

李宏毅 (Hung-yi Lee) · HYLEE | Machine Learning (2021)

HYLEE(2021) · 课程资料包 @ShowMeAI



视频

课件

笔记

代码

中英双语字幕

一键打包下载

官方笔记翻译

作业项目解析



视频 · B 站 [扫码或点击链接]

<https://www.bilibili.com/video/BV1fM4y137M4>



课件 & 代码 · 博客 [扫码或点击链接]

<http://blog.showmeai.tech/ntu-hylee-ml>

机器学习

Auto-encoder

生成式对抗网络

学习率

深度学习

卷积神经网络

GAN

自监督

自注意力机制

批次标准化

神经网络压缩

强化学习

元学习

Transformer

Awesome AI Courses Notes Cheatsheets 是 [ShowMeAI](#) 资料库的分支系列，覆盖最具知名度的 **TOP50+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

点击课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

机器学习	深度学习	自然语言处理	计算机视觉
Stanford · CS229	Stanford · CS230	Stanford · CS224n	Stanford · CS231n
# Awesome AI Courses Notes Cheatsheets · 持续更新中			
知识图谱	图机器学习	深度强化学习	自动驾驶
Stanford · CS520	Stanford · CS224W	UCBerkeley · CS285	MIT · 6.S094



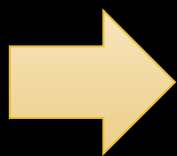
微信公众号

资料下载方式 2: 扫码点击**底部菜单栏**

称为 **AI 内容创作者**? 回复 [添砖加瓦]

EXPLAINABLE MACHINE LEARNING

Hung-yi Lee 李宏毅



This is a
"cat" .

Because ...

Why we need Explainable ML?

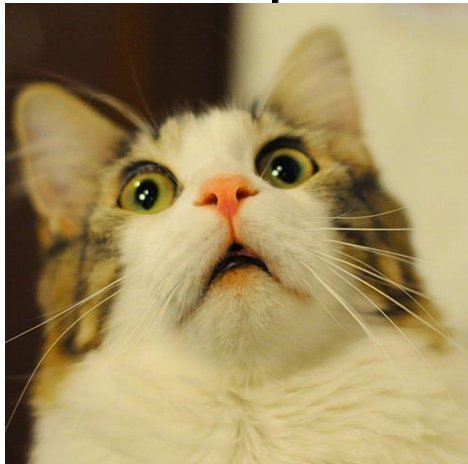
- Correct answers \neq Intelligent



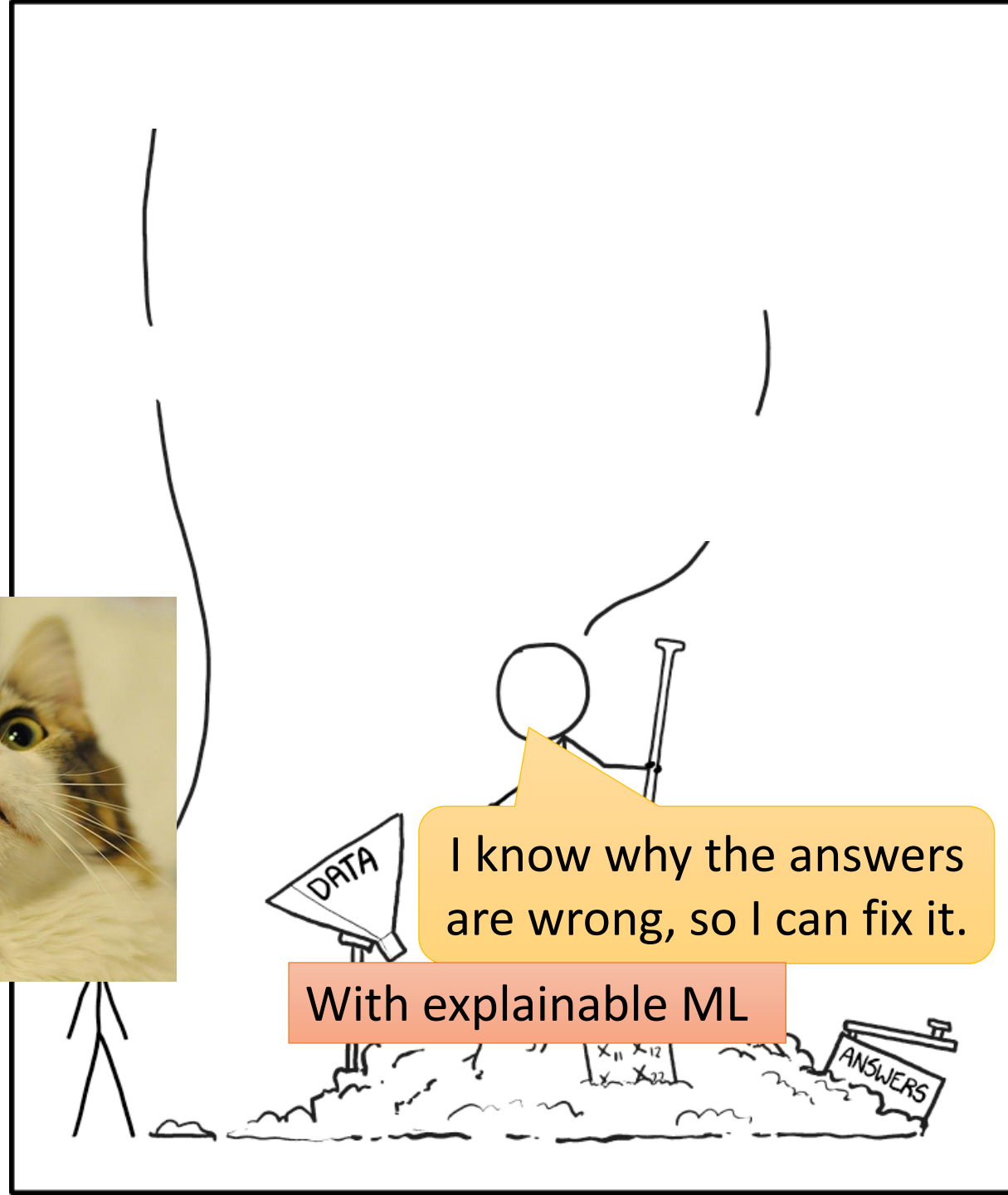
Why we need Explainable ML?

- 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.



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



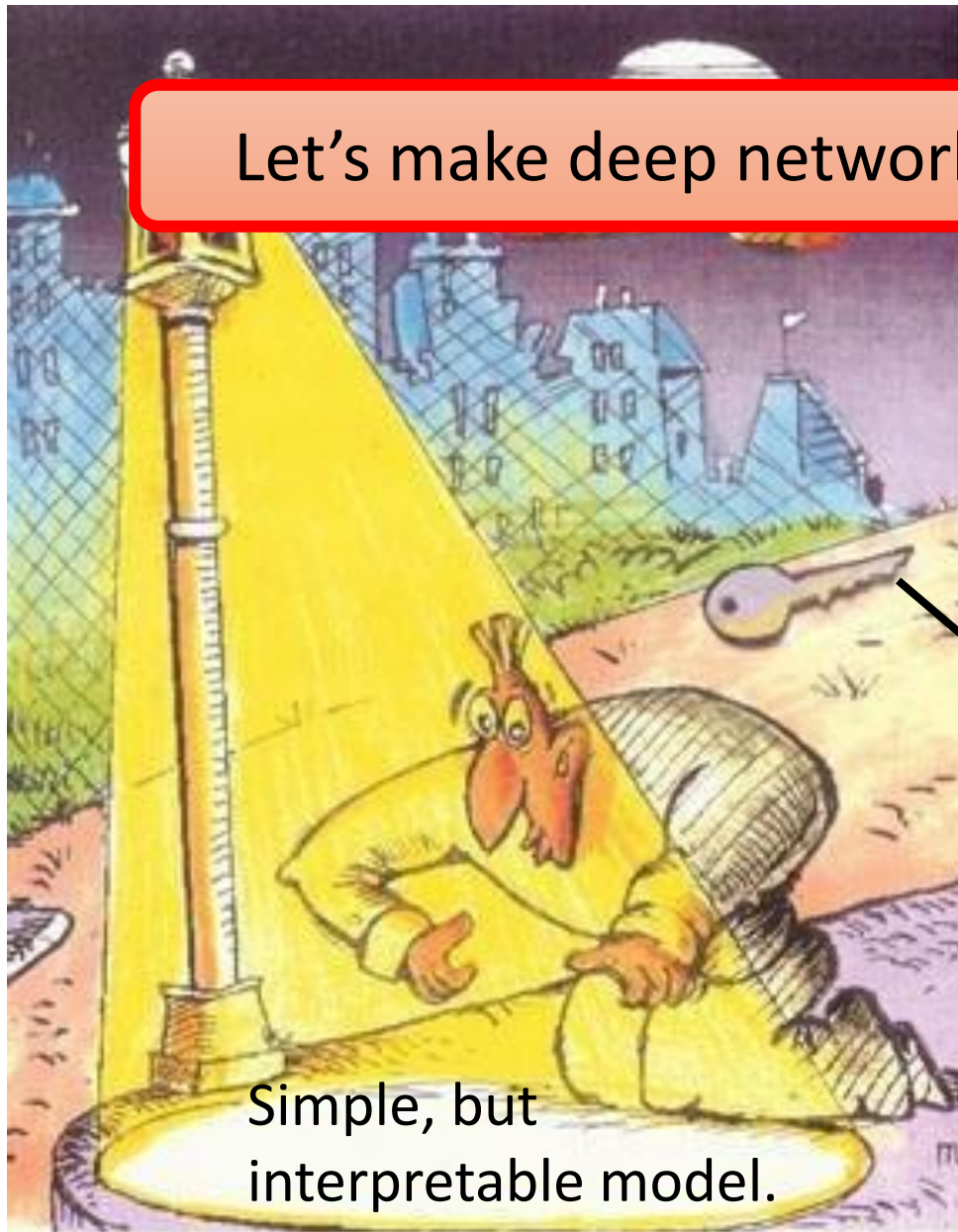
Interpretable v.s. Powerful

- 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 interpret. 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.



Simple, but
interpretable model.

