VISUALIZING DEEP NEURAL NETWORK DECISIONS: PREDICTION DIFFERENCE ANALYSIS

高久怡　2019/11/24

CONTENTS

目录

Introduction

Introduction

Figure 1: Example of our visualization method: explains why the DCNN (GoogLeNet) predicts "cockatoo". Shown is the evidence for (red) and against (blue) the prediction. We see that the facial features of the cockatoo are most supportive for the decision, and parts of the body seem to constitute evidence against it. In fact, the classifier most likely considers them evidence for the second-highest scoring class, white wolf.

从图片中移除特征并评估被移除部分的影响

Introduction

prediction difference analysis

- 衡量input feature对target class ***c*** 的relevance

- 当该输入特征被移除后预测结果的变化： Difference between $p(c \mid \mathbf{x})$ and $p(c \mid \mathbf{x}_{\setminus i})$

$$p(c \setminus \mathbf{x}_{\setminus i}) = \sum_{x_i} p(x_i) p(c \mid \mathbf{x}_{\setminus i}, x_i) = \sum_{x_i = a_1}^{a_m} p(x_i = a_j) p(c \mid \mathbf{x} \leftarrow x_i = a_j)$$

prior

全概率公式

Introduction

$$relevance \quad score = WE_i(c \mid \mathbf{x}) = \log_2(odds(c \mid \mathbf{x}) - \log_2(odds(c \mid \mathbf{x}_{\setminus i}))$$

$$odds(p) = \frac{p}{1-p}$$

score： large or small positive or negative

Improvement

Improvement

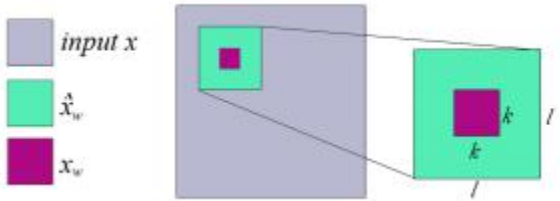
➢ Conditional Sampling

在图片中，像素之间的相关性很高，不能用相互独立近似

条件：

① 像素**强**依赖于其周围小部分区域的像素

② 像素领域**不**依赖于像素在图片中的位置



Improvement

- **Multivariate Analysis**

  移除 feature sets：相关像素集，而不是单个像素；

  以$k \times k$大小的 patch 遍历整张图片，采用滑动窗口模式；

Improvement

k=2, l=k+4, S=4

patch

3	0	1	2	7	4
1	5	8	9	3	1
2	7	2	5	1	3
0	1	3	1	7	8
4	2	1	6	2	8
2	4	5	2	3	9

for $s = 1$ **to** S **do**
  $\mathbf{x}'_w \leftarrow \mathbf{x}_w$ sampled from $p(\mathbf{x}_w|\hat{\mathbf{x}}_w \backslash \mathbf{x}_w)$
  $\text{sum}_w \mathrel{+}= p(c|\mathbf{x}')$
end for
$p(c|\mathbf{x}\backslash\mathbf{x}_w) := \text{sum}_w / S$

$$WE = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \vdots & r_1 & r_1 & \vdots \\ \vdots & r_1 & r_1 & \vdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}_{n\times n}$$

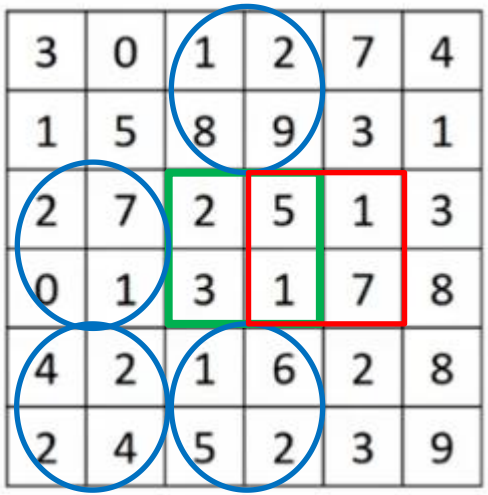
$$counts = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \vdots & 1 & 1 & \vdots \\ \vdots & 1 & 1 & \vdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}_{n\times n}$$

$$WE_i(c \mid \mathbf{x}) = \log_2(odds(c \mid \mathbf{x}) - \log_2(odds(c \mid \mathbf{x}_{\backslash i}))$$

