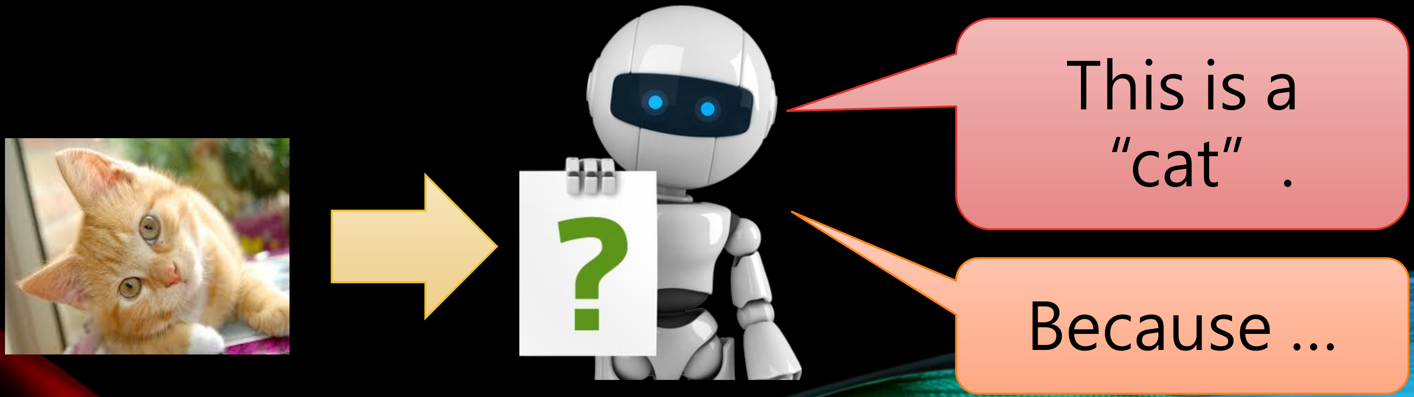
EXPLAINABLE MACHINE LEARNING

Hung-yi Lee 李宏毅



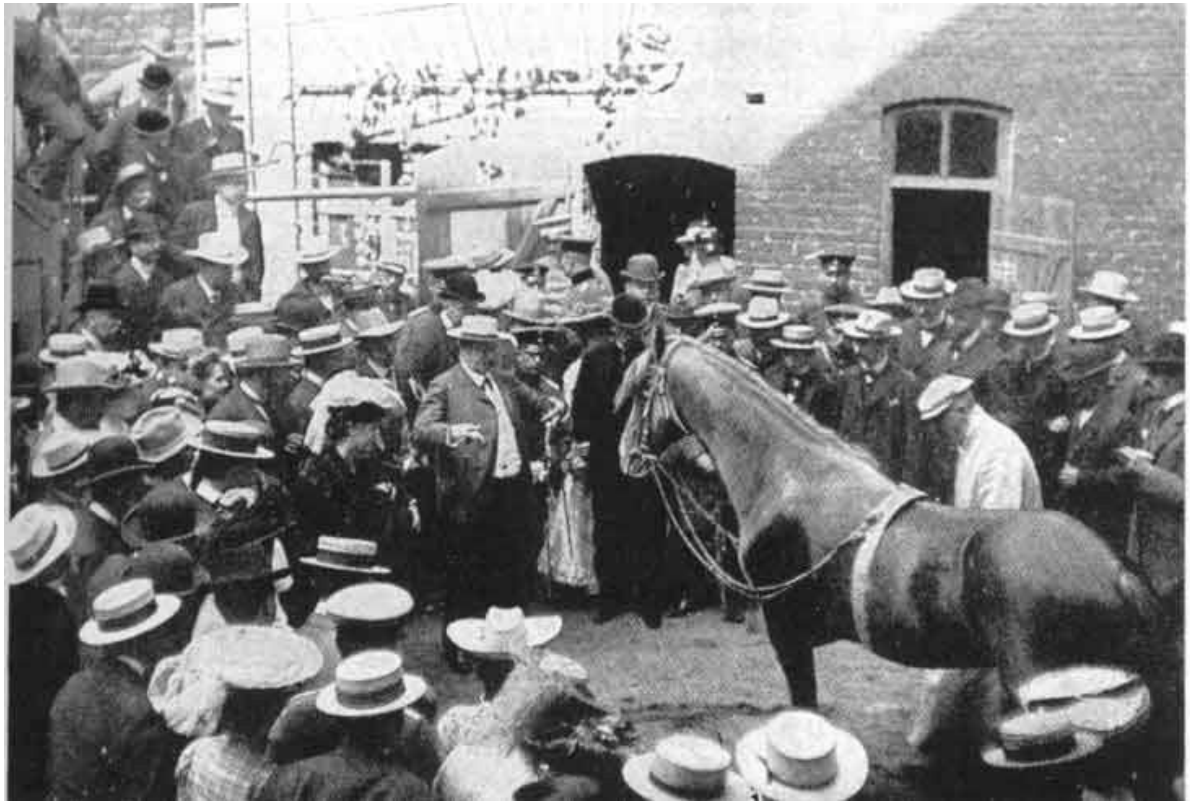
This is a "cat" .

Because ...

Why we need Explainable ML?

- Correct answers ≠ Intelligent

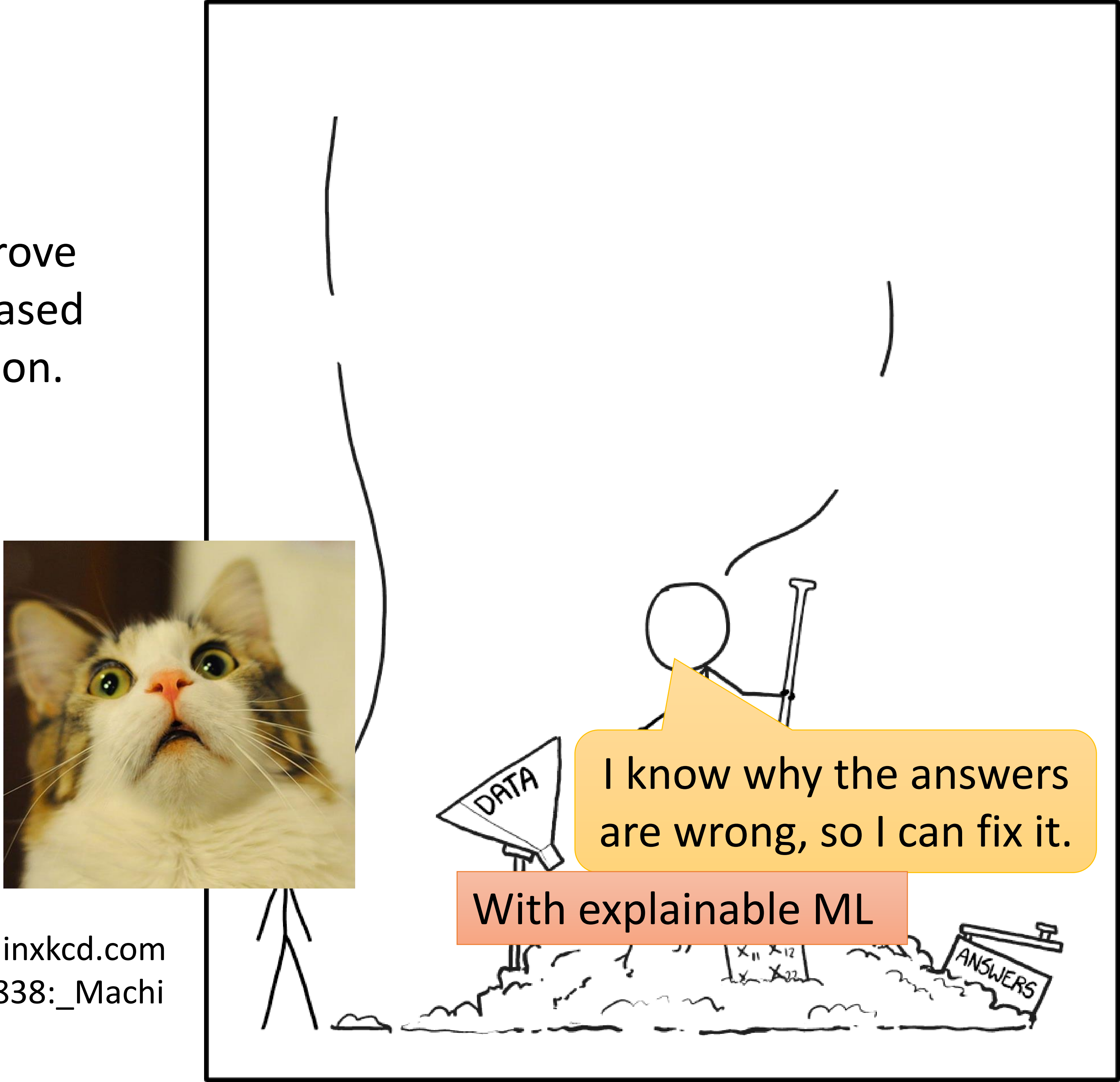
有一隻馬
他會在眾人面前算數學
例如簡易的加法
（他會用踩踏次數給出答案）
但他其實不是真的會算
而是看周遭人們的反應
而去看什麼時候要停止踩踏



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.

We can improve ML model based on explanation.



I know why the answers are wrong, so I can fix it.

With explainable ML

https://www.explainxkcd.com/wiki/index.php/1838:_Machine_Learning

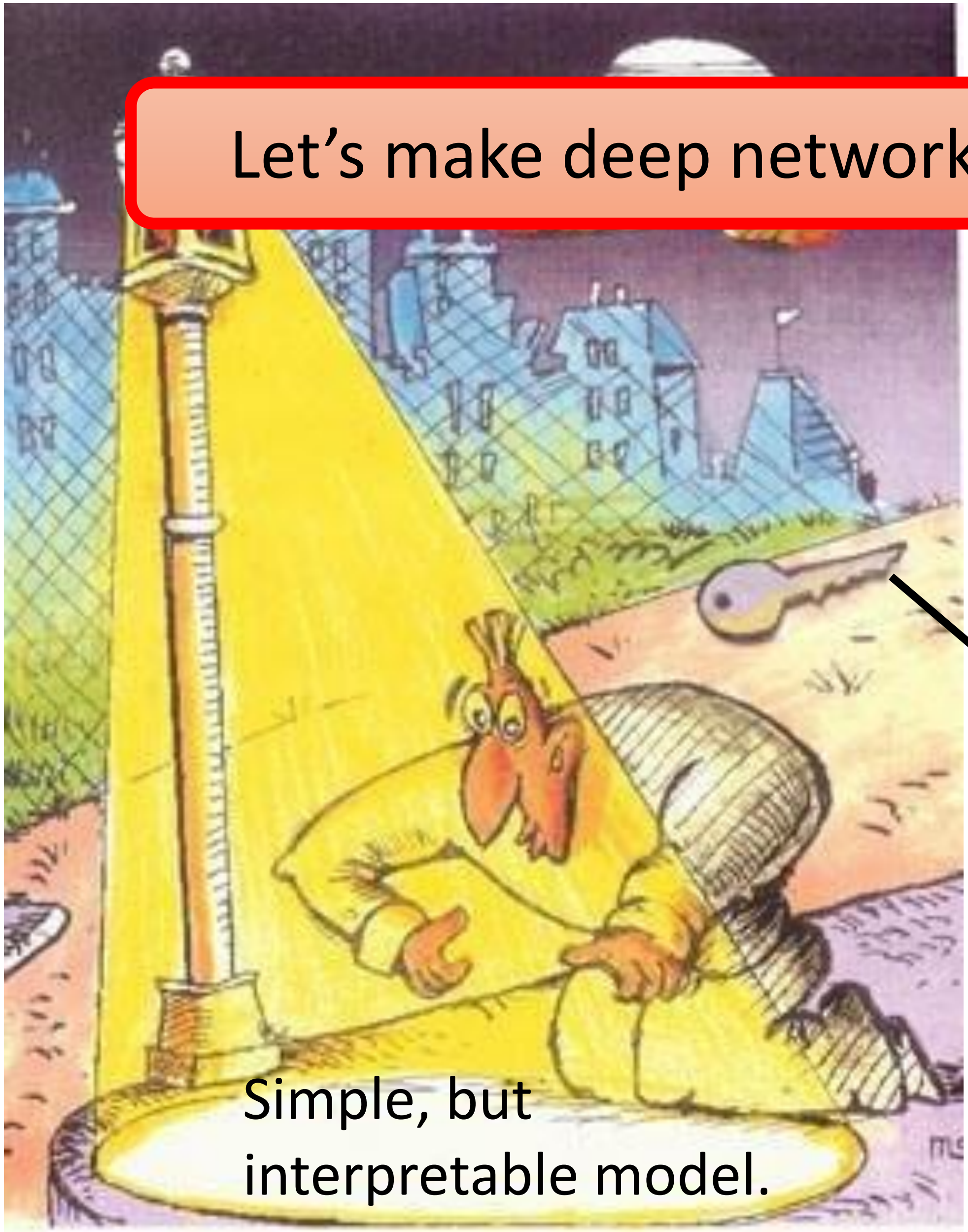
Interpretable v.s. Powerful

- 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)

Let’s make deep network explainable.



