

机器学习的可解释性

为什么需要可解释性的机器学习： 机器学习不仅得到答案，更需要给出决策的**理由**。我们可以从解释中改进模型和判错。

目标就是让人看的爽信服（readable）。

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

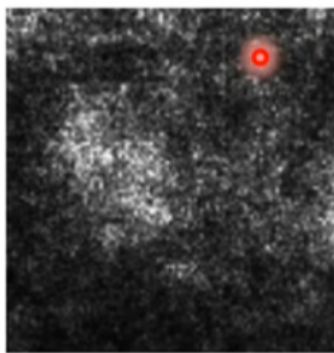
那一个组成成分对于做出决定是关键，如果删除或者改变这个组件会有很大的判断变化。

- e.g.: 把图片每一个像素写成 $\{x_1, \dots, x_n, \dots, x_N\}$ ，得到的结果是 e ，如果 $\{x_1, \dots, x_n + \Delta x, \dots, x_N\}$ ，结果变成 $e + \Delta e$ 。
 - 如果 Δx 小，但是 Δe 大，说明是关键，那么可以用 $|\frac{\Delta e}{\Delta x}| \rightarrow |\frac{\partial e}{\partial x}|$ 判断（越大越关键）

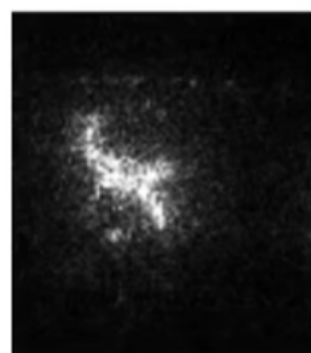
如何把显著性图画得更好（让人看的更加信服）？



Gazelle



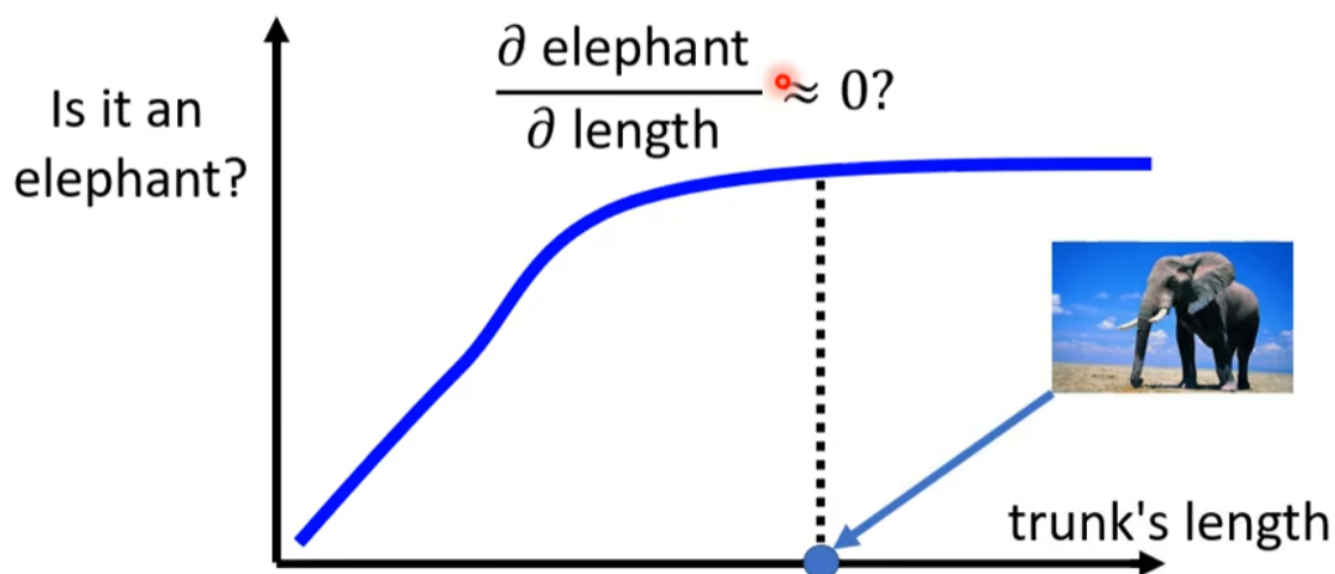
Typical



SmoothGrad

利用SmoothGrad，随机向输入图像添加噪声，获取噪声图像的显著性热力图并对其进行平均。

但利用 $|\frac{\partial e}{\partial x}|$ 这种方法判断也不是完全正确的，如果根据鼻子的长度判断是否是大象。当鼻子长度增加到一定程度，变化就变得很小了。