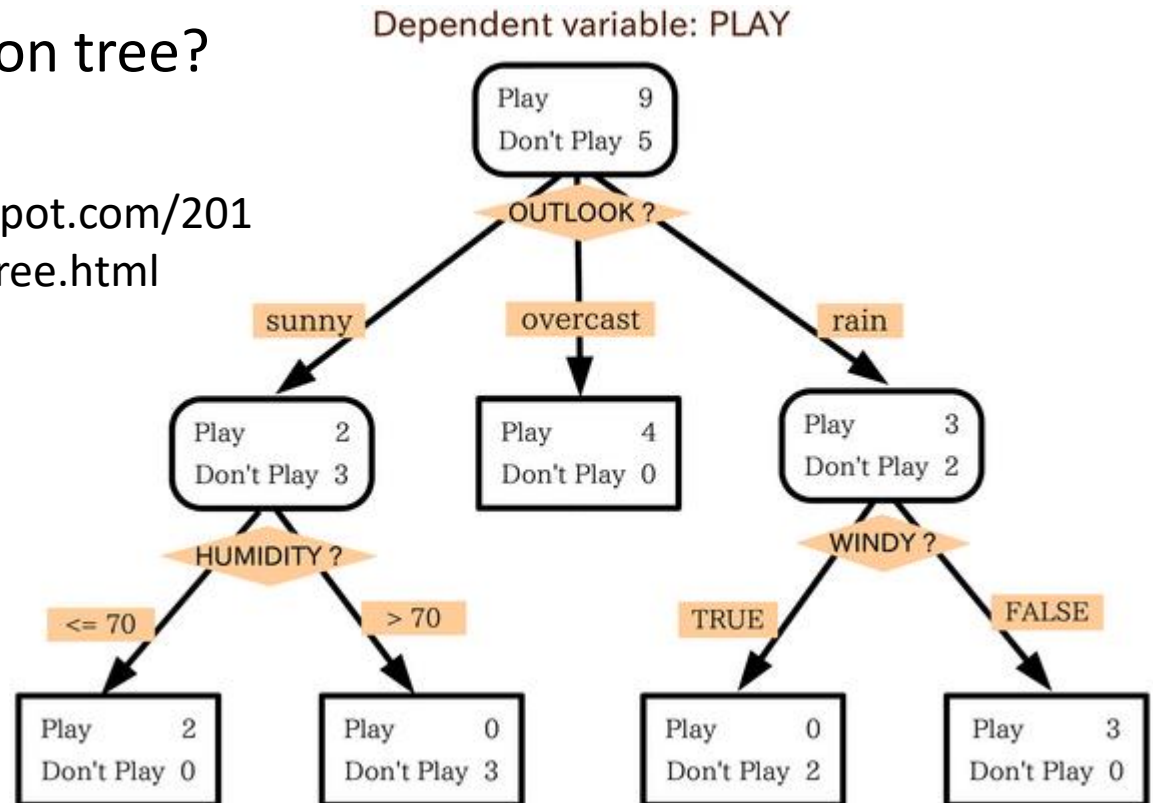
Powerful model

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>

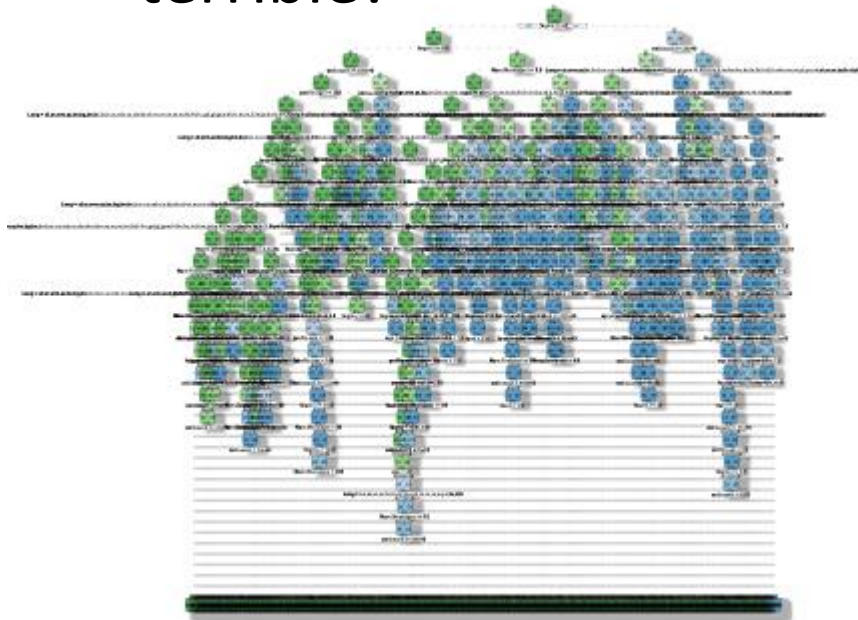


A photograph of a forest path that splits into two directions. The path is covered in fallen yellow and brown leaves. The forest is dense with green trees and foliage. The text "Decision tree is all you need!?" is overlaid in white, sans-serif font in the center of the image.

Decision tree is all you need!?

Interpretable v.s. Powerful

- A tree can still be terrible!



Rattle 2016-Aug-18 16:15:42 sklisarov

<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

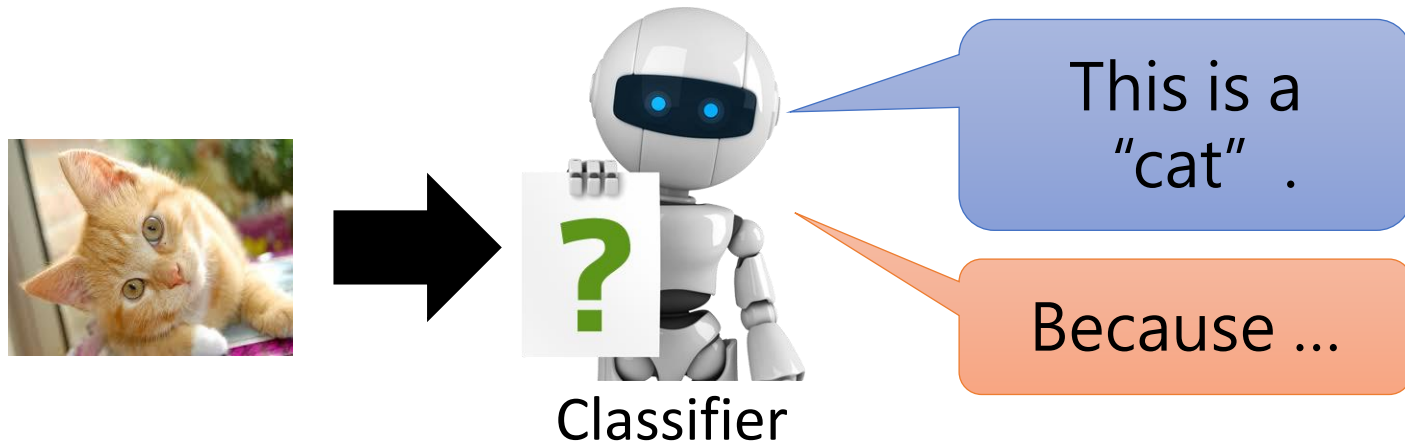
“Excuse me, I have 5 pages. May I use the Xerox machine,
because I have to make copies?”

93% accept

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

(my two cents)

Explainable ML



Local Explanation

Why do you think this image is a cat?

Global Explanation

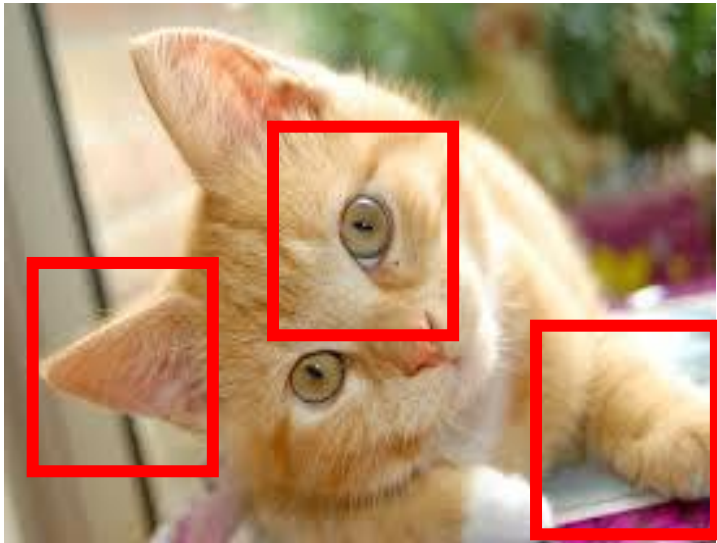
What does a “cat” look like?

(not referred to a specific image)

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 $x \longrightarrow$ Image, text, etc.

Components:

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

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

- Removing or modifying the components
 - Large decision change
- ➡** Important component



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

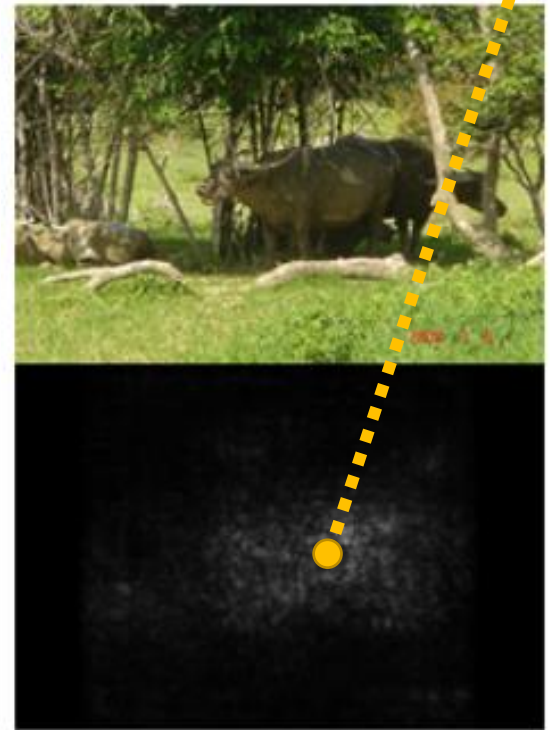
$$\{x_1, \dots, x_n, \dots, x_N\} \longrightarrow \{x_1, \dots, x_n + \Delta x, \dots, x_N\}$$