Powerful model

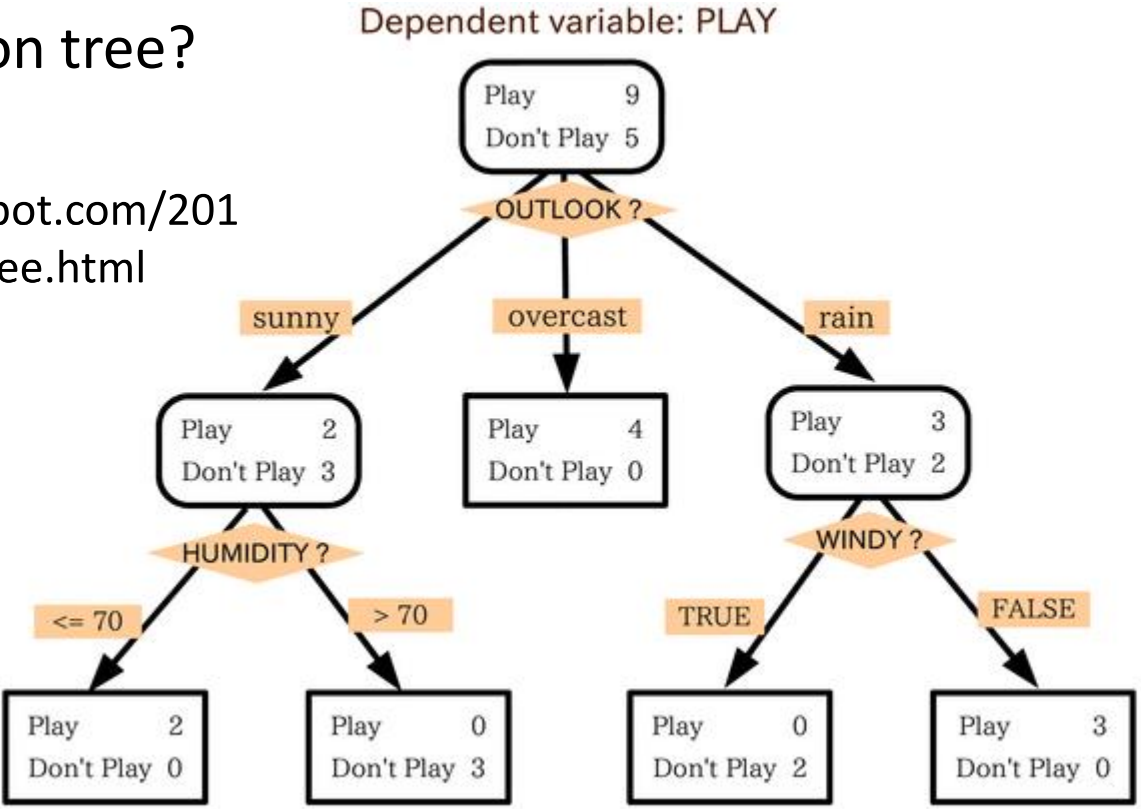
Simple, but interpretable model.

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

Interpretable v.s. Powerful

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

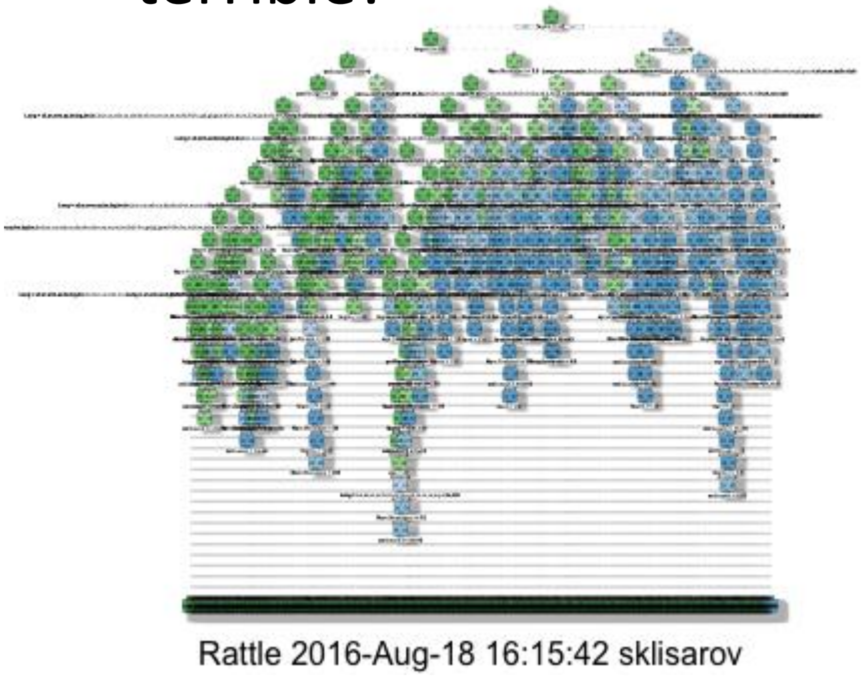
Source of image: https://mropengate.blogspot.com/2015/06/ai-ch13-2-decision-tree.html



Decision tree is all you need!?

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!

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**?"

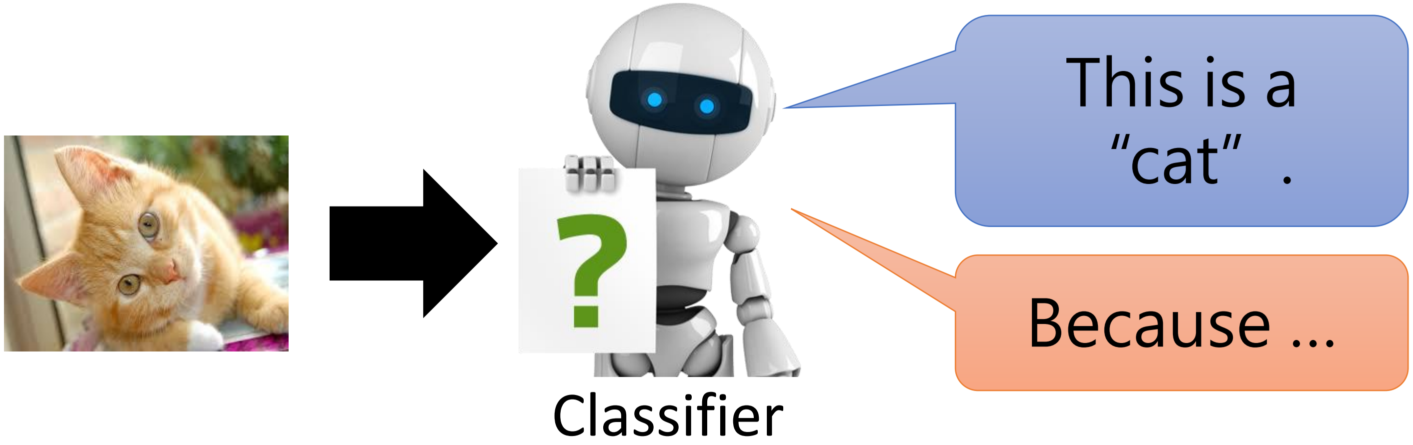
93% accept

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)

Explainable ML



This is a "cat" .

Because ...

Classifier

Local Explanation

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

Global Explanation

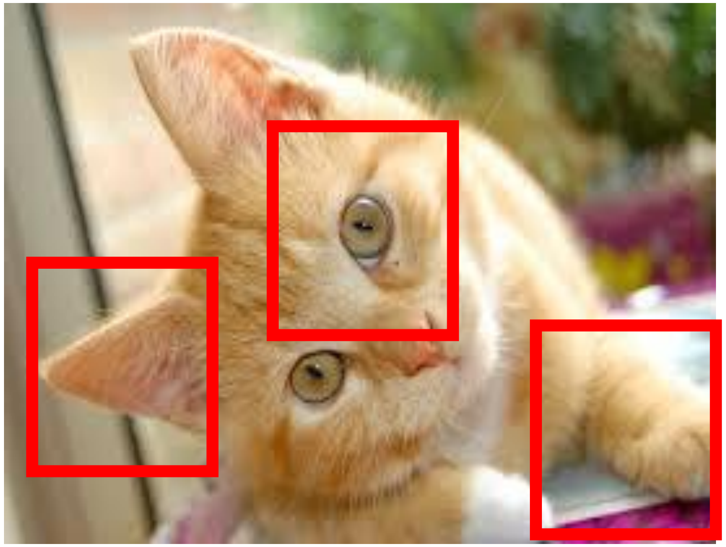
What does a "cat" look like?

(not referred to a specific image)

Local Explanation: Explain the Decision

Questions: Why do you think this image is a cat?

Which component is critical?



Which component is critical for making decision?

Object x ⟶ Image, text, etc.

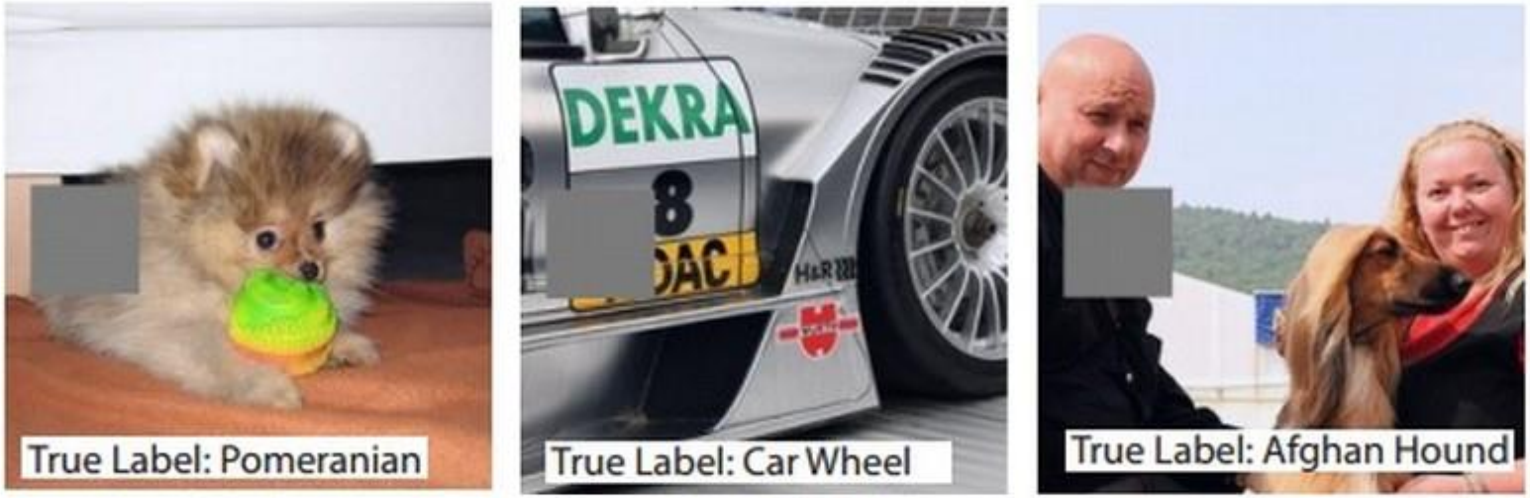
Components:

$$\{x_1, \cdots, x_n, \cdots, x_N\}$$

Image: pixel, segment, etc.
Text: a word

- Removing or modifying the components
- Large decision change

➡ Important component



使用灰色框框，然後去覆蓋掉圖片中每一個位置
若model的confidence變低了，就代表機器是看那個地方做決定的

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

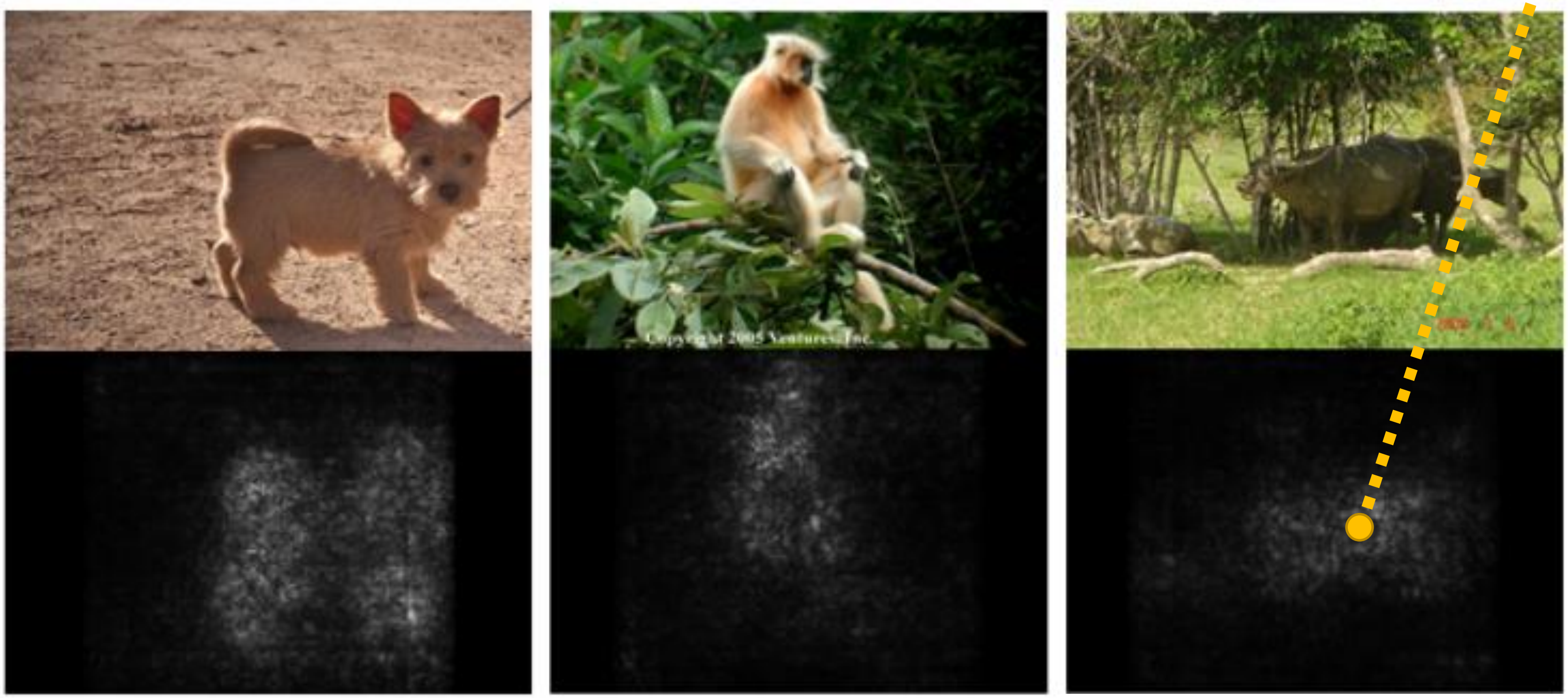
$$\{x_1, \cdots, x_n, \cdots, x_N\} \longrightarrow \{x_1, \cdots, x_n + \Delta x, \cdots, x_N\}$$

pixels

$$e \longrightarrow e + \Delta e$$

loss of an example (the difference between model output and ground truth)

$$\left|\frac{\Delta e}{\Delta x}\right| \longrightarrow \left|\frac{\partial e}{\partial x_n}\right|$$



