022 李宏毅_機器學習