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

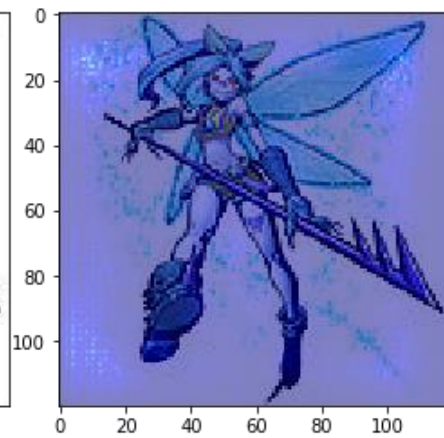
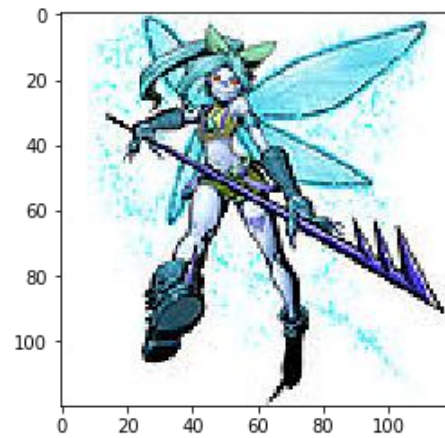
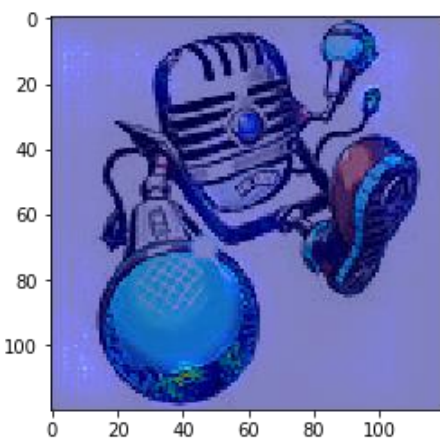
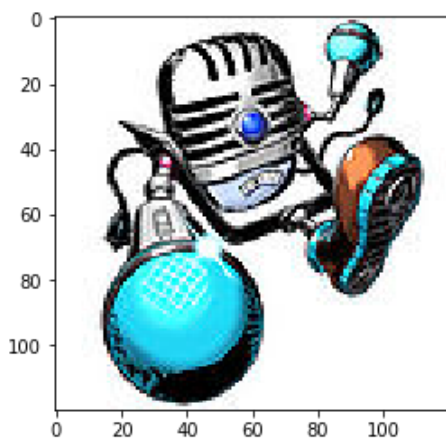
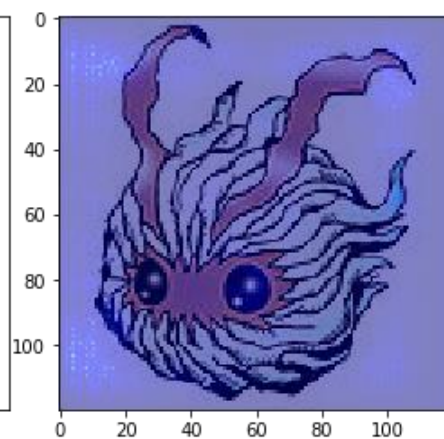
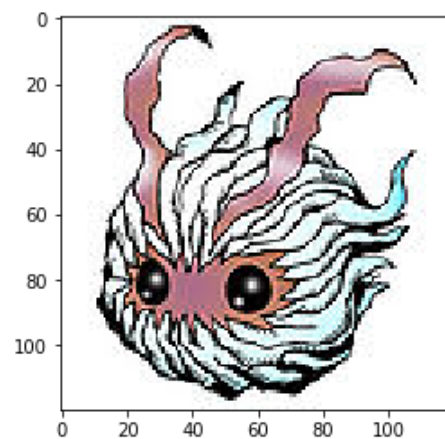
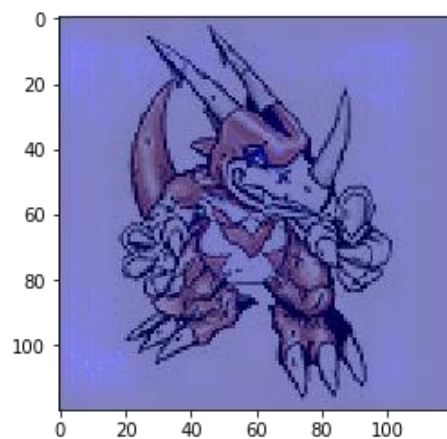
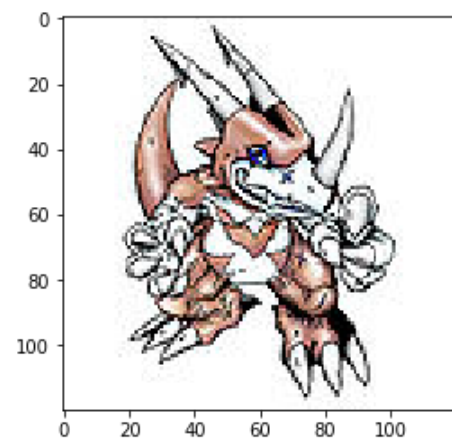
```
model = Sequential()  
model.add(Conv2D(32, (3, 3), padding='same', input_shape=(120,120,3)))  
model.add(Activation('relu'))  
model.add(Conv2D(32, (3, 3)))  
model.add(Activation('relu'))  
model.add(MaxPooling2D(pool_size=(2, 2)))  
  
model.add(Conv2D(64, (3, 3), padding='same'))  
model.add(Activation('relu'))  
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%

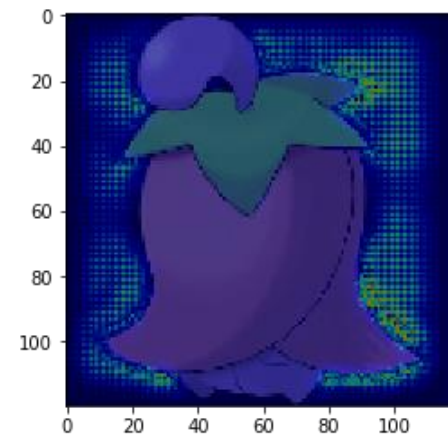
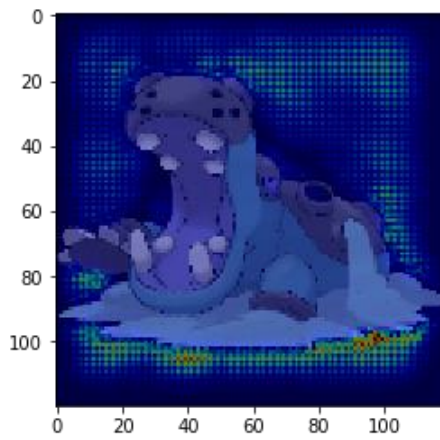
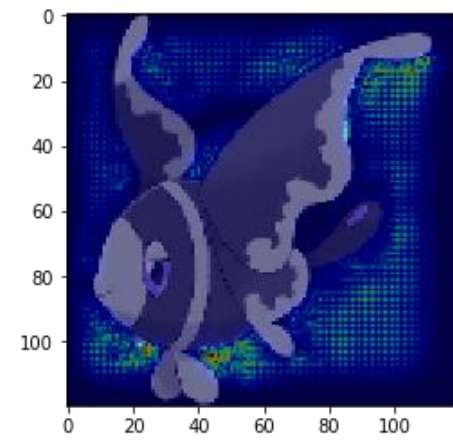
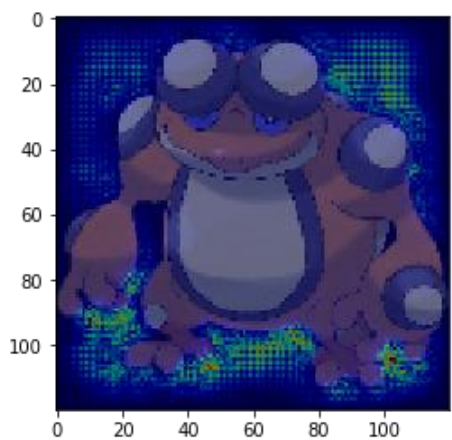
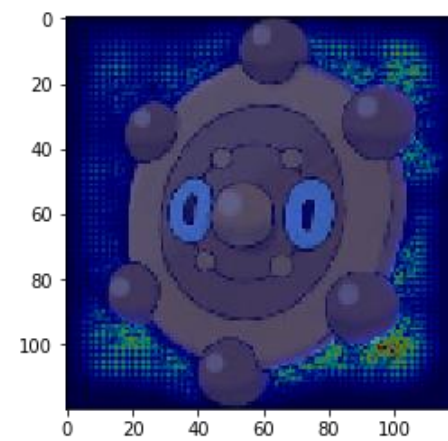
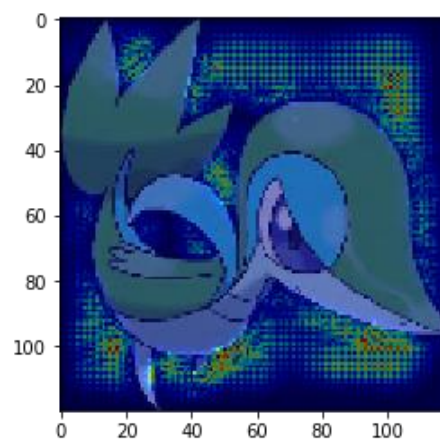
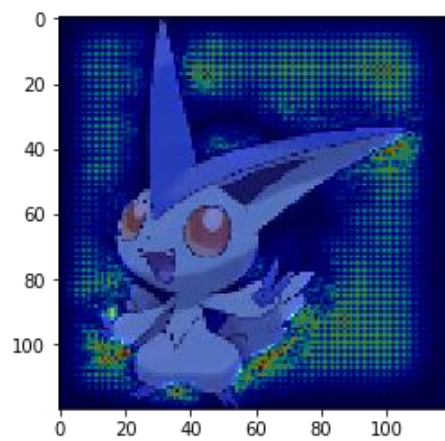
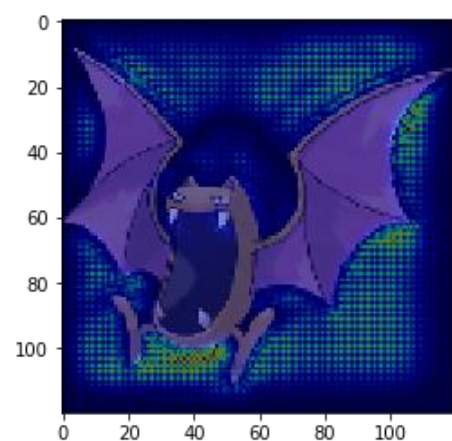