Saliency Map

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

Case Study: Pokémon v.s. Digimon

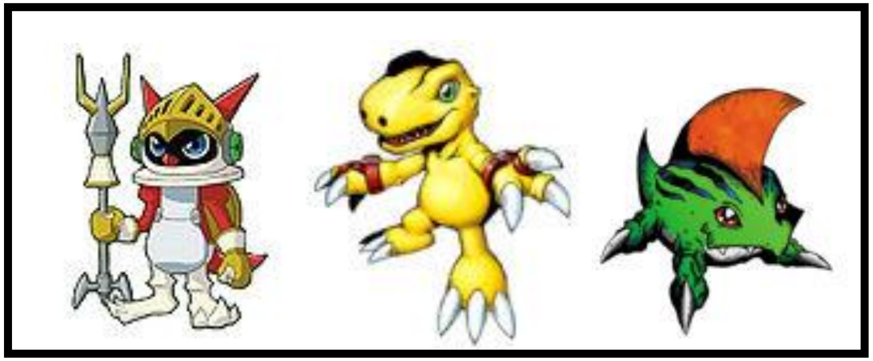
https://medium.com/@tyreeostevenson/teaching-a-computer-to-classify-anime-8c77bc89b881

Task

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

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

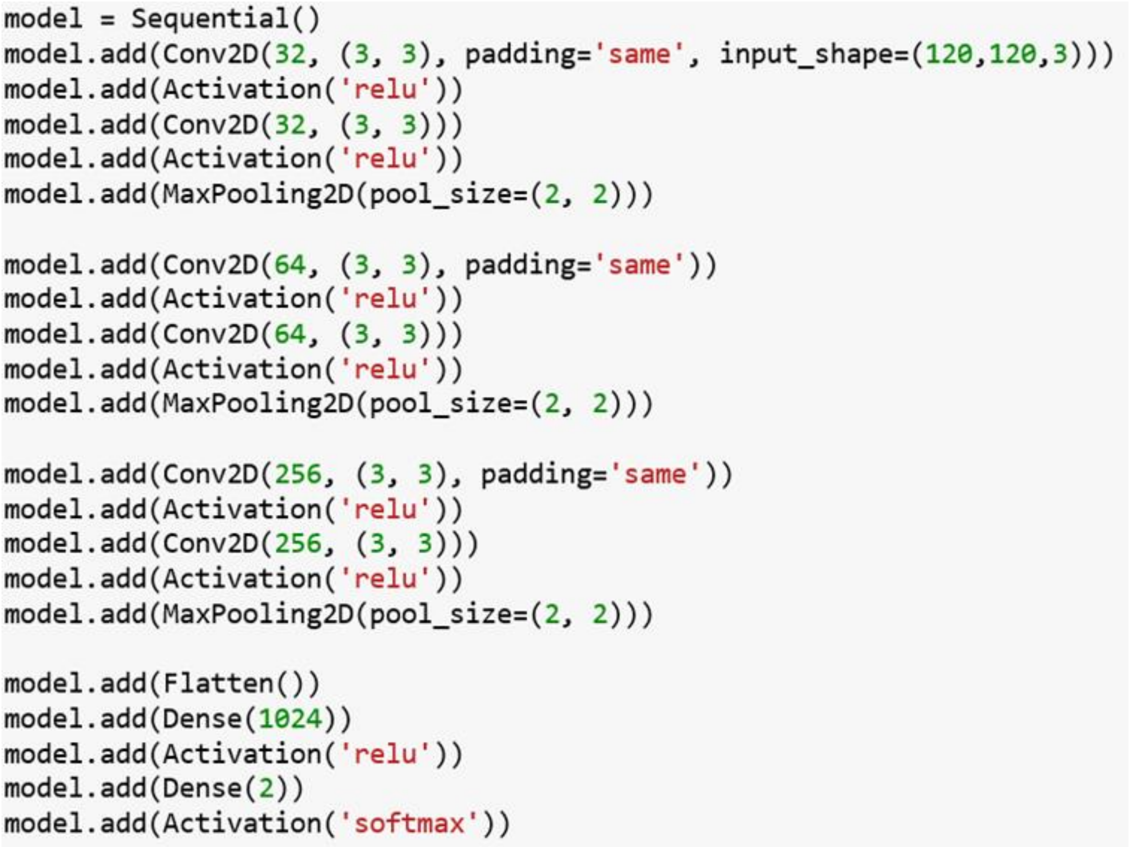
Pokémon



Digimon

Testing Images:

Experimental Results



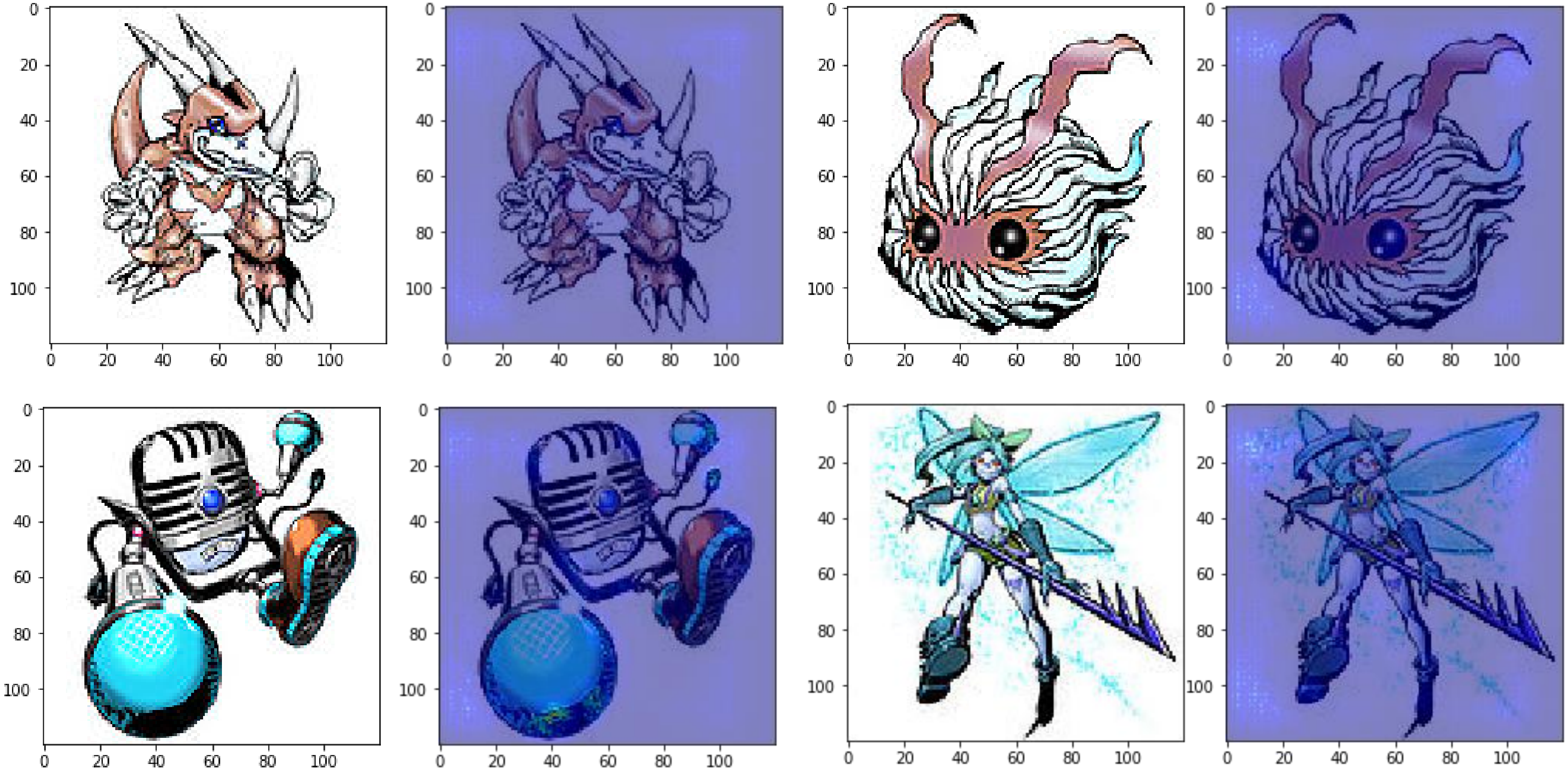
```
model = Sequential()
model.add(Conv2D(32, (3, 3), padding='same', input_shape=(120,120,3)))
model.add(Activation('relu'))
model.add(Conv2D(32, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

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%

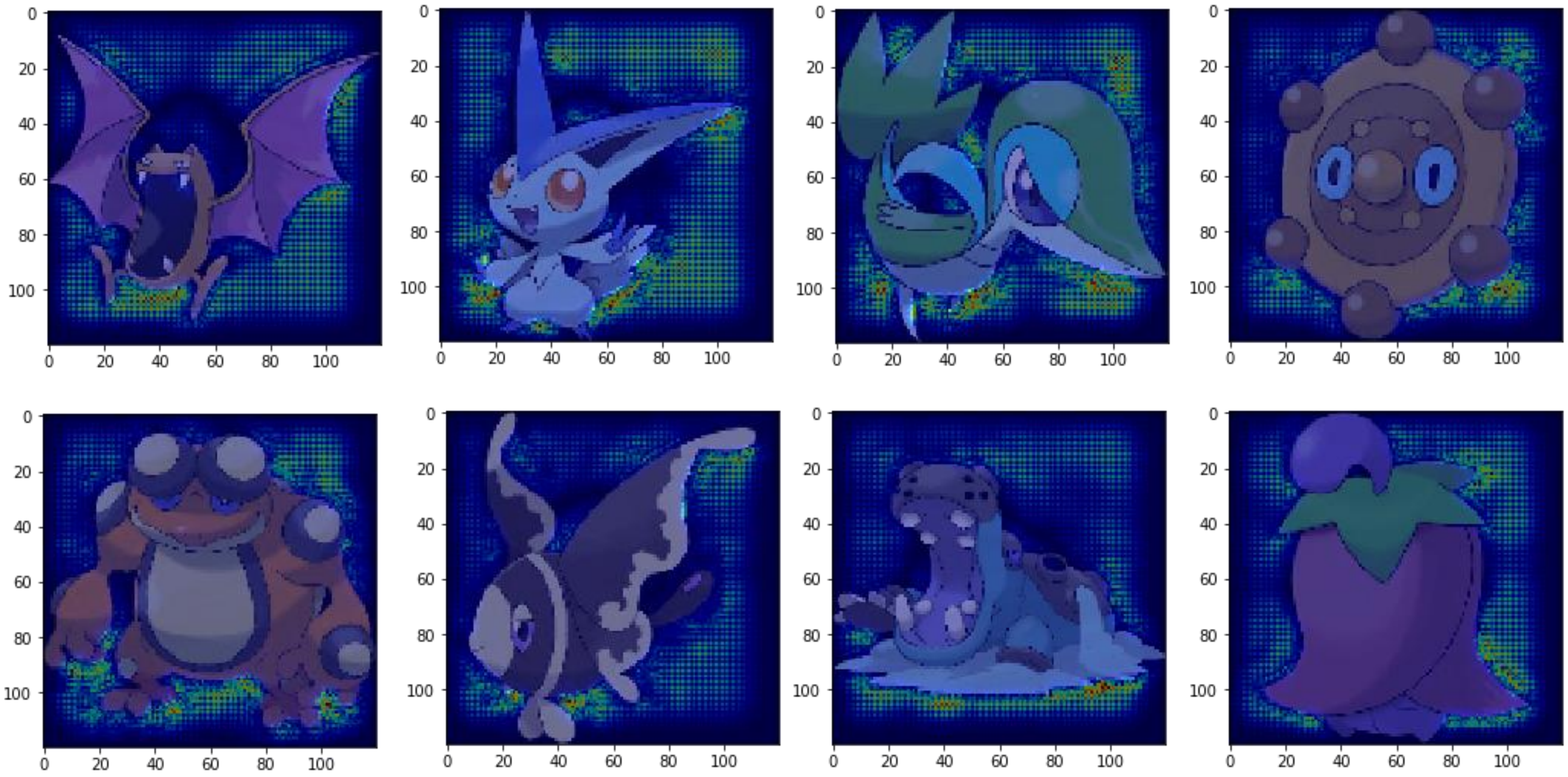
Testing Accuracy: 98.4%

Amazing!!!!!!

Saliency Map

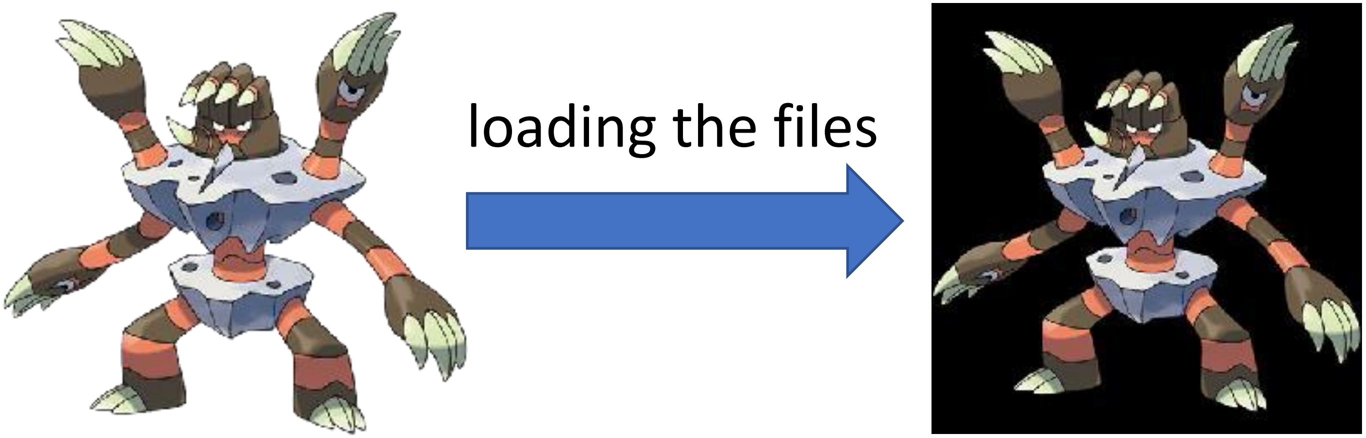


Saliency Map



What Happened?

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



loading the files

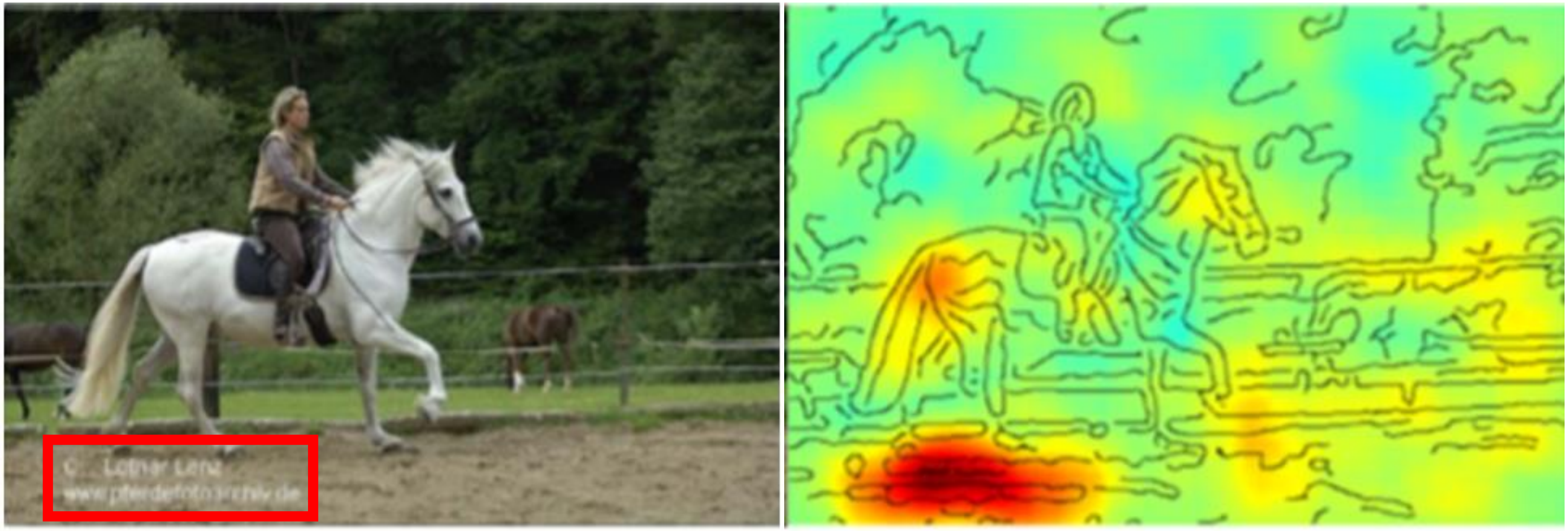
png files have transparent background

transparent background becomes black

Machine discriminates Pokémon and Digimon based on the background colors.

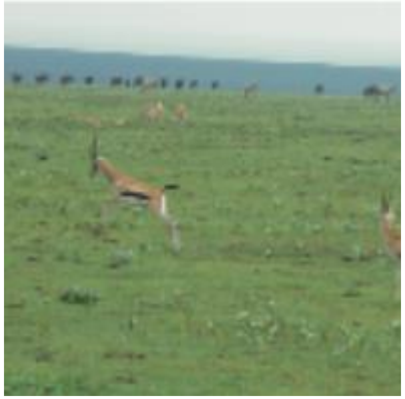
More Examples …

- PASCAL VOC 2007 data set

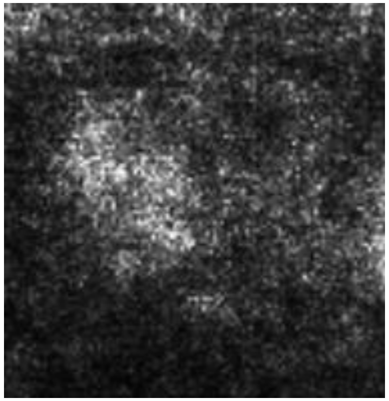


This slide is from: GCPR 2017 Tutorial — W. Samek & K.-R. Müller

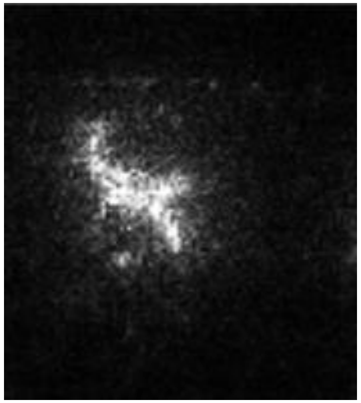
Limitation: Noisy Gradient



Gazelle
(瞪羚)



Typical



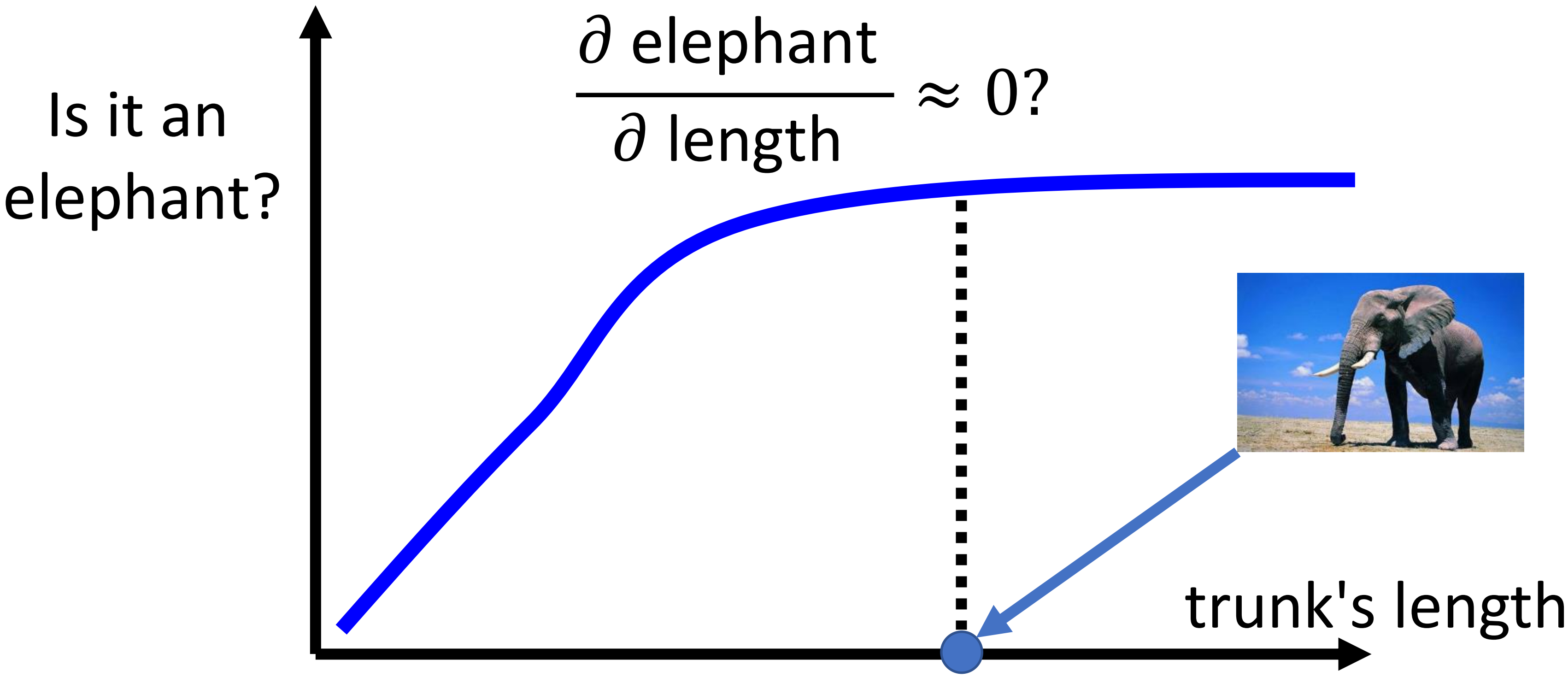
SmoothGrad

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

https://arxiv.org/abs/1706.03825

Limitation: Gradient Saturation

Gradient cannot always reflect importance