Improvement

k=2, l=k+4, S=4

patch



$p(c|\mathbf{x}\backslash\mathbf{x}_w)$

$$WE = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & r_1 & r_1+r_2 & r_1+r_2 & \vdots \\ \vdots & r_1 & r_1+r_2 & r_1+r_2 & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

$$\odot$$

$$counts = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & 1 & 1+1 & 0+1 & \vdots \\ \vdots & 1 & 1+1 & 0+1 & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Improvement

- **Deep Visualization of Hidden Layers**

$\mathbf{h}$ 为某隐藏层的激活向量，$\mathbf{z}$ 为与 $\mathbf{h}$ 相关的后一（几）层的节点的激活值，$z = z(\mathbf{h})$

$$p(c \mid \mathbf{x}_{\setminus i}) = \sum_{x_i} p(x_i \mid \mathbf{x}_{\setminus i})\, p(c \mid \mathbf{x}_{\setminus i}, x_i)$$

$$g(z \mid \mathbf{h}_{\setminus i}) = E_{p(h_i \mid \mathbf{h}_{\setminus i})}[z(\mathbf{h})] = \sum_{h_i} p(h_i \mid \mathbf{h}_{\setminus i}) z(\mathbf{h}_{\setminus i}, h_i)$$

activation

Improvement

- **Deep Visualization of Hidden Layers**

$\mathbf{h}$ 为某隐藏层的激活向量，$\mathbf{z}$ 为与 $\mathbf{h}$ 相关的后一（几）层的节点的激活值，$z = z(\mathbf{h})$

$$g(z \mid \mathbf{h}_{\setminus i}) = E_{p(h_i \mid \mathbf{h}_{\setminus i})}[z(\mathbf{h})] = \sum_{h_i} p(h_i \mid \mathbf{h}_{\setminus i}) z(\mathbf{h}_{\setminus i}, h_i)$$

weight of evidence

$$relevance \quad score = WE_i(c \mid \mathbf{x}) = \log_2(odds(c \mid \mathbf{x}) - \log_2(odds(c \mid \mathbf{x}_{\setminus i}))$$

activation difference :

$$AD_i(z \mid \mathbf{h}) = g(z \mid \mathbf{h}) - g(z \mid \mathbf{h}_{\setminus i})$$

不使用概率值，而使用激活值

Experiment

Experiment

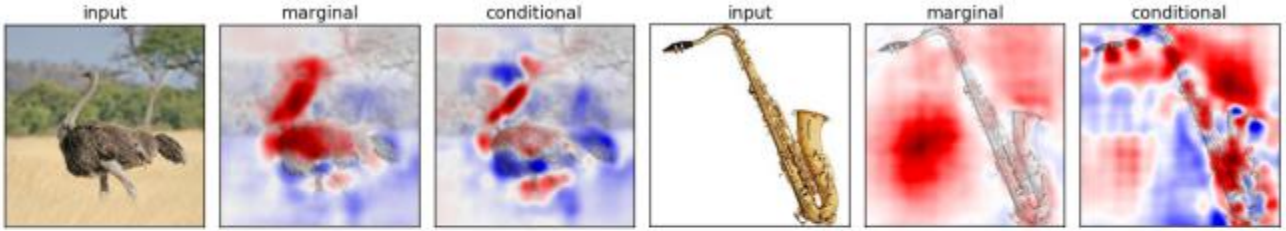


Figure 3: **Visualization of the effects of marginal versus conditional sampling** using the GoogLeNet classifier. The classifier makes correct predictions (ostrich and saxophone), and we show the evidence for (red) and against (blue) this decision at the output layer. We can see that conditional sampling gives more targeted explanations compared to marginal sampling. Also, marginal sampling assigns too much importance on pixels that are easily predictable conditioned on their neighboring pixels.

Experiment

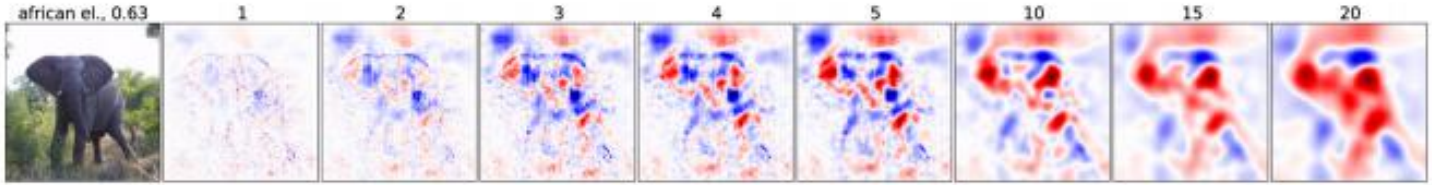


Figure 4: **Visualization of how different window sizes influence the visualization result.** We used the conditional sampling method and the AlexNet classifier with $l = k + 4$ and varying k. We can see that even when removing single pixels ($k = 1$), this has a noticeable effect on the classifier and more important pixels get a higher score. By increasing the window size we can get a more easily interpretable, smooth result until the image gets blurry for very large window sizes.