Testing Accuracy: 98.4%

Amazing!!!!!!

Saliency Map



Saliency Map



What Happened?

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



png files have transparent background

loading the files

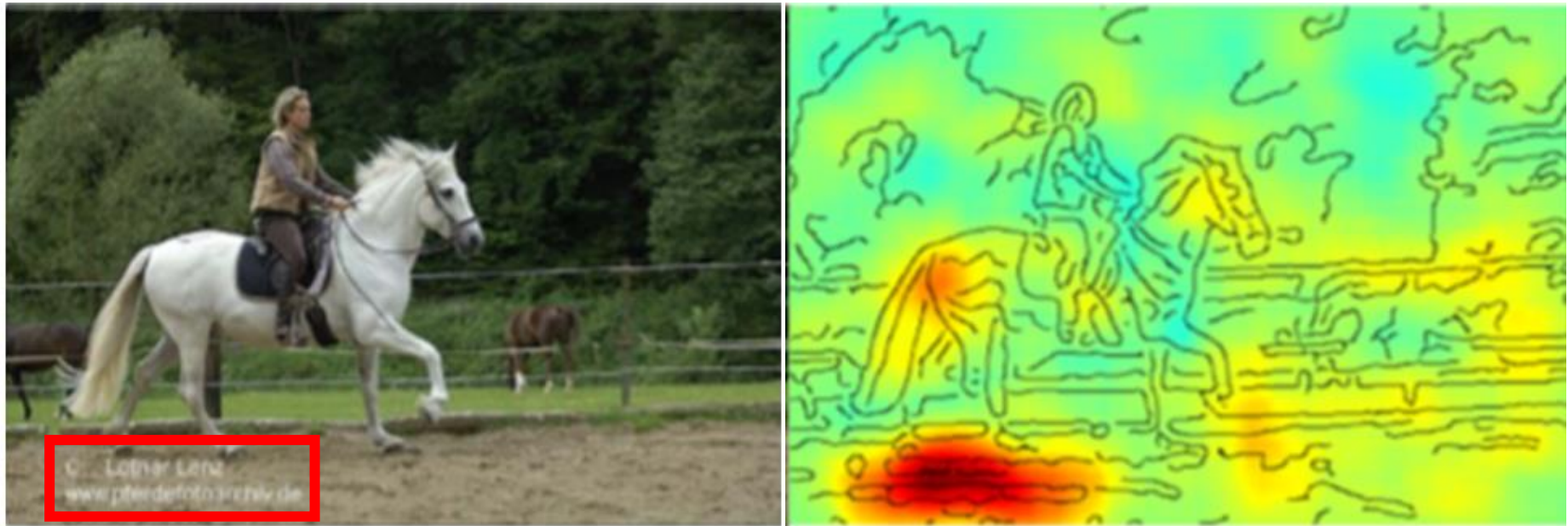


transparent background becomes black

Machine discriminates Pokémon and Digimon based on the background colors.

More Examples ...

- PASCAL VOC 2007 data set

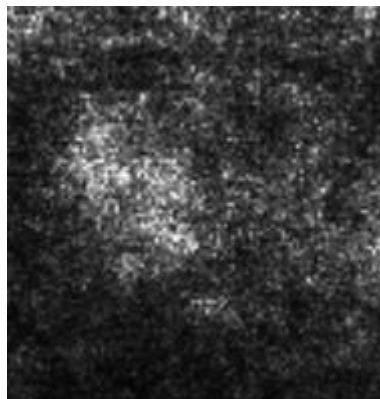


This slide is from: GCPR 2017 Tutorial — W. Samek & K.-R. Müller

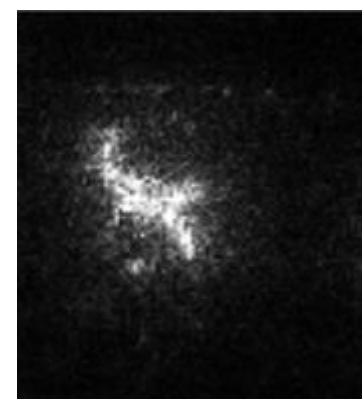
Limitation: Noisy Gradient



Gazelle
(瞪羚)



Typical



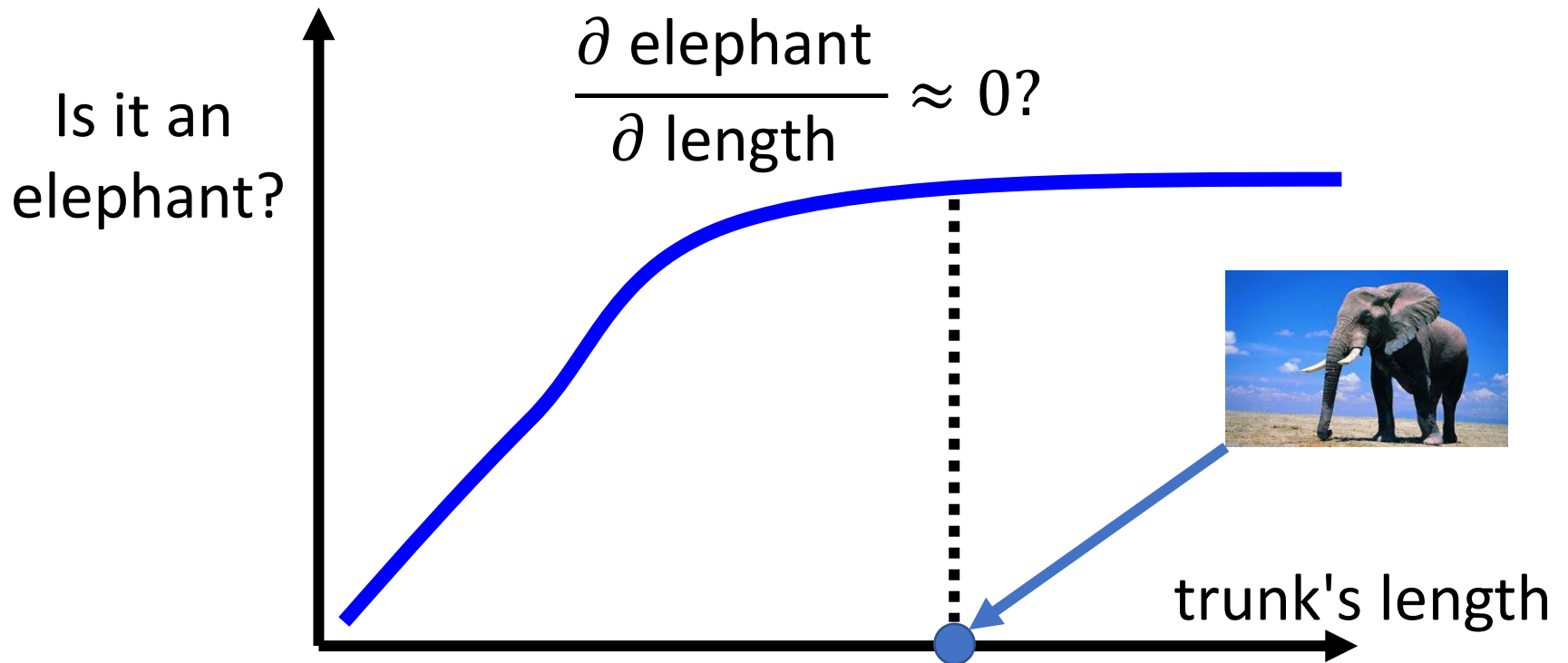
SmoothGrad

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

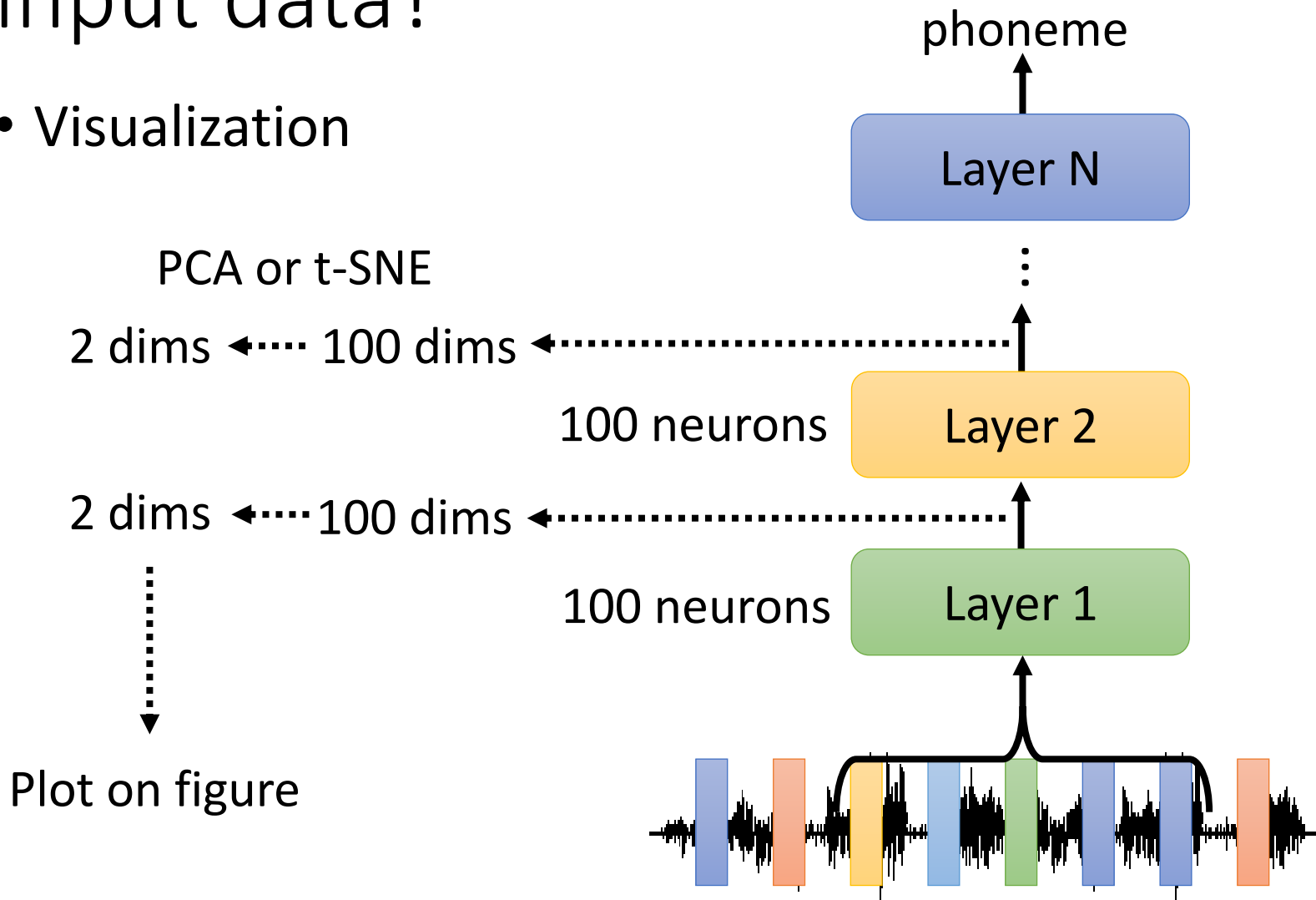


Alternative: Integrated gradient (IG)

<https://arxiv.org/abs/1611.02639>

How a network processes the input data?

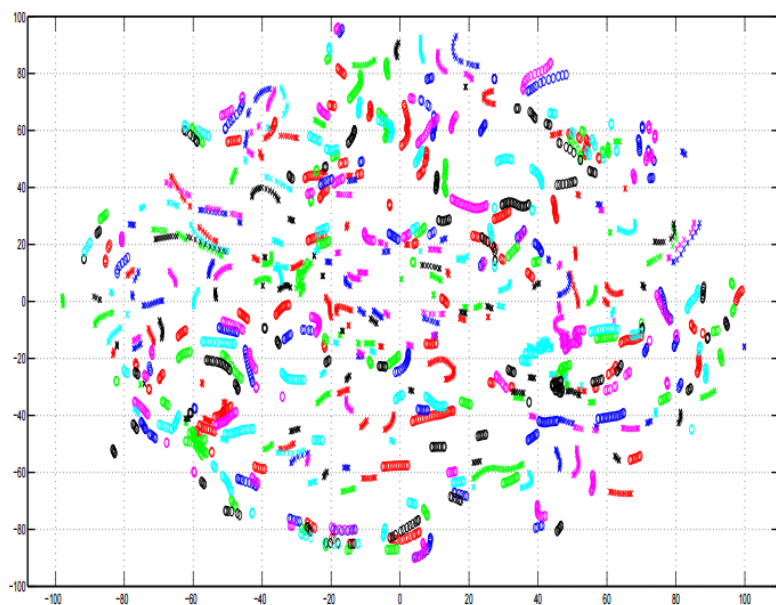
- Visualization



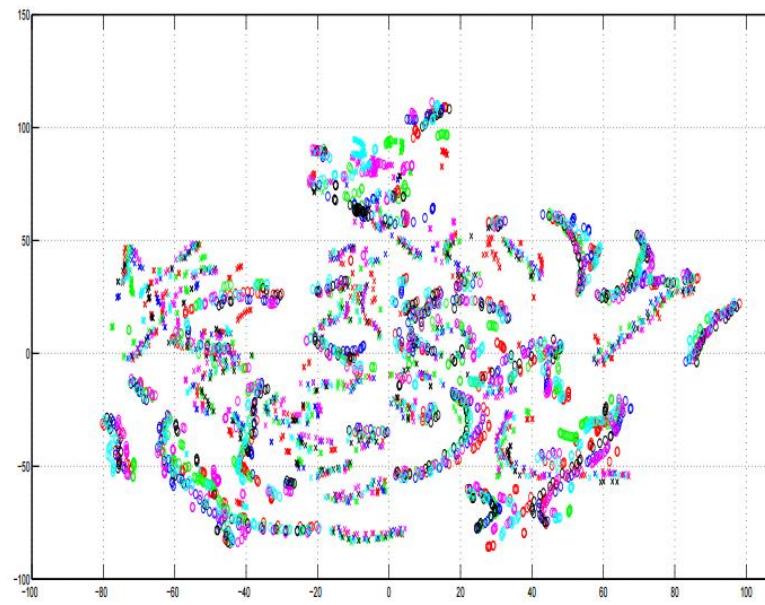
How a network processes the input data?

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

- Visualization