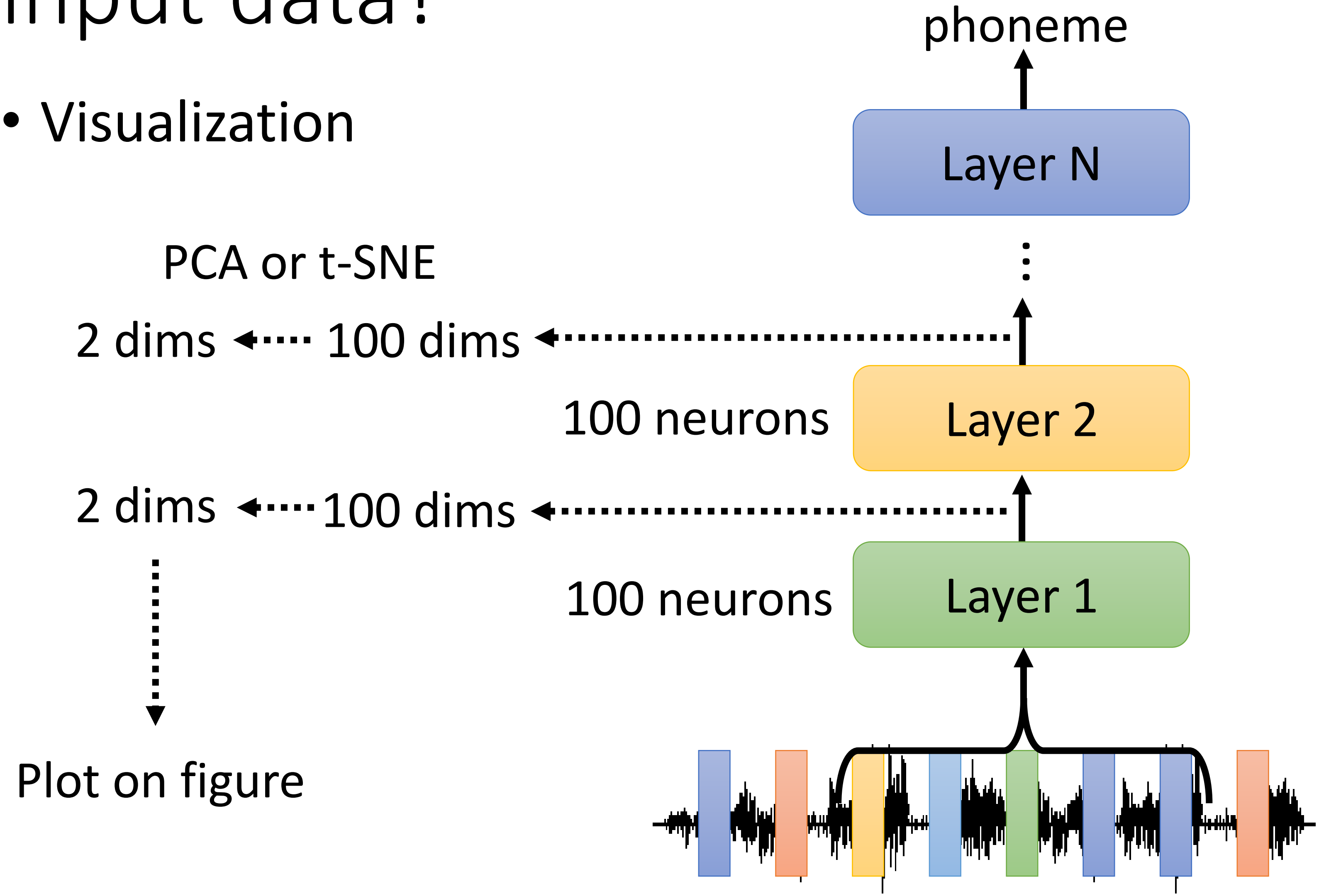
$$\frac{\partial \text{ elephant}}{\partial \text{ length}} \approx 0?$$

Alternative: Integrated gradient (IG)

https://arxiv.org/abs/1611.02639

How a network processes the input data?

- Visualization



phoneme

Layer N

PCA or t-SNE

2 dims ← 100 dims ←

100 neurons

Layer 2

2 dims ← 100 dims ←

100 neurons

Layer 1

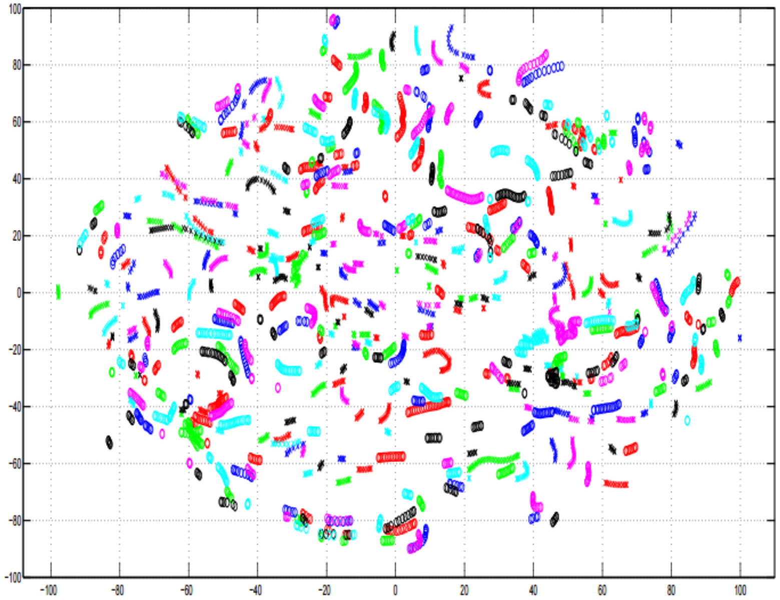
Plot on figure

How a network processes the input data?

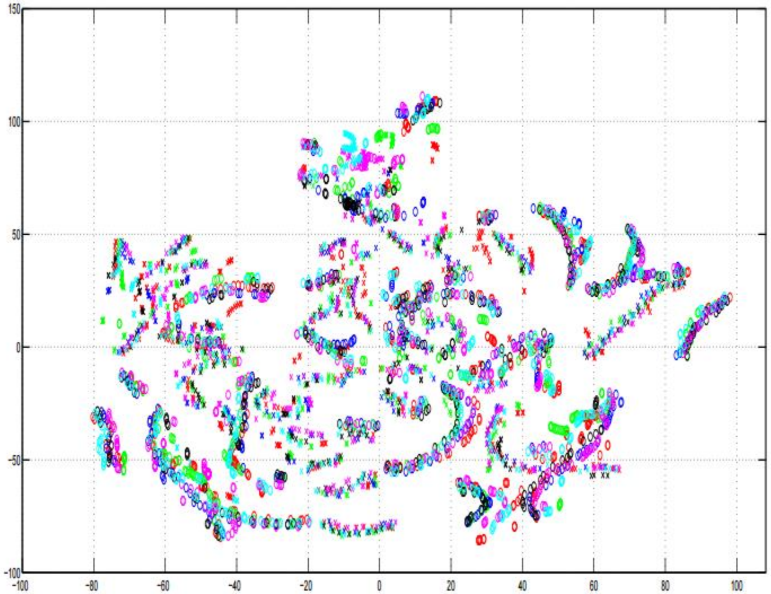
A. Mohamed, G. Hinton, and G. Penn, “Understanding how Deep Belief Networks Perform Acoustic Modelling,” in ICASSP, 2012.

- Visualization

  Colors: speakers



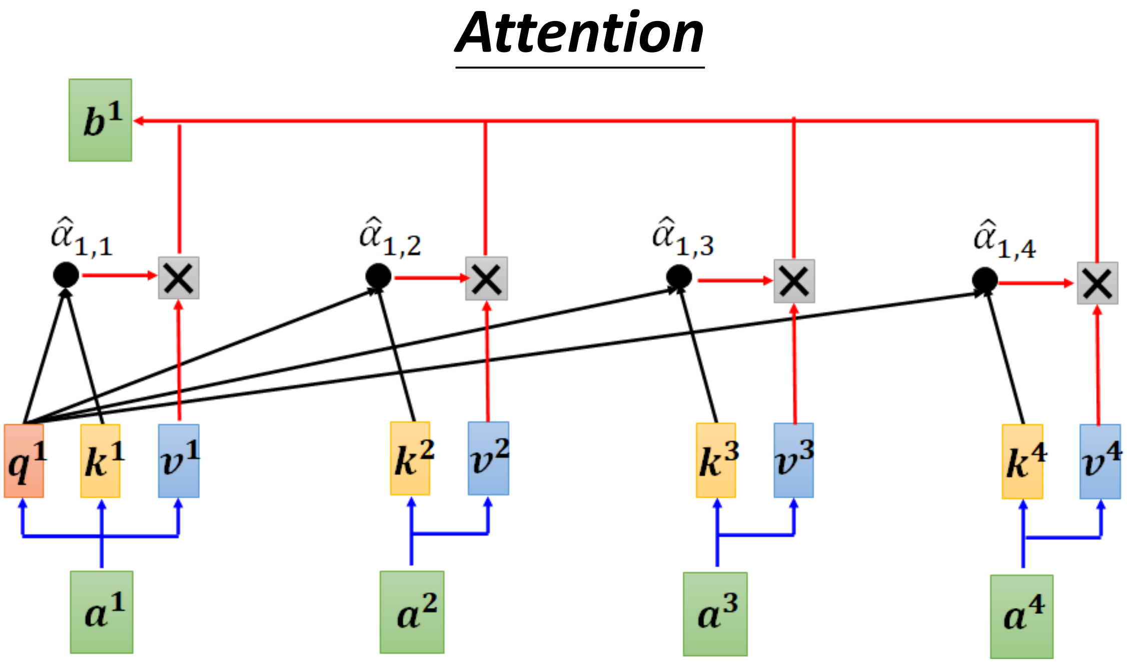
Input Acoustic Feature (MFCC)



8-th Hidden Layer

How a network processes the input data?

- Visualization

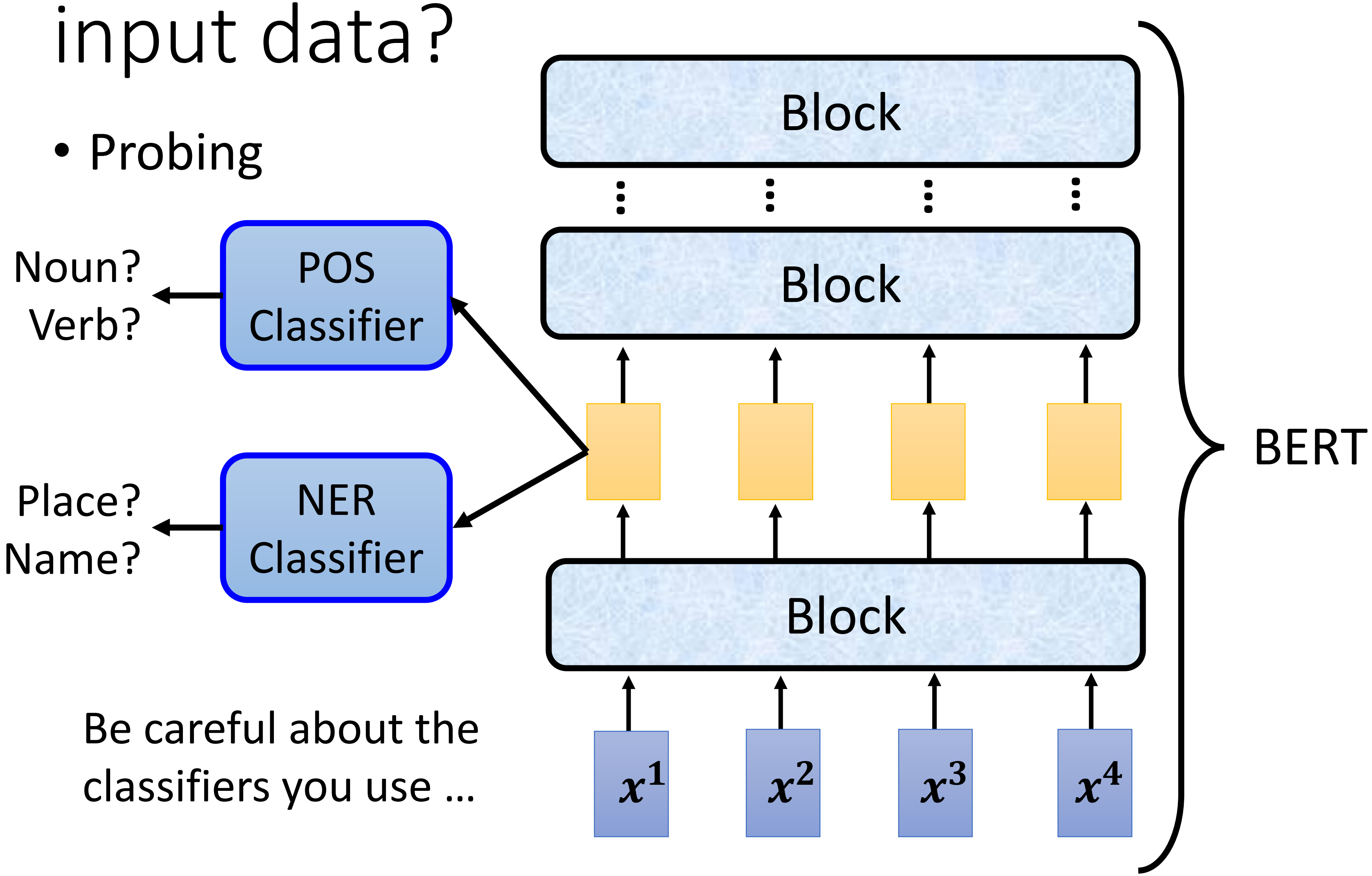


Attention is not Explanation
https://arxiv.org/abs/1902.10186

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

How a network processes the input data?

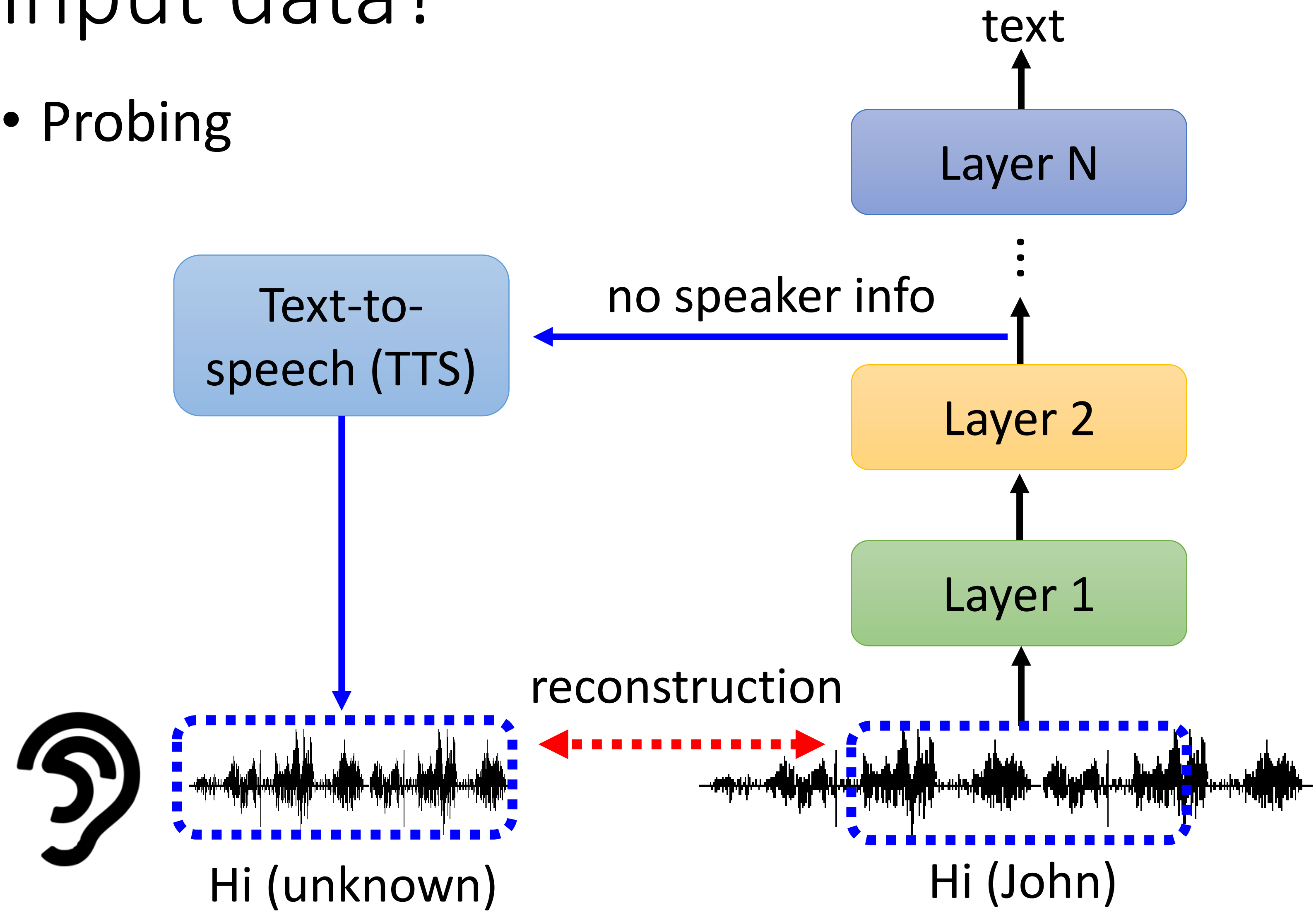
- Probing



Be careful about the classifiers you use …

How a network processes the input data?

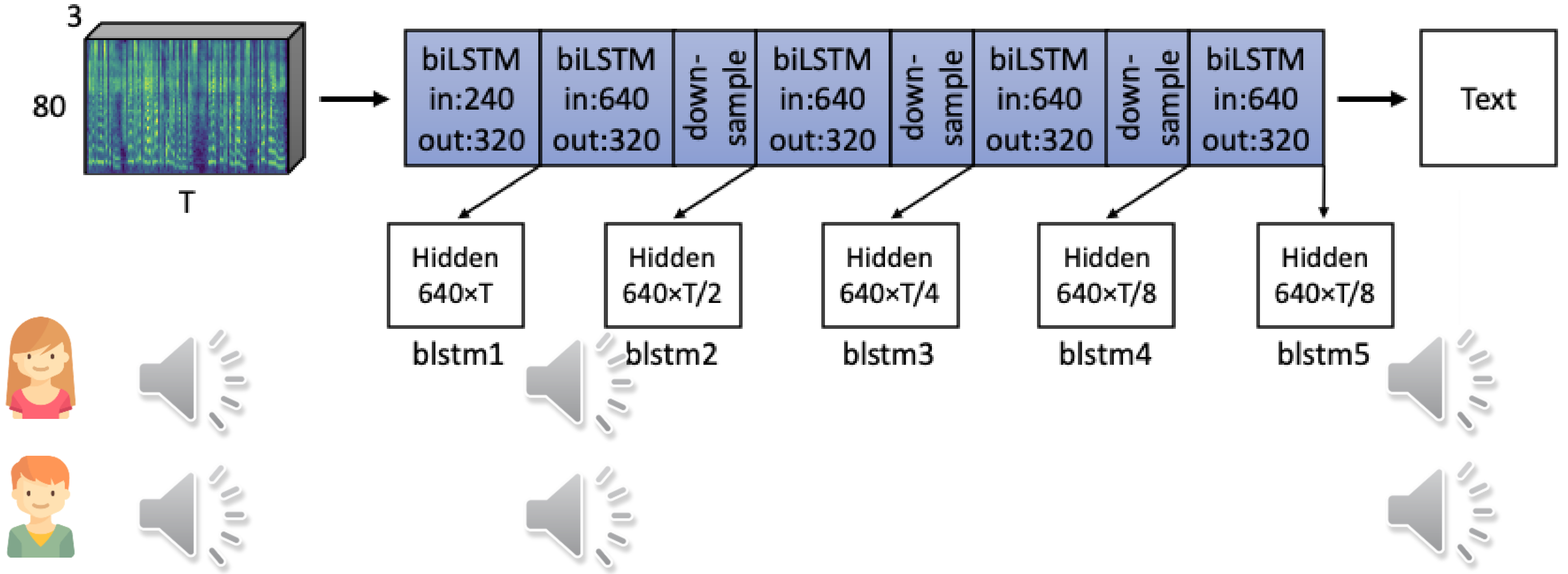
- Probing



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