Experiment

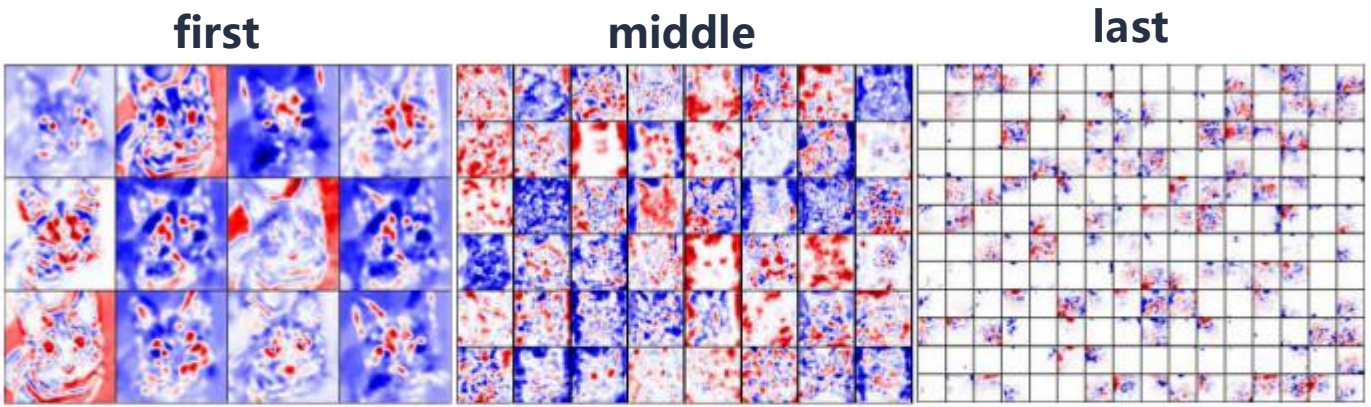
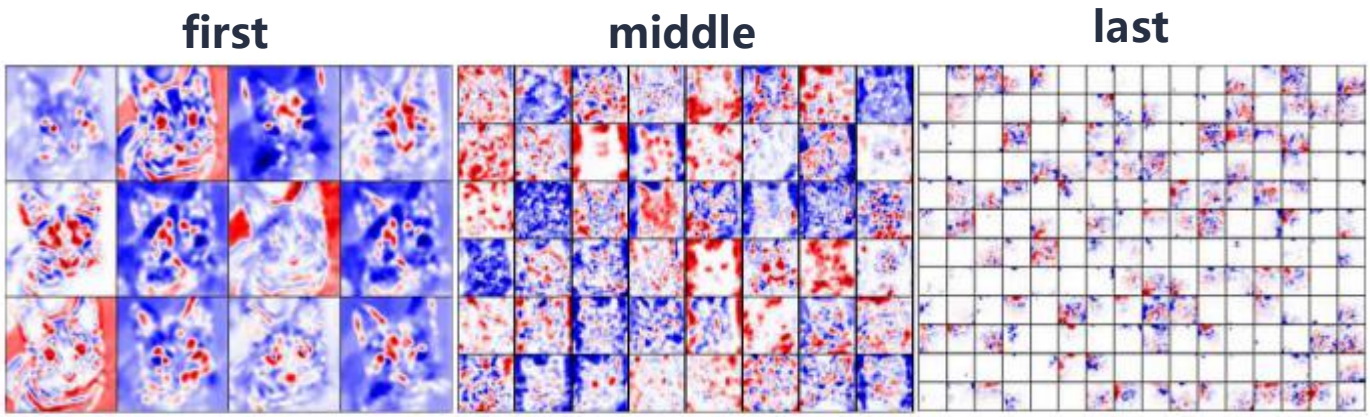


Figure 5: Visualization of feature maps from thee different layers of the GoogLeNet (l.t.r.: "conv1/7x7_s2", "inception_3a/output", "inception_5b/output"), using conditional sampling and patch sizes $k = 10$ and $l = 14$ (see alg. 1). For each feature map in the convolutional layer, we first evaluate the relevance for every single unit, and then average the results over all the units in one feature map to get a sense of what the unit is doing as a whole. *Red* pixels activate a unit, *blue* pixels decreased the activation.