Colors: speakers



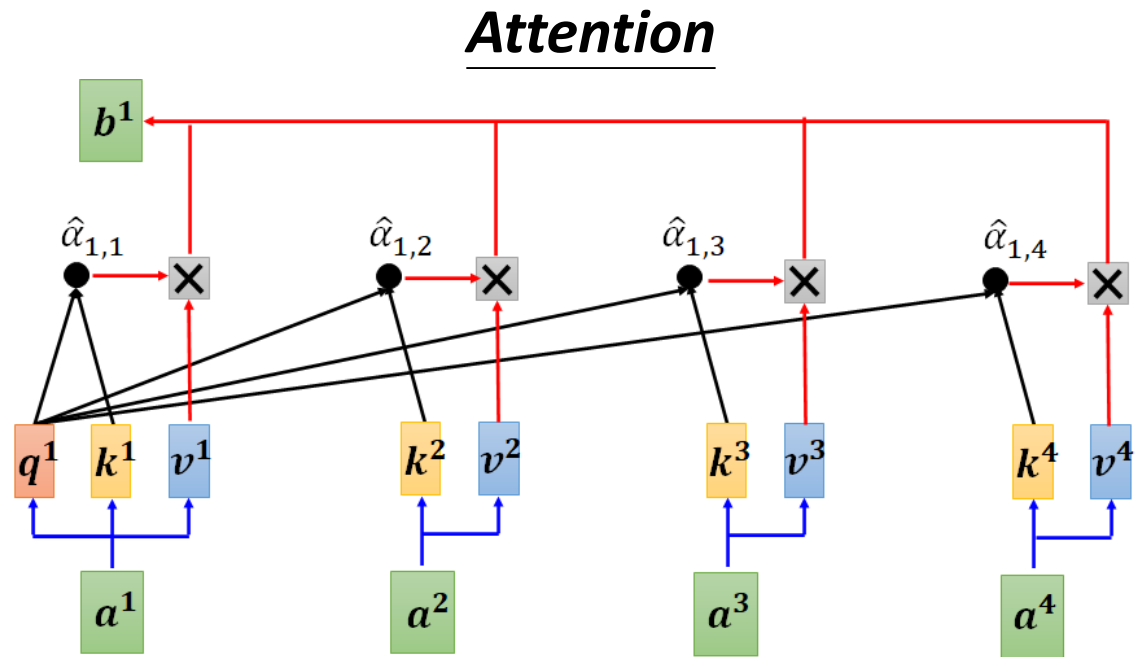
Input Acoustic Feature (MFCC)



8-th Hidden Layer

How a network processes the input data?

- Visualization



Attention is not Explanation

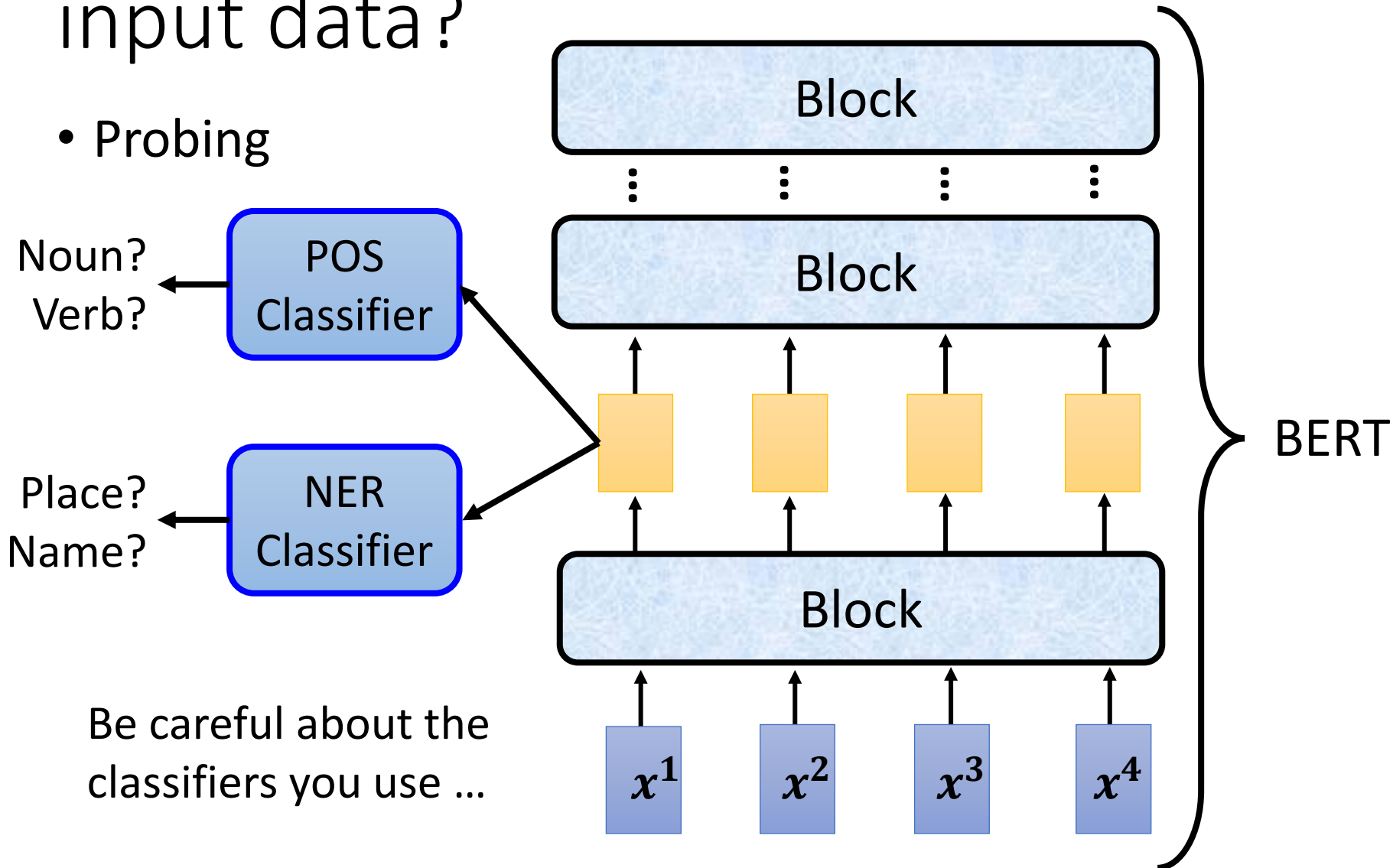
<https://arxiv.org/abs/1902.10186>

Attention is not not Explanation

<https://arxiv.org/abs/1908.04626>

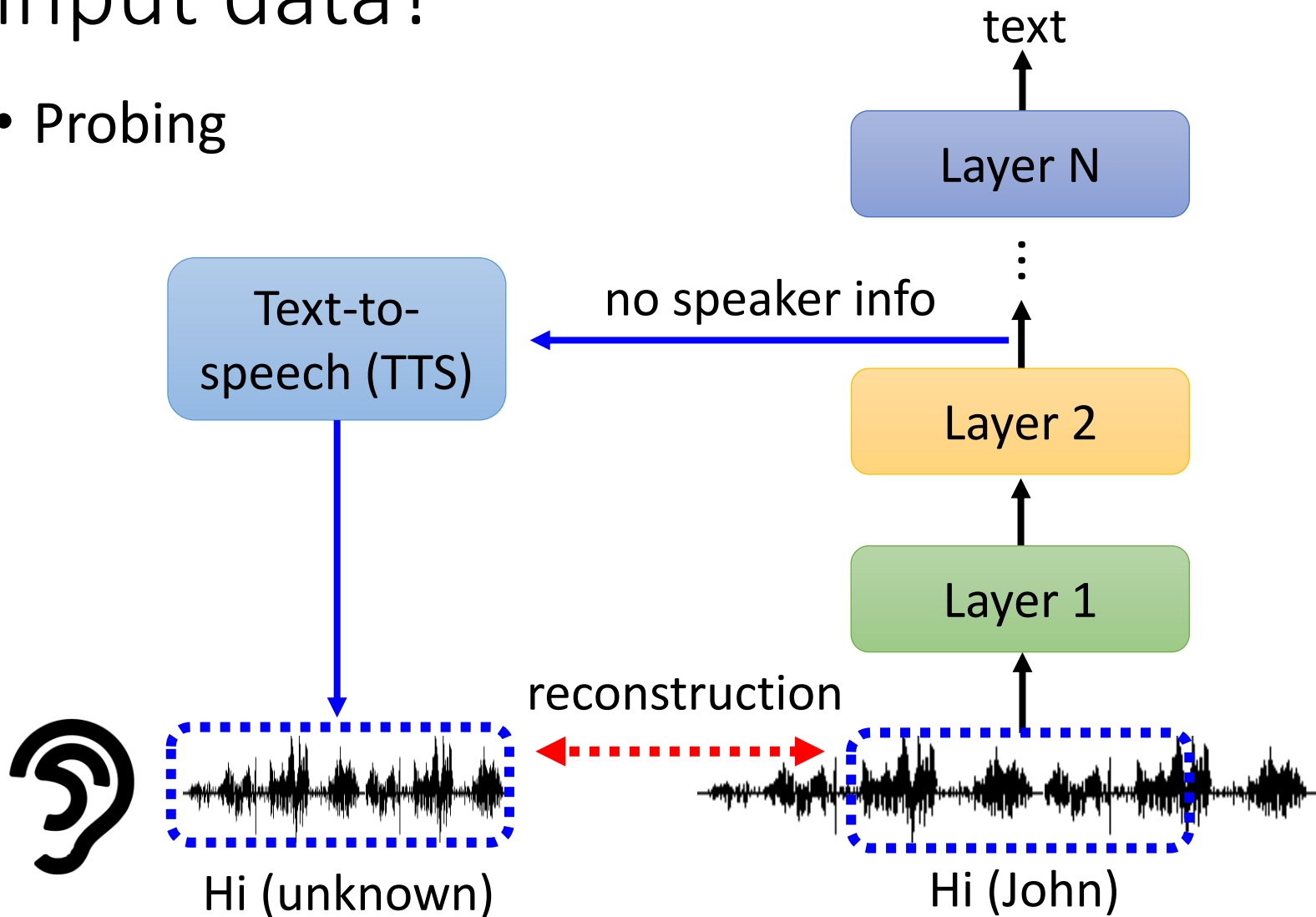
How a network processes the input data?

- Probing



How a network processes the input data?

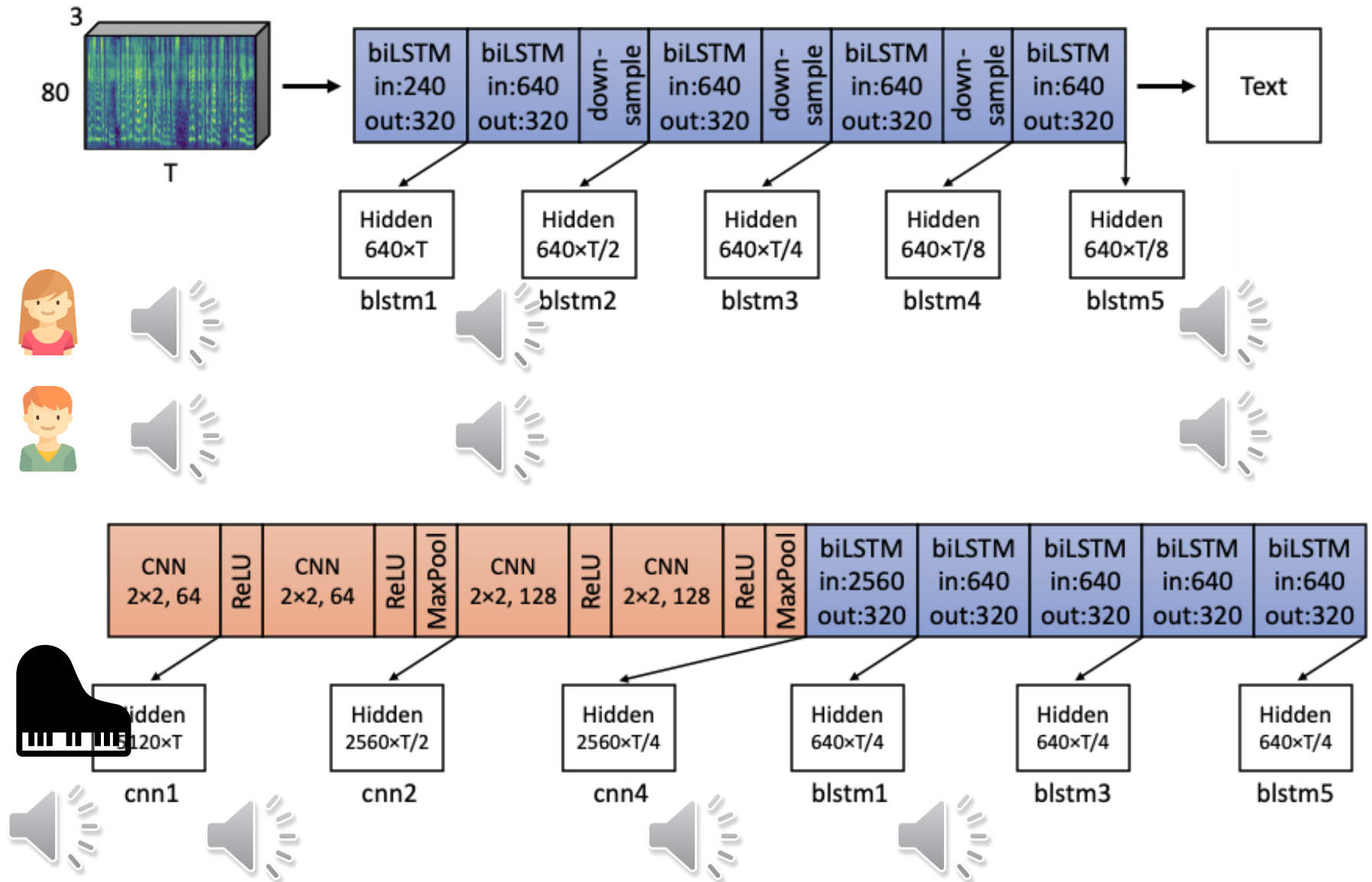
- Probing



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>





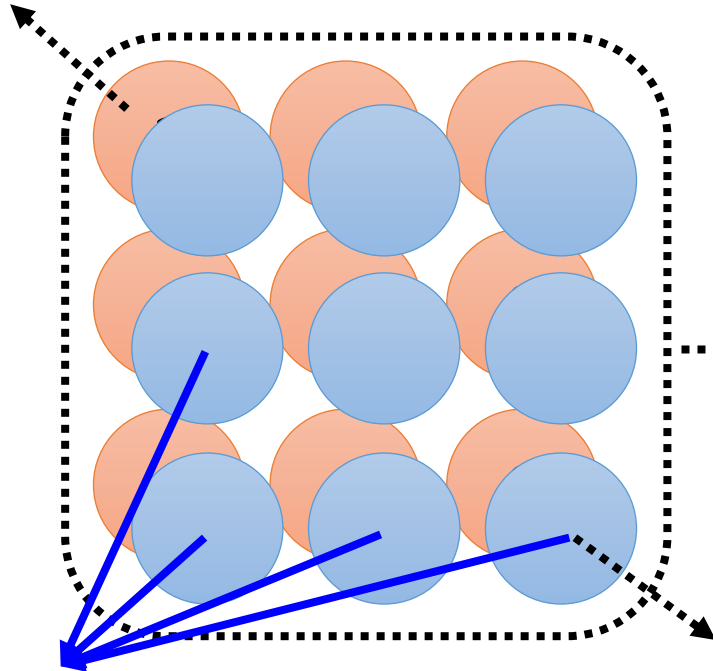
GLOBAL EXPLANATION: EXPLAIN THE WHOLE MODEL

Question: What does a “cat” look like?



What does a filter detect?

output of filter 2



Large values

output of
filter 1

➡ Image X contains the patterns
filter 1 can detect.

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

unknown

image X

input

filters

Convolution

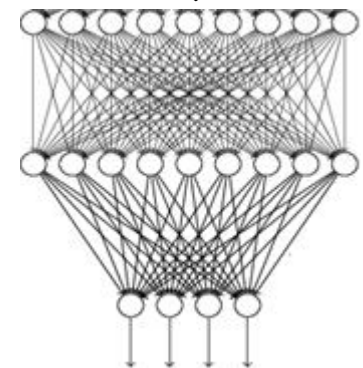
Max Pooling

filters

Convolution

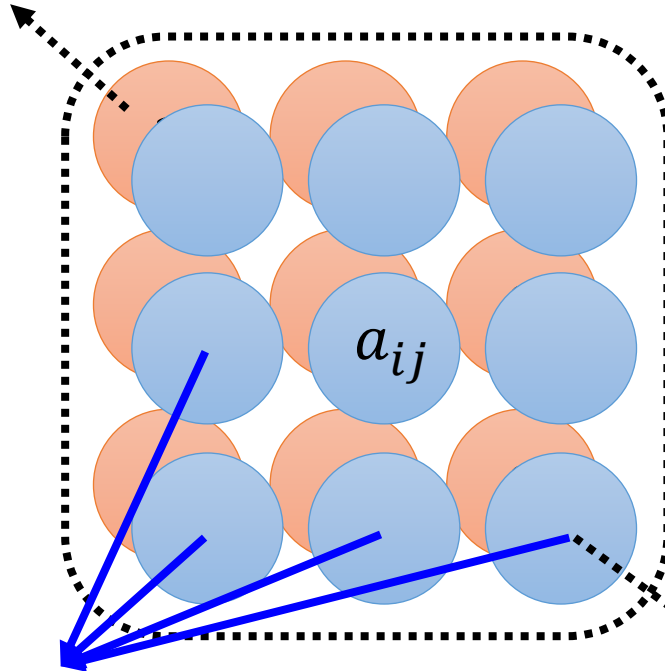
Max Pooling