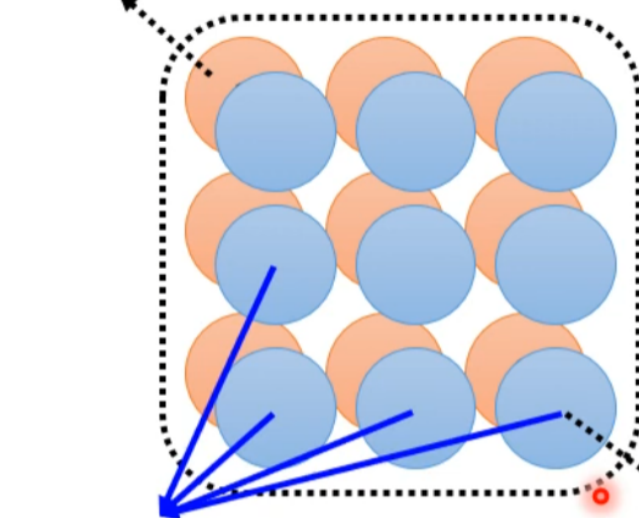


Global Explanation

对于每一个filter进行解释，如果一个filter算出来的feature map的值很大，说明这个filter很重要，图上的很多pattern都需要filter检测。

What does a filter detect?

output of filter 2



Large values

➡ Image X contains the patterns filter 1 can detect.

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

unknown

image X input

filters

Convolution

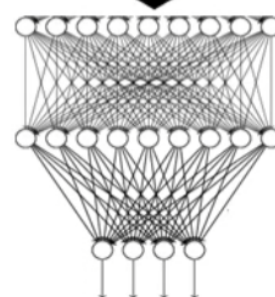
Max Pooling

filters

Convolution

Max Pooling

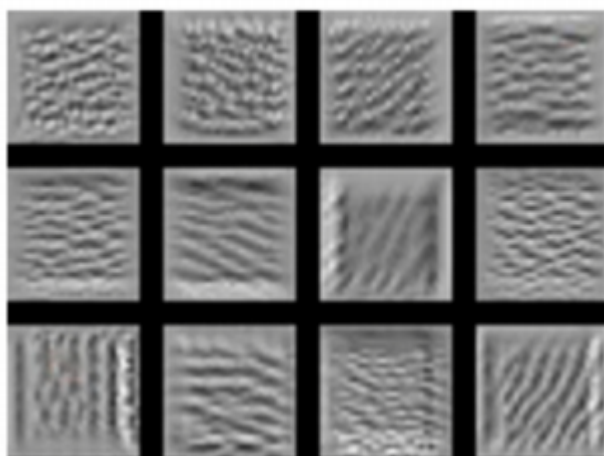
flatten



- 如何找到每一个filter关注的pattern?

我们可以自己创建一个输入 X^* 使得, $X^* = \arg \max_x \sum_i \sum_j a_{ij}$, 此时 X^* 就是关注的特征。例如在手写数字辨识中的某个卷积层, 可以看出不同滤波器的关注点同 (横向、纵向) :

X^* for each filter

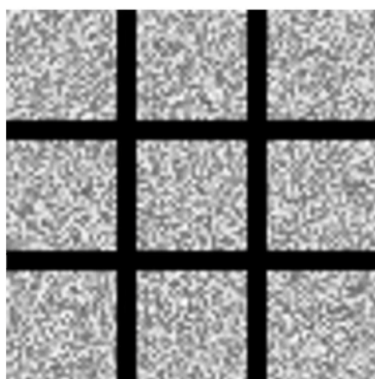


但即使最后输出层的判断，也不一定会出现readable的图像，如下图（左）（考虑危险攻击，有噪声，不一定完全判断具象的数字）。但为了可视化，我们可以添加限制，要求白色的点越少越好。

What does a digit look like for CNN?

Find the image that maximizes class probability

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

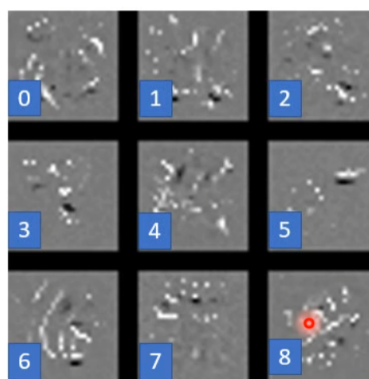


The image should look like a digit.

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

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

How likely
X is a digit



还可以收到生成器的约束

Constraint from Generator

- Training a generator

(by GAN, VAE, etc.)



Training Examples

low-dim
vector
z

Image
Generator

G

Image
X

$$X = G(z)$$



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

从而变成创建一个输入 X^* 的流形 z^* 。