3

80

T

biLSTM in:240 out:320

biLSTM in:640 out:320

down-sample

biLSTM in:640 out:320

down-sample

biLSTM in:640 out:320

down-sample

biLSTM in:640 out:320

Text

Hidden 640×T

blstm1

Hidden 640×T/2

blstm2

Hidden 640×T/4

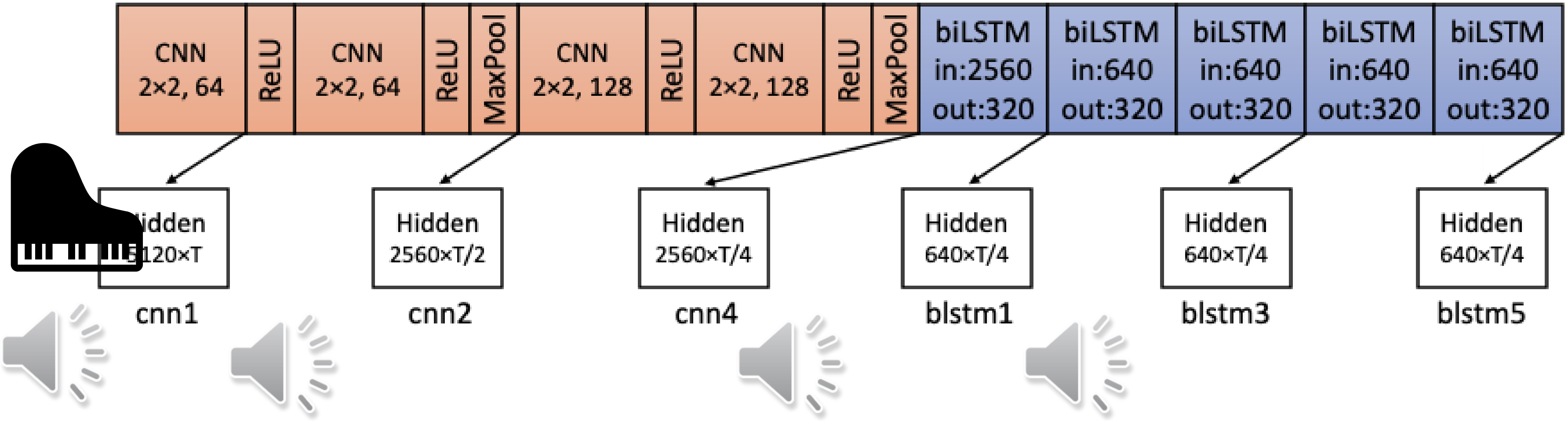
blstm3

Hidden 640×T/8

blstm4

Hidden 640×T/8

blstm5



CNN 2×2, 64

ReLU

CNN 2×2, 64

ReLU

MaxPool

CNN 2×2, 128

ReLU

CNN 2×2, 128

ReLU

MaxPool

biLSTM in:2560 out:320

biLSTM in:640 out:320

biLSTM in:640 out:320

biLSTM in:640 out:320

biLSTM in:640 out:320

idden 5120×T

cnn1

Hidden 2560×T/2

cnn2

Hidden 2560×T/4

cnn4

Hidden 640×T/4

blstm1

Hidden 640×T/4

blstm3

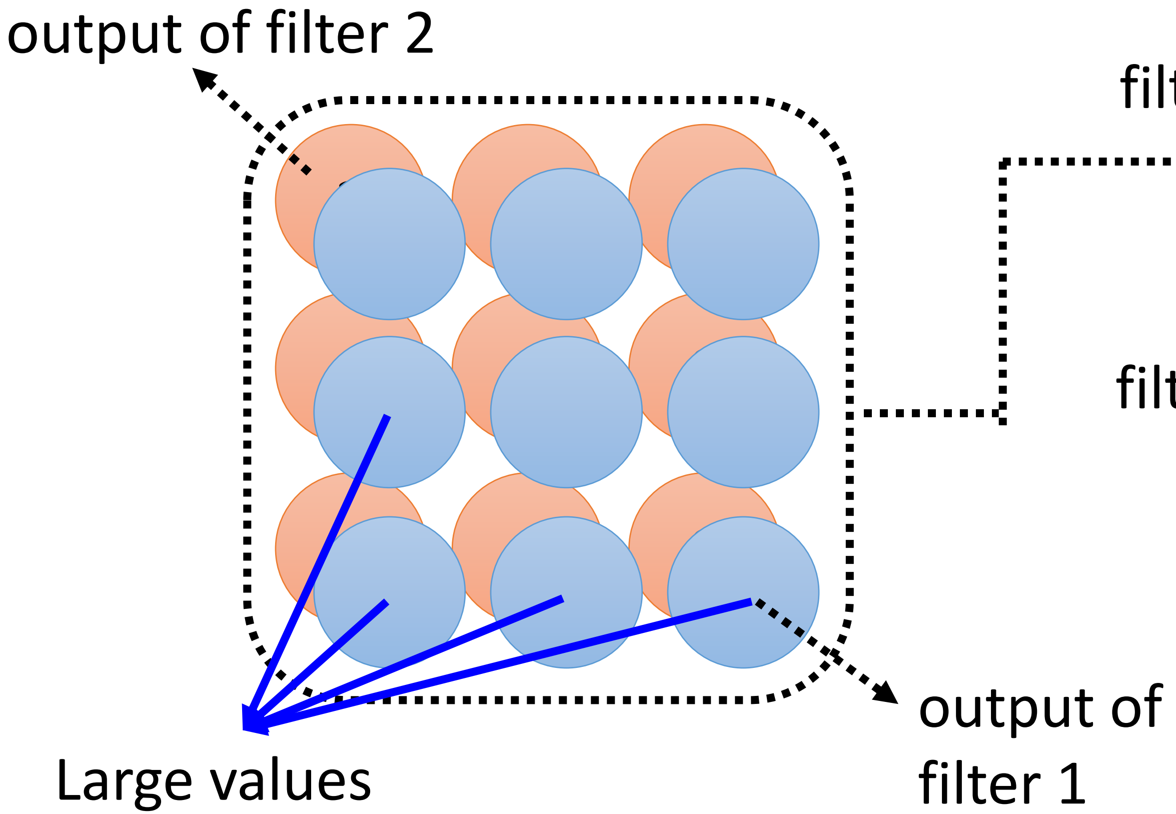
Hidden 640×T/4

blstm5

GLOBAL EXPLANATION: EXPLAIN THE WHOLE MODEL

Question: What does a “cat” look like?

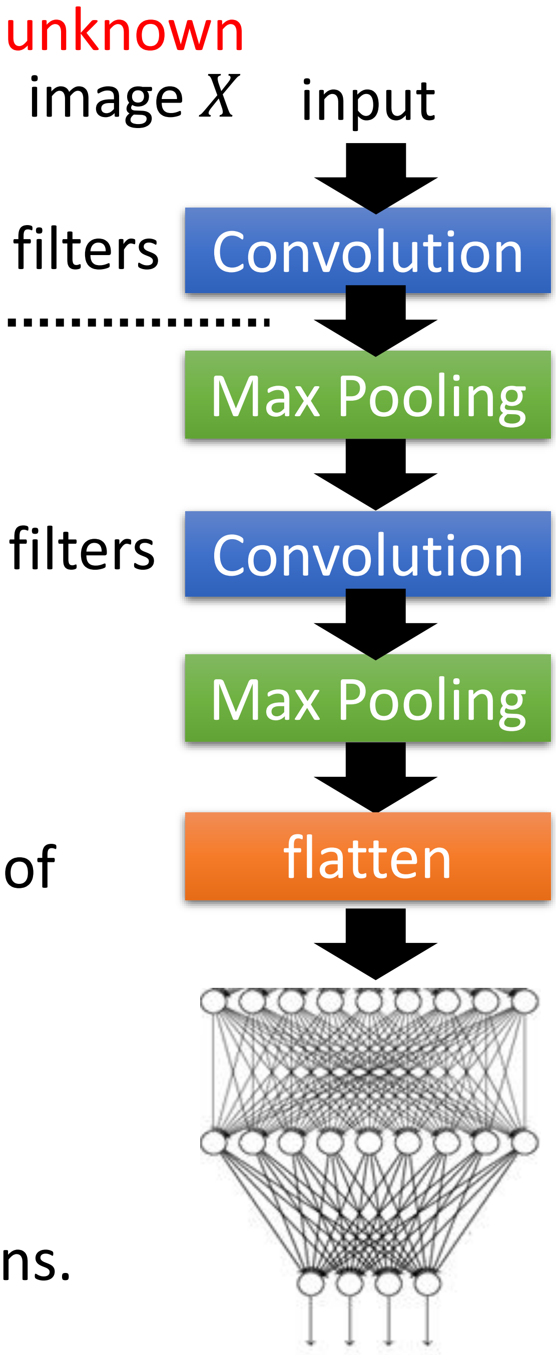
What does a filter detect?



output of filter 2

Large values

output of filter 1



unknown image X

input

filters

Convolution

Max Pooling

filters

Convolution

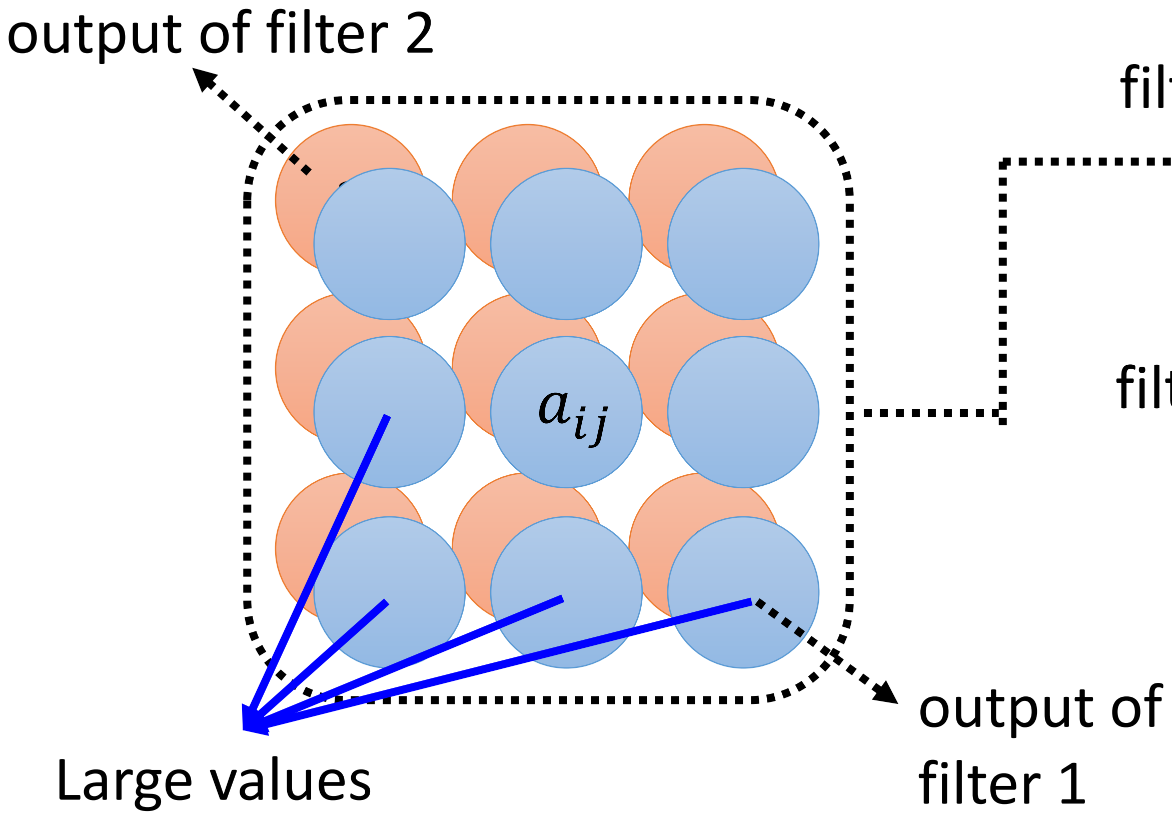
Max Pooling

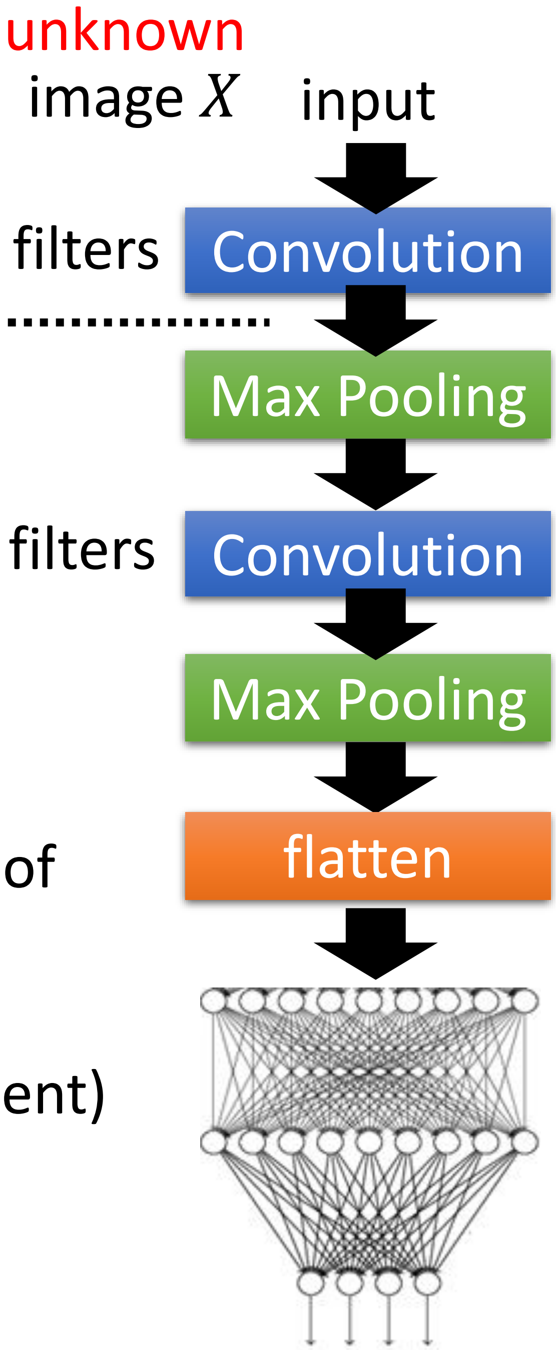
flatten

➡ Image X contains the patterns filter 1 can detect.

Let’s **create** an image including the patterns.