flatten



What does a filter detect?

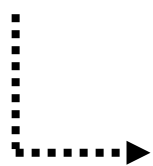
output of filter 2



Large values

output of
filter 1

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



The image contains the patterns
filter 1 can detect.

unknown

image X

input

filters

Convolution

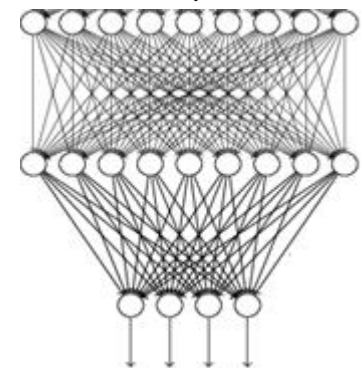
Max Pooling

filters

Convolution

Max Pooling

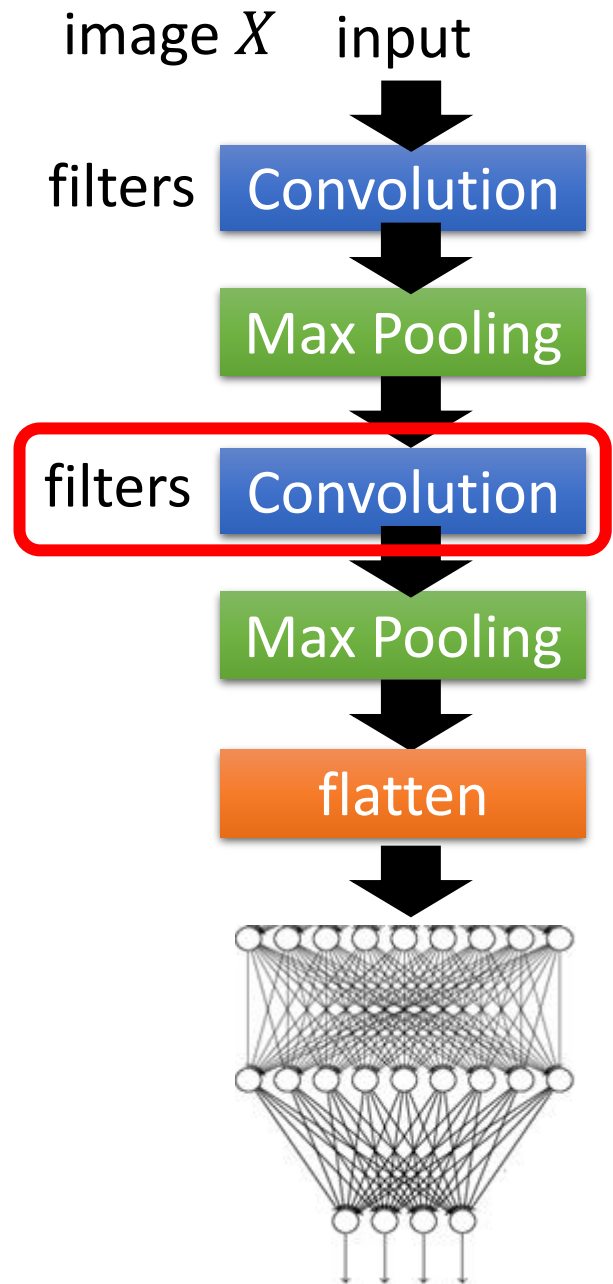
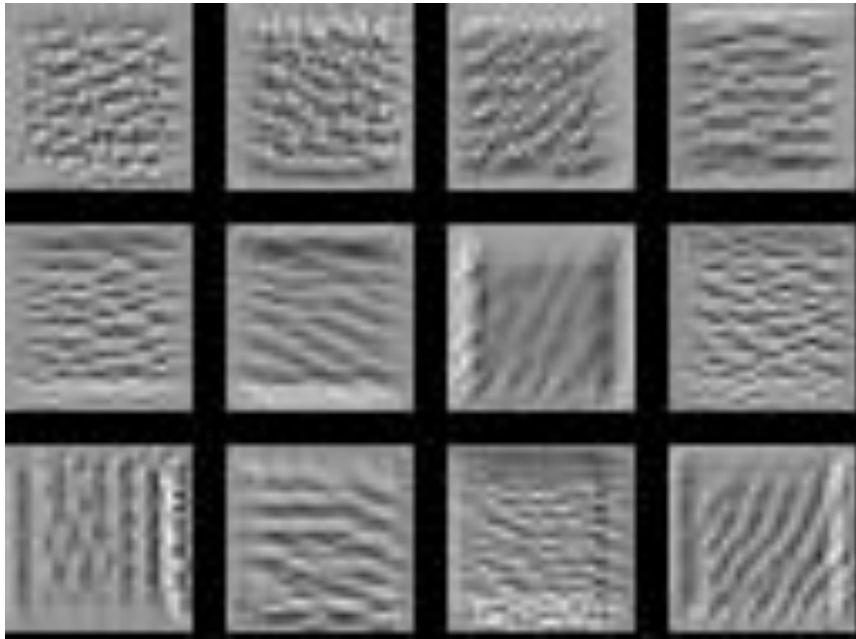
flatten



What does a filter detect?

E.g., Digit classifier

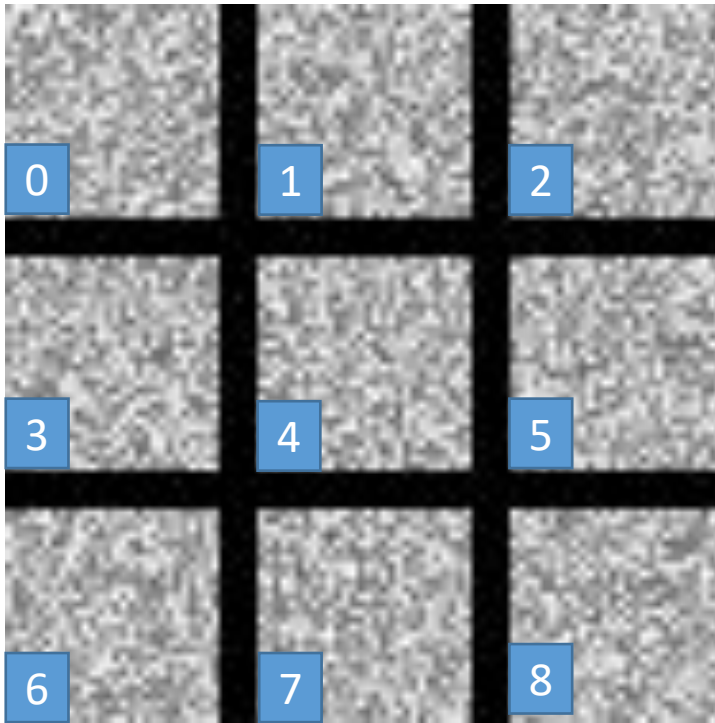
X^* for each filter



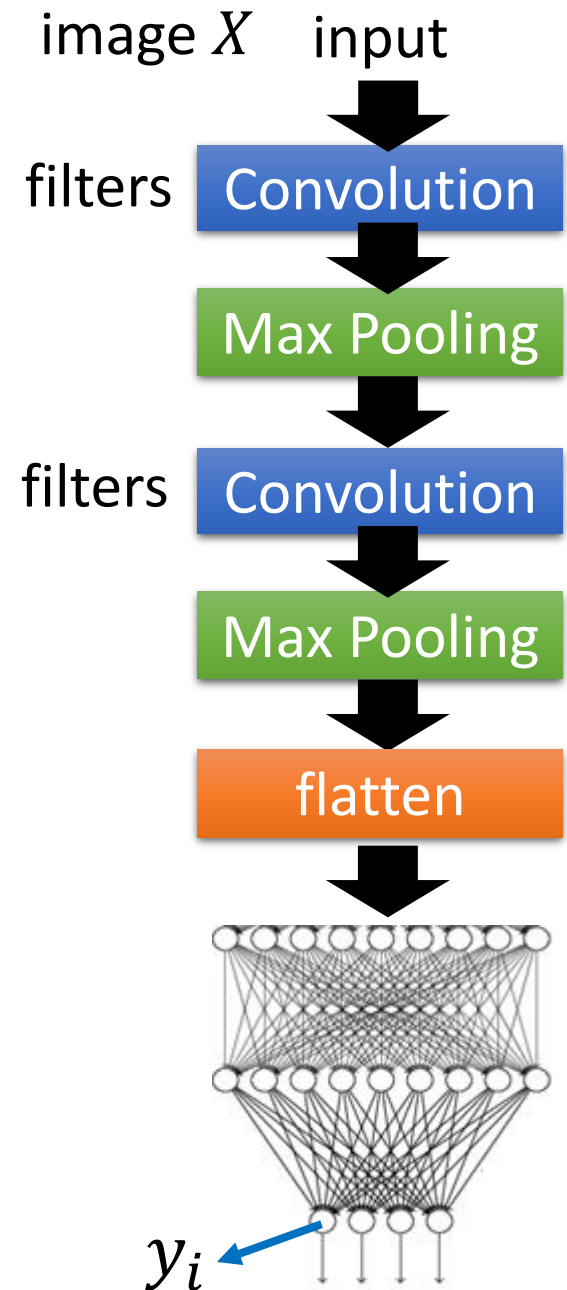
What does a digit look like for CNN?

E.g., Digit classifier

$$X^* = \arg \max_X y_i \quad \text{Can we see digits?}$$



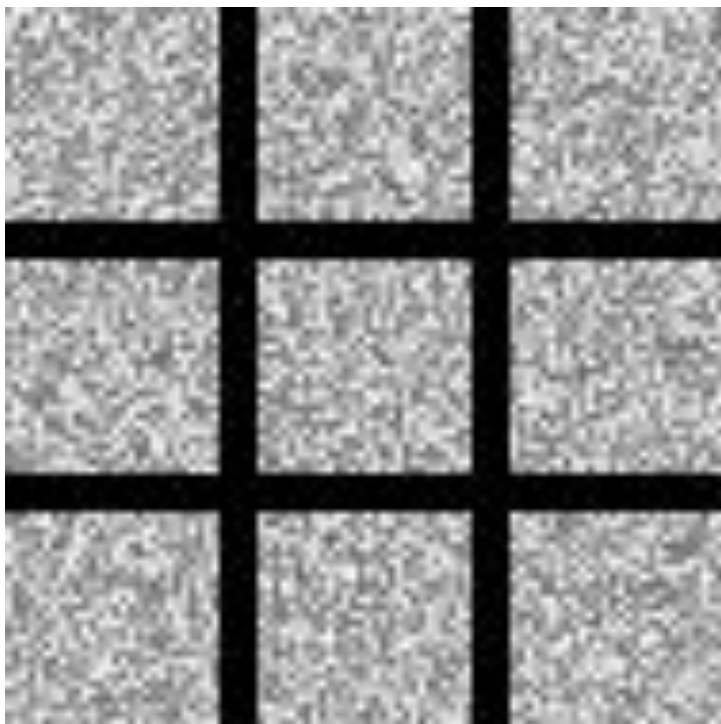
Surprise? Consider adversarial attack!



What does a digit look like for CNN?

Find the image that
maximizes class probability

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

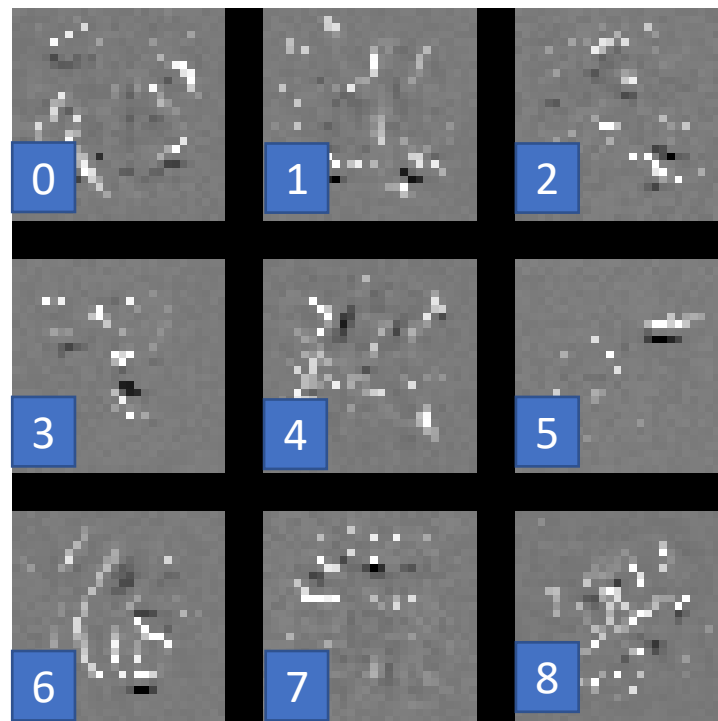


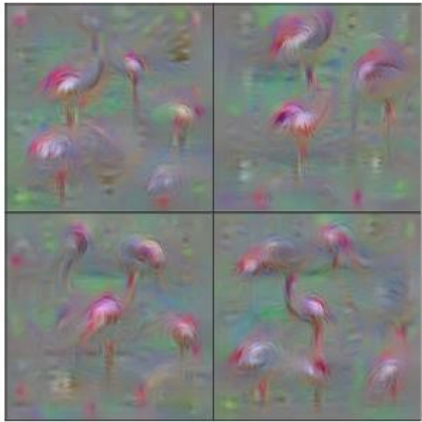
The image should look like a digit.

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

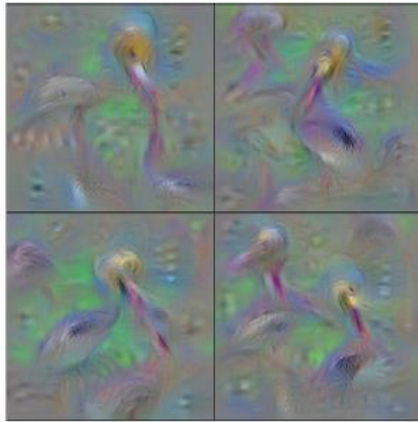
$$R(X) = - \sum_{i,j} |X_{ij}|$$

How likely
 X is a digit

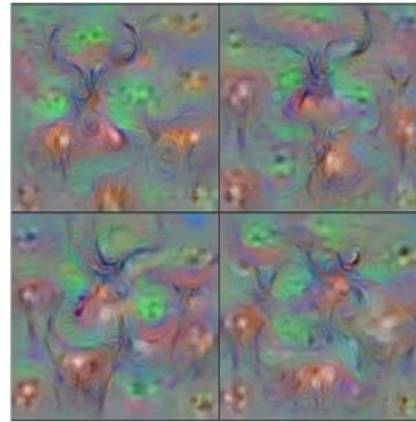




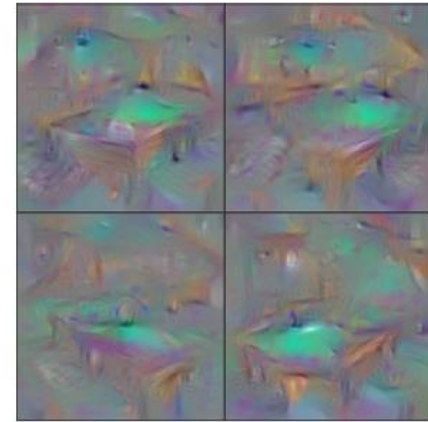
Flamingo



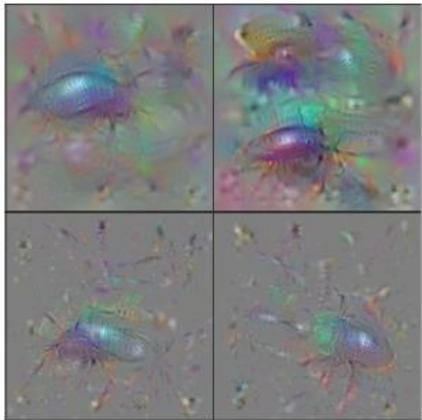
Pelican



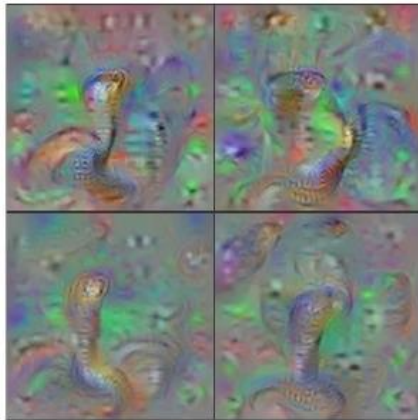
Hartebeest



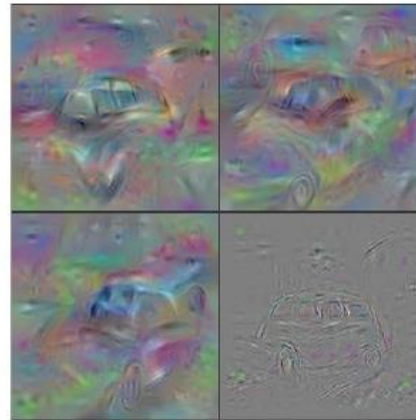
Billiard Table



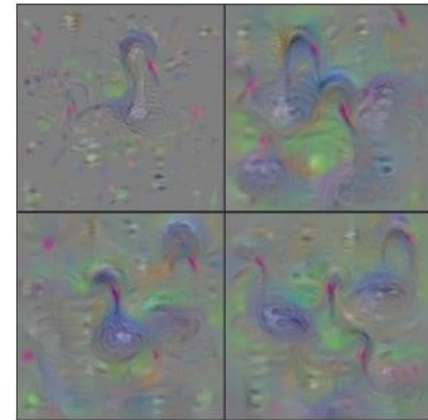
Ground Beetle



Indian Cobra



Station Wagon



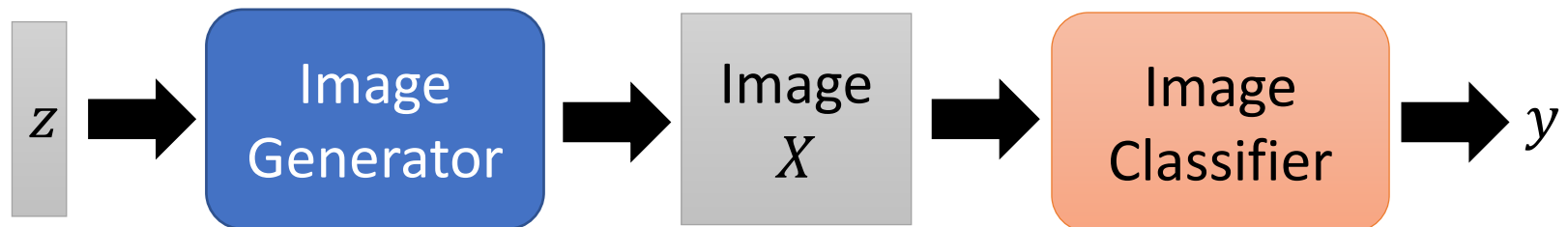
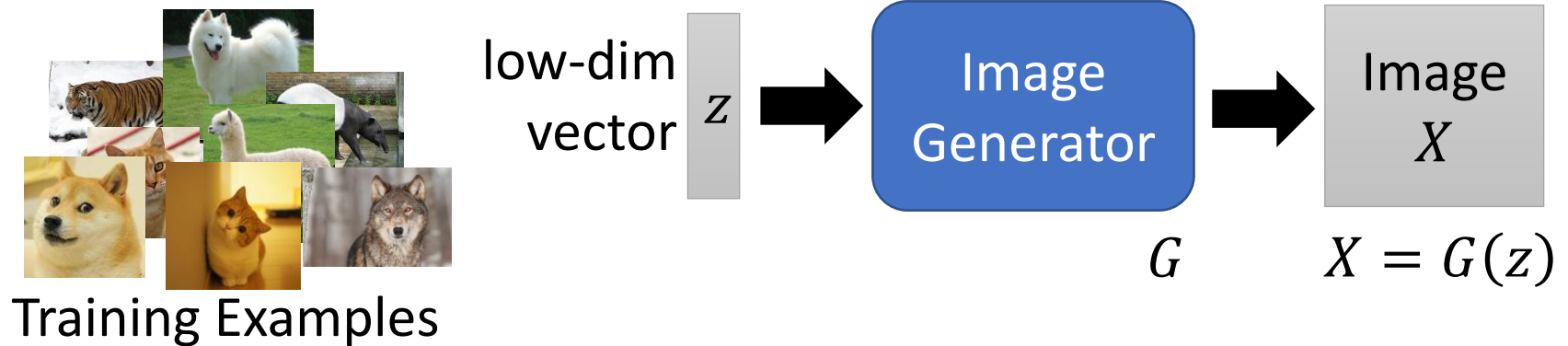
Black Swan

With several regularization terms, and hyperparameter tuning

<https://arxiv.org/abs/1506.06579>

Constraint from Generator

- Training a generator



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

Show image:

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



redshank

ant

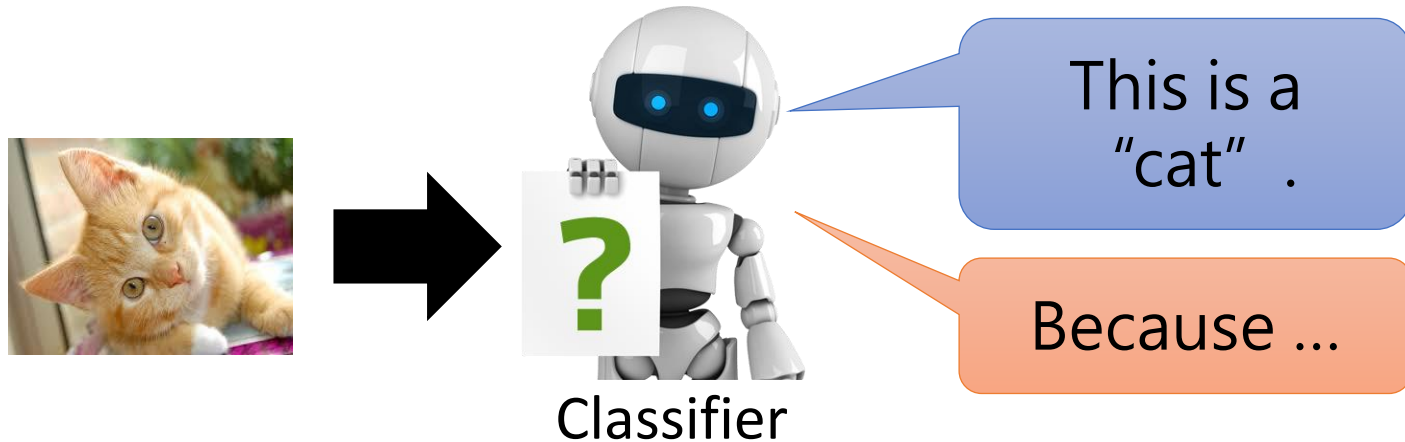
monastery



volcano

<https://arxiv.org/abs/1612.00005>

Concluding Remarks



Local Explanation

Why do you think this image is a cat?