What does a filter detect?





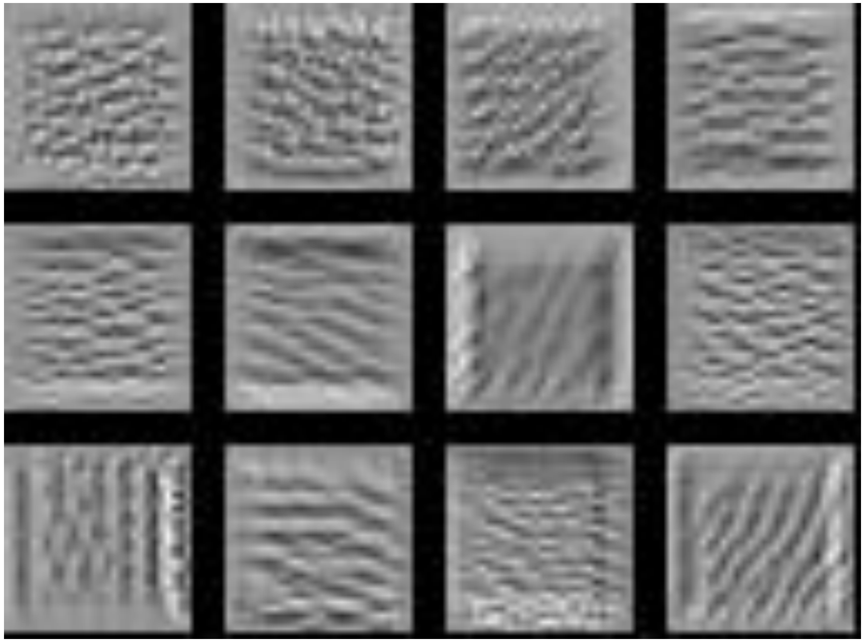
$$X^* = arg\max_X \sum_i \sum_j a_{ij}$$ (gradient ascent)

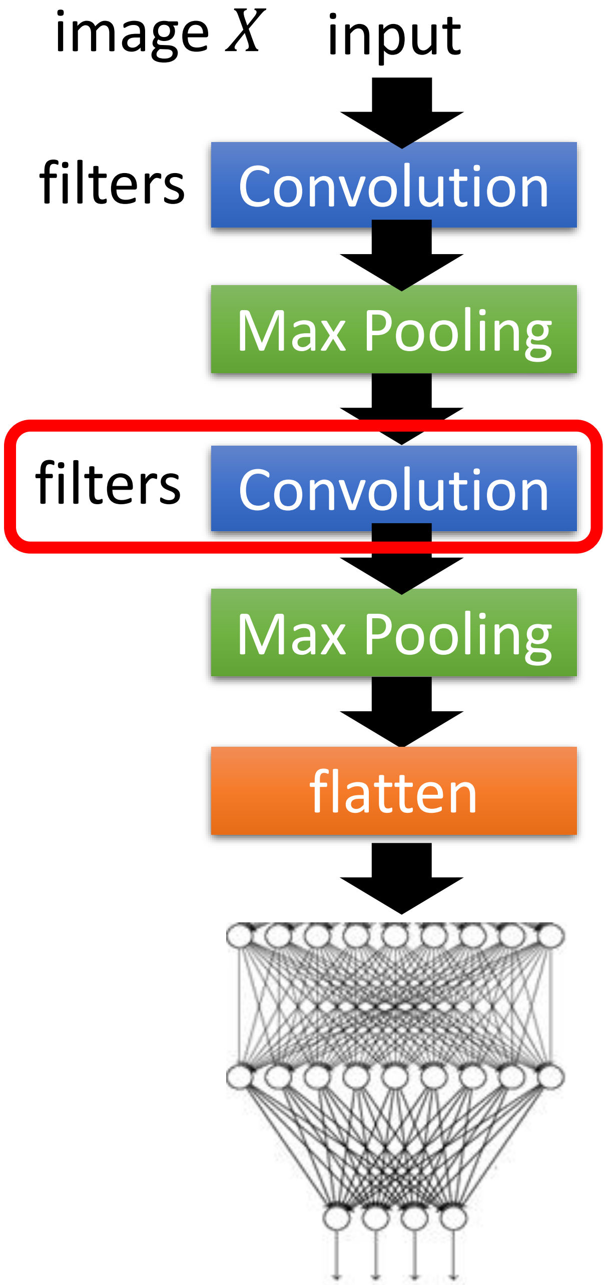
The image contains the patterns filter 1 can detect.

What does a filter detect?

E.g., Digit classifier

X^* for each filter



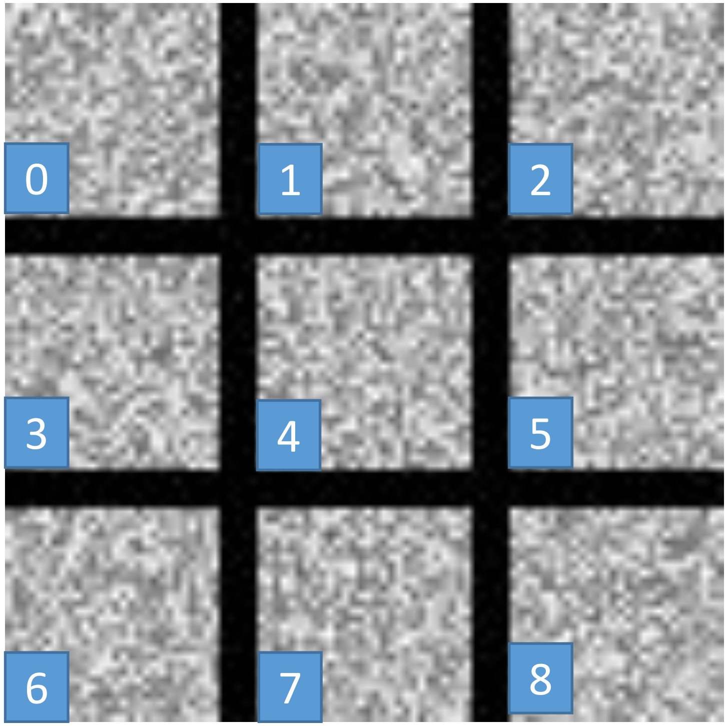


What does a digit look like for CNN?

E.g., Digit classifier

$$X^* = arg \max_X y_i$$

Can we see digits?



Surprise? Consider adversarial attack!

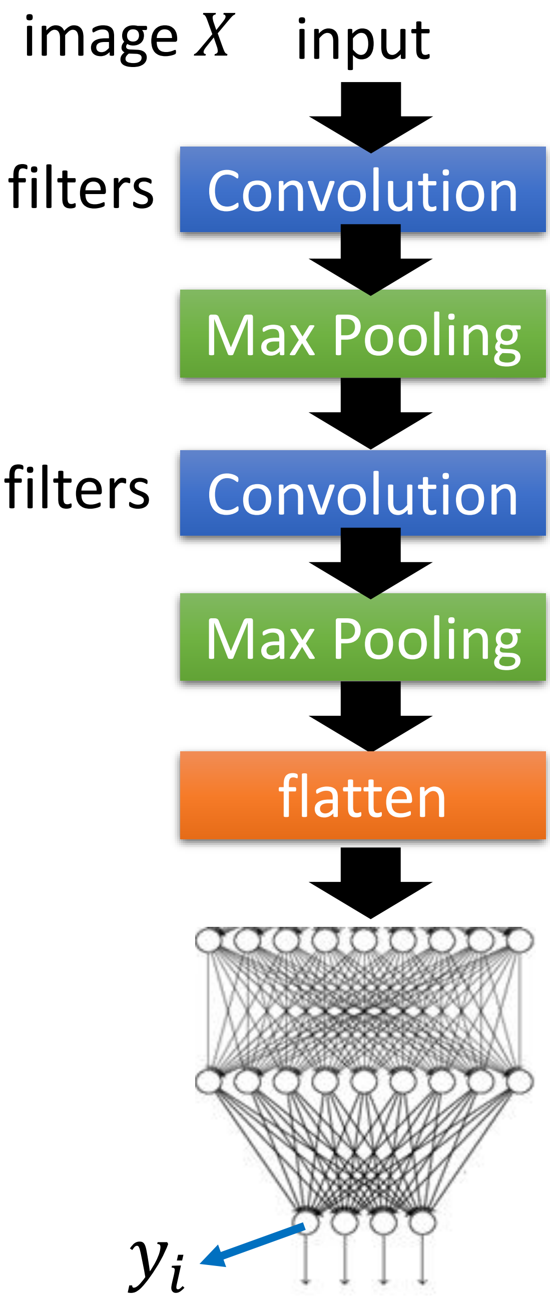


image X

input

filters

Convolution

Max Pooling

filters

Convolution

Max Pooling

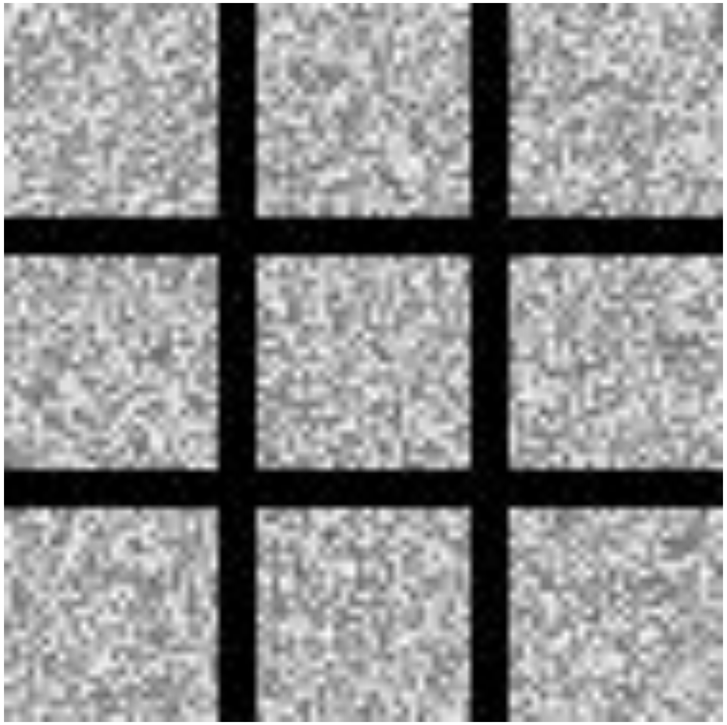
flatten

y_i

What does a digit look like for CNN?

Find the image that maximizes class probability

$$X^* = arg \max_X y_i$$

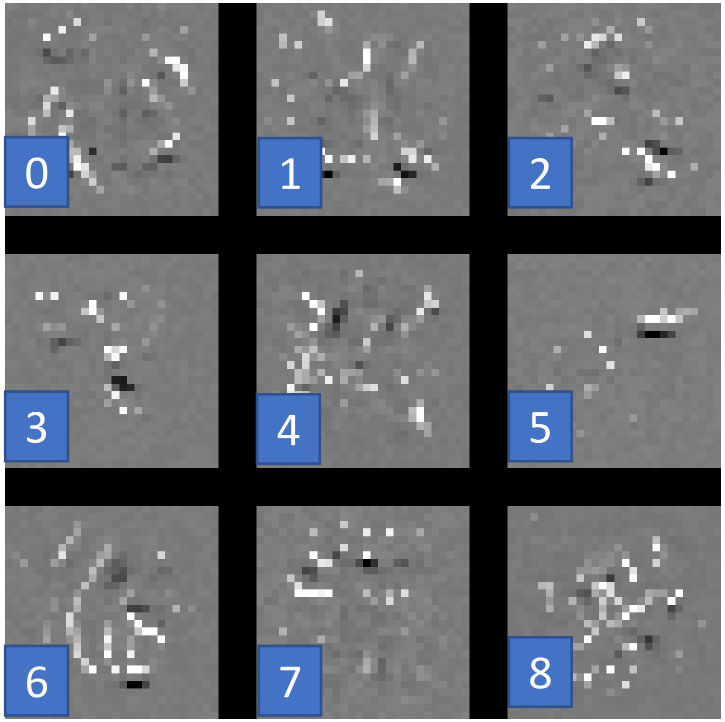


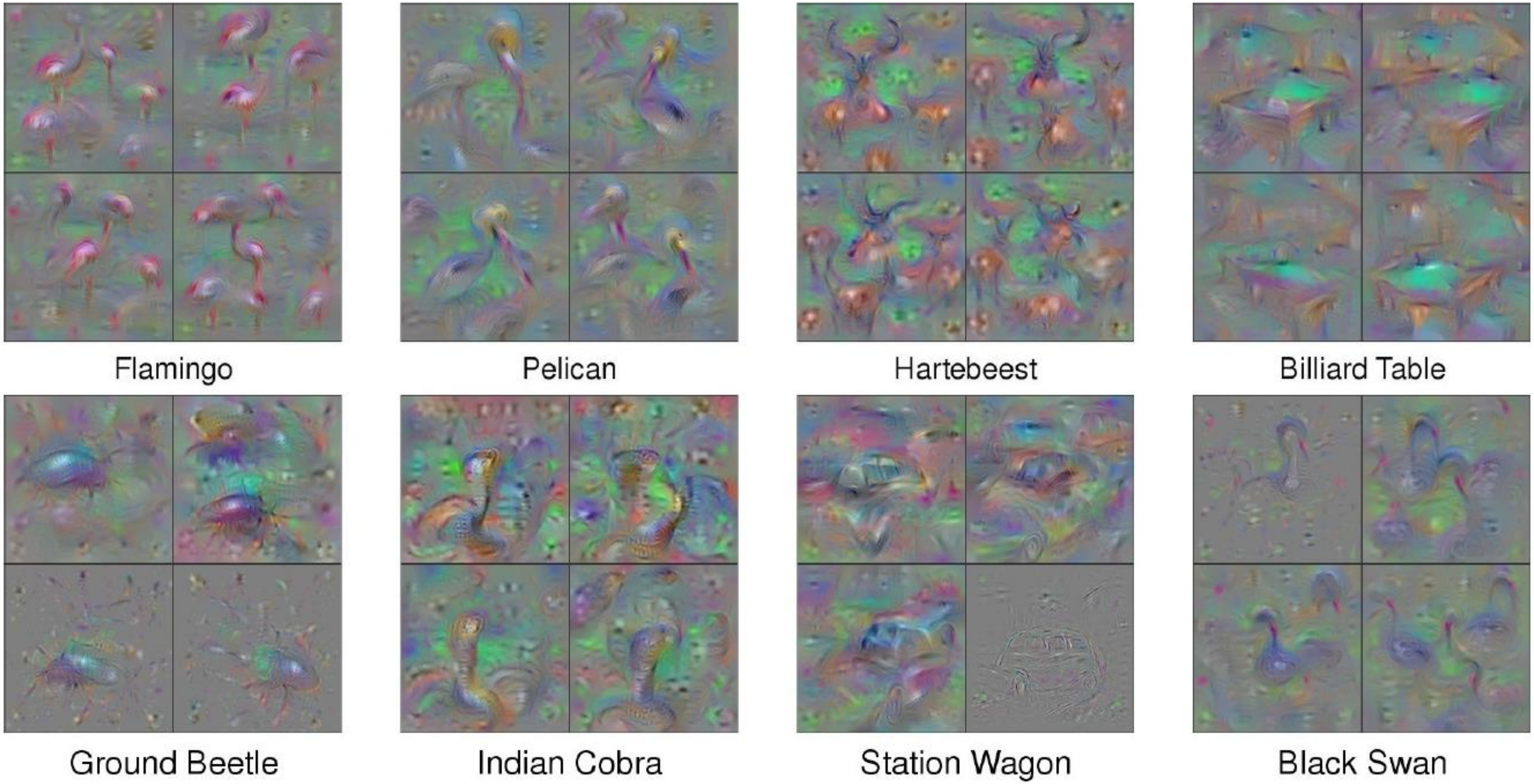
The image should looks like a digit.

$$X^* = arg \max_X y_i \quad + \quad \underline{R(X)}$$

$$R(X) = -\sum_{i,j} |X_{ij}|$$

How likely X is a digit





Flamingo

Pelican

Hartebeest

Billiard Table

Ground Beetle

Indian Cobra

Station Wagon

Black Swan

With several regularization terms, and hyperparameter tuning …..

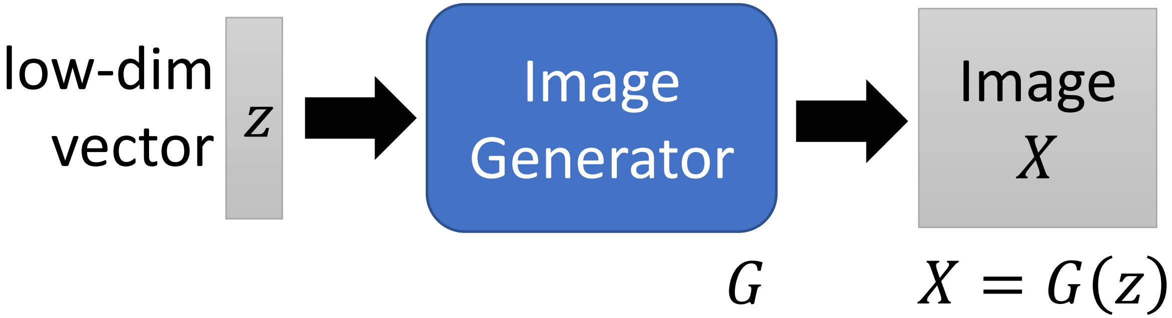
https://arxiv.org/abs/1506.06579