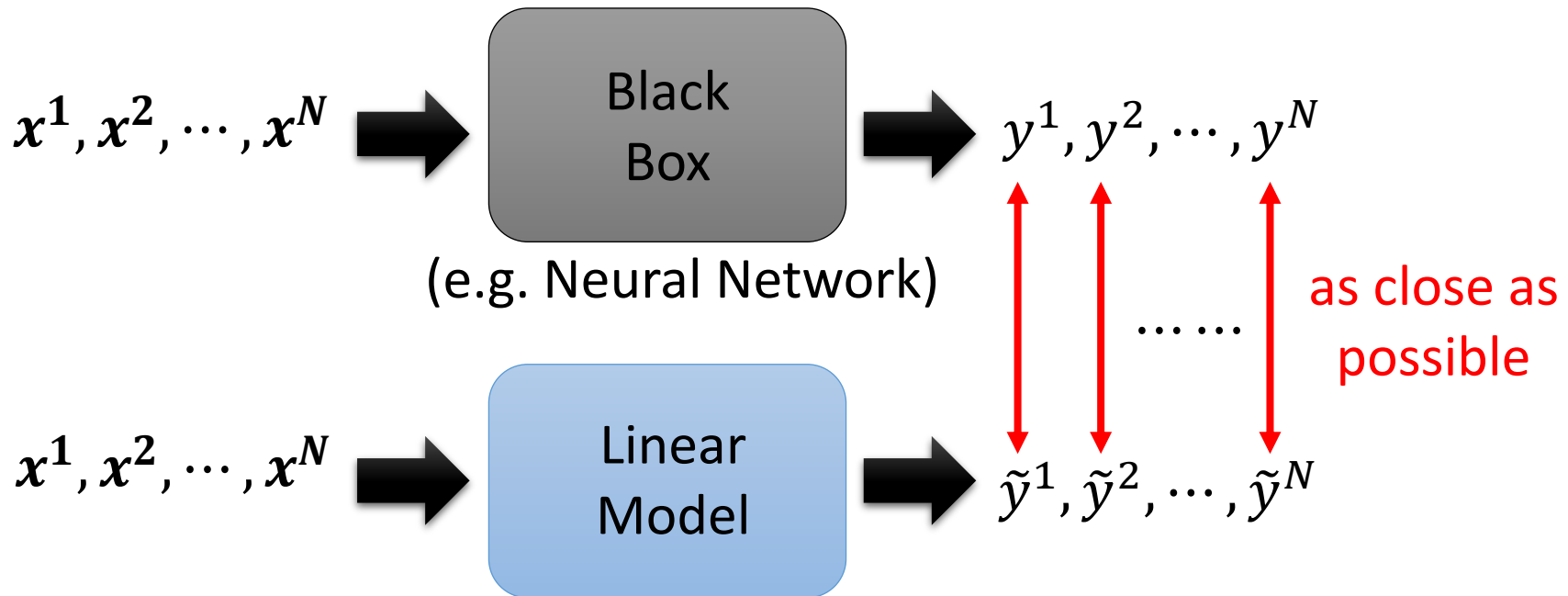
Global Explanation

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.



Local Interpretable Model-Agnostic Explanations (LIME)

<https://youtu.be/K1mWgthGS-A>

<https://youtu.be/OjqIVSwly4k>

李宏毅 (Hung-yi Lee) · HYLEE | Machine Learning (2021)

HYLEE(2021) · 课程资料包 @ShowMeAI



视频

课件

笔记

代码

中英双语字幕

一键打包下载

官方笔记翻译

作业项目解析



视频 · B 站 [扫码或点击链接]

<https://www.bilibili.com/video/BV1fM4y137M4>



课件 & 代码 · 博客 [扫码或点击链接]

<http://blog.showmeai.tech/ntu-hylee-ml>

机器学习

Auto-encoder

生成式对抗网络

学习率

深度学习

卷积神经网络

GAN

自监督

自注意力机制

批次标准化

神经网络压缩

强化学习

元学习

Transformer

Awesome AI Courses Notes Cheatsheets 是 [ShowMeAI](#) 资料库的分支系列，覆盖最具知名度的 **TOP50+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

点击课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

机器学习	深度学习	自然语言处理	计算机视觉
Stanford · CS229	Stanford · CS230	Stanford · CS224n	Stanford · CS231n
# Awesome AI Courses Notes Cheatsheets · 持续更新中			
知识图谱	图机器学习	深度强化学习	自动驾驶
Stanford · CS520	Stanford · CS224W	UCBerkeley · CS285	MIT · 6.S094



微信公众号

资料下载方式 2: 扫码点击**底部菜单栏**

称为 **AI 内容创作者**? 回复 [添砖加瓦]