(Simplified Version)

Constraint from Generator

- Training a generator (by GAN, VAE, etc.)

Training Examples



low-dim vector

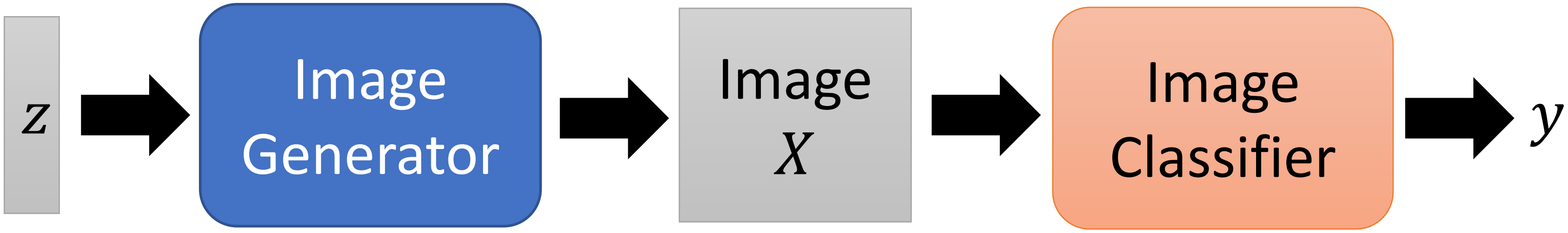
z

Image Generator

Image X

G

$X = G(z)$



z

Image Generator

Image X

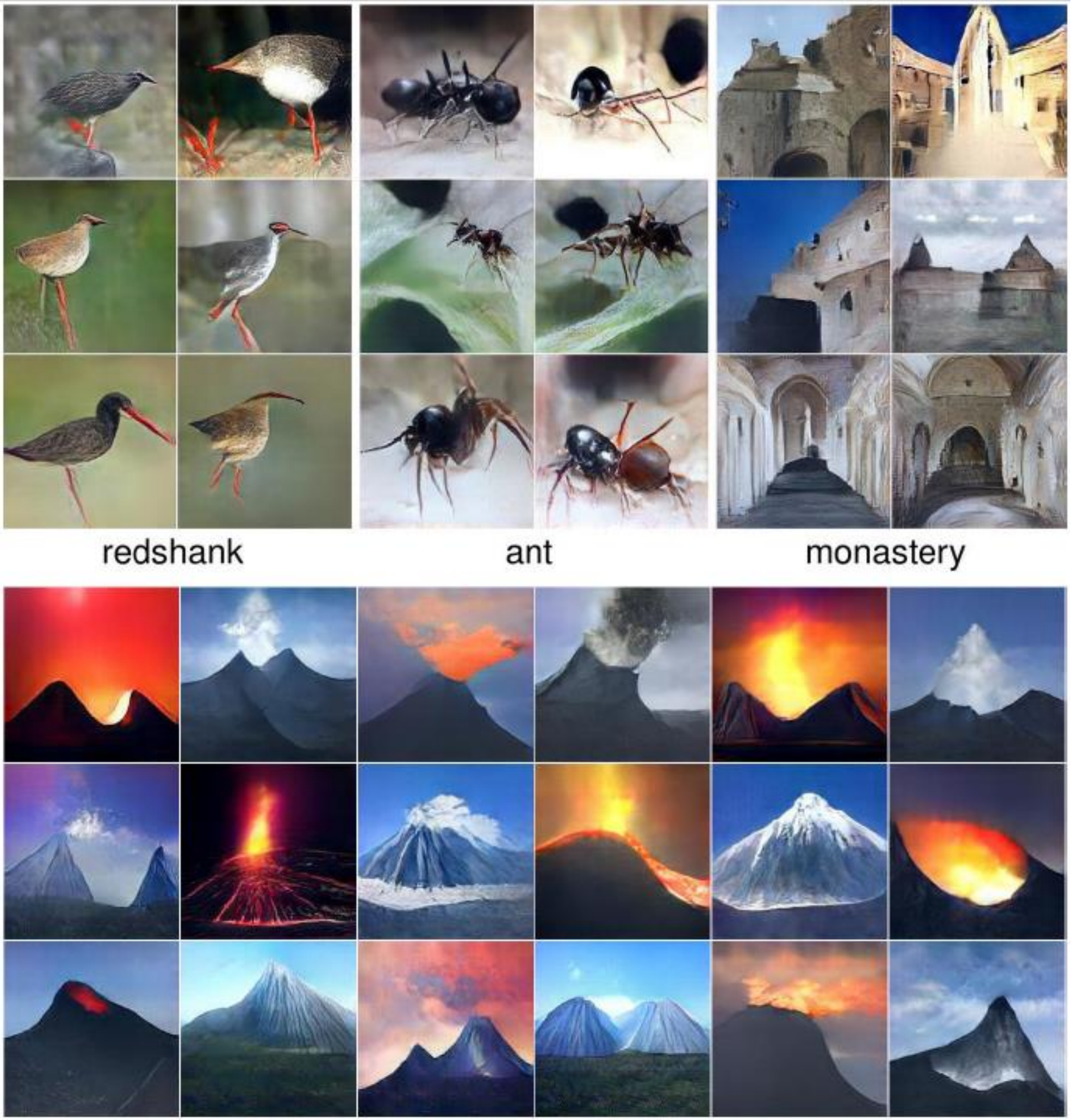
Image Classifier

y

$$X^* = arg \max_X y_i \Rightarrow z^* = arg \max_z y_i$$

Show image:

$$X^* = G(z^*)$$



redshank

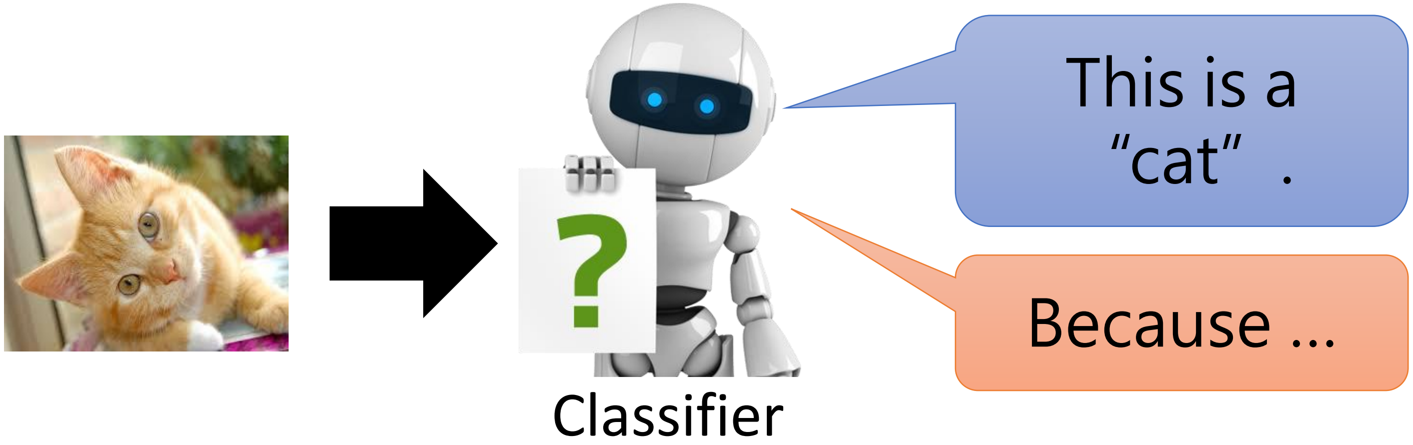
ant

monastery

volcano

https://arxiv.org/abs/
1612.00005

Concluding Remarks



This is a "cat" .

Because ...

Classifier

Local Explanation

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

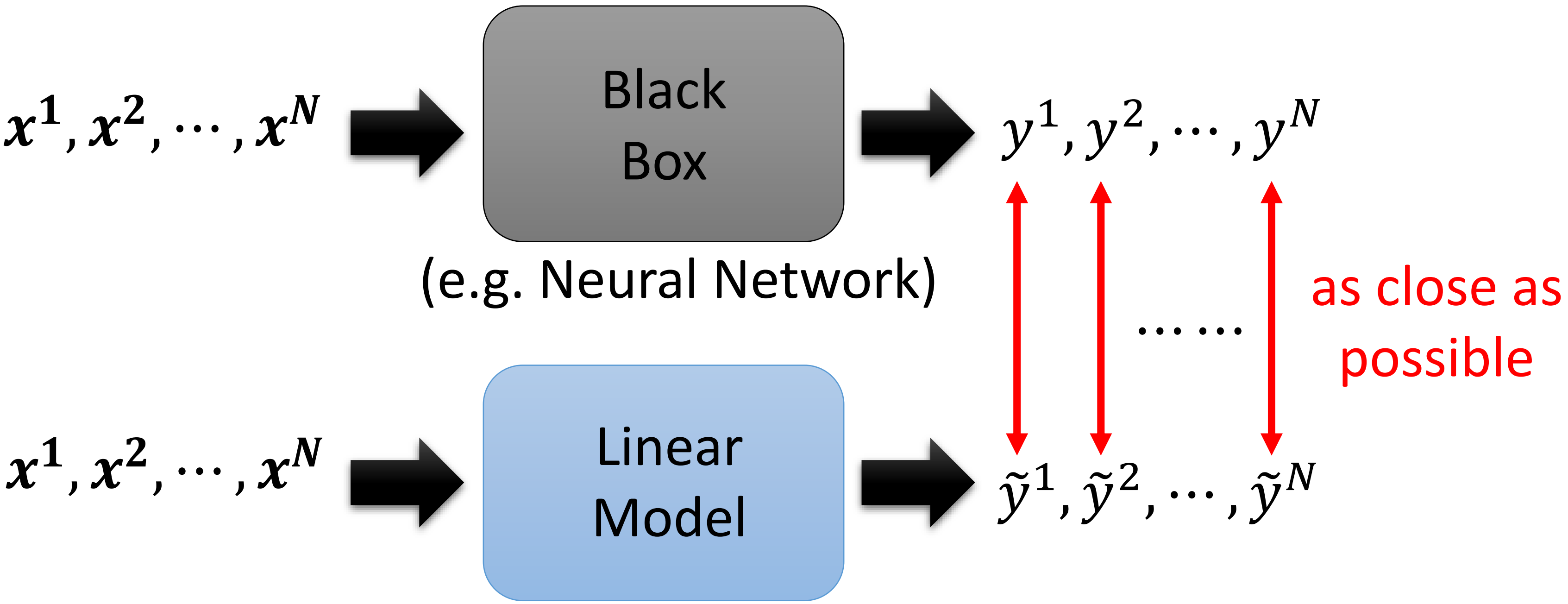
Global Explanation

What does a "cat" look like?

(not referred to a specific image)

Outlook

Using an interpretable model to mimic the behavior of an uninterpretable model.



$\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^N$

Black Box

$y^1, y^2, \cdots, y^N$

(e.g. Neural Network)

…… ……

as close as possible

$\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^N$

Linear Model

$\tilde{y}^1, \tilde{y}^2, \cdots, \tilde{y}^N$

Local Interpretable Model-Agnostic Explanations (LIME)

https://youtu.be/K1mWgthGS-A
https://youtu.be/OjqIVSwly4k