Experiment



观察**不同层**学到了什么

conv/7x7_s2：图片的不同部分对不同的边缘检测卷积核有正向或负向的反应

inception_3a/output：抽取更高水平的特征（猫的脸部特征）

inception_5b/output：在channel维上特征更为稀疏，表明最后一层卷积抽取的特征更为抽象，每一个unit包含的空间语义信息更丰富

Experiment

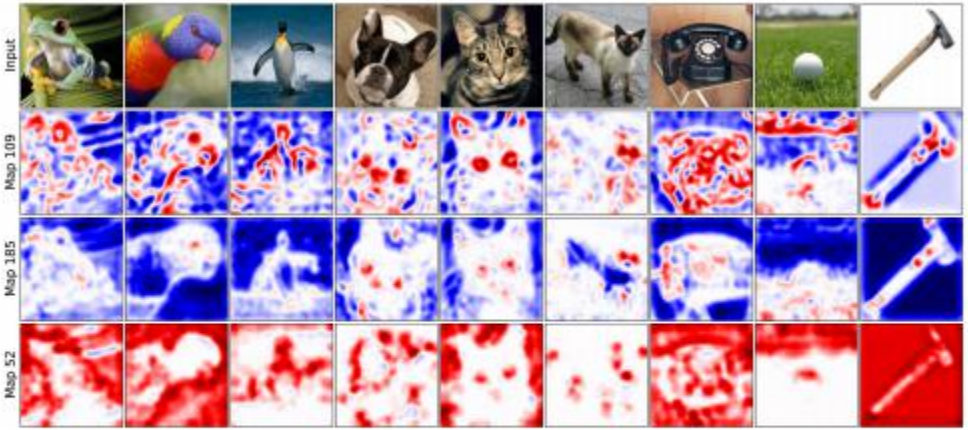
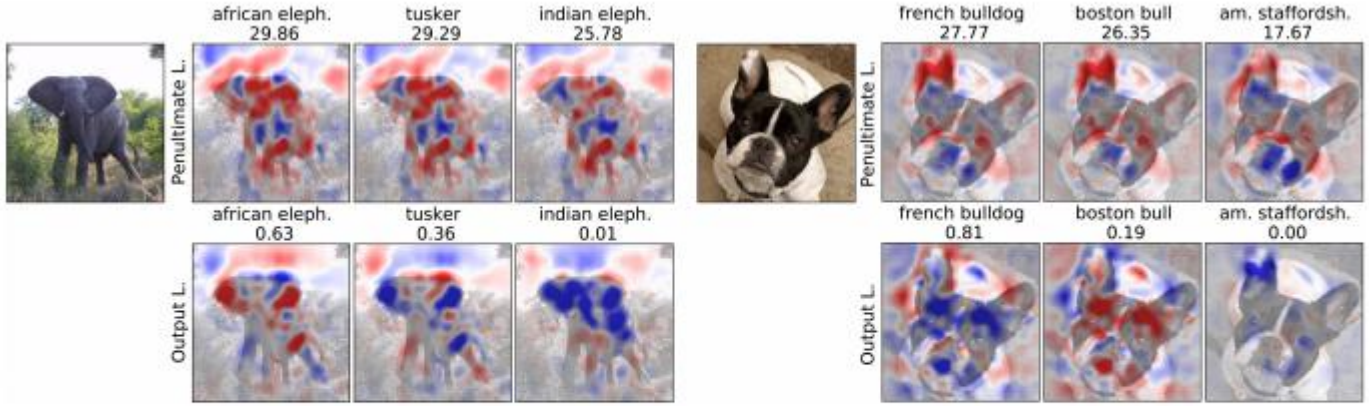


Figure 6: **Visualization of three different feature maps**, taken from the "inception_3a/output" layer of the GoogLeNet (from the middle of the network). Shown is the average relevance of the input features over all activations of the feature map. We used patch sizes $k = 10$ and $l = 14$ (see alg. 1). *Red* pixels activate a unit, *blue* pixels decreased the activation.

Experiment



pre-softmax

softmax

Figure 7: **Visualization of the support for the top-three scoring classes in the penultimate- and output layer**. Next to the input image, the first row shows the results with respect to the penultimate layer; the second row with respect to the output layer. For each image, we additionally report the values of the units. We used the AlexNet with conditional sampling and patch sizes $k = 10$ and $l = 14$ (see alg. 1). *Red* pixels are evidence for a class, and *blue* against it.

Top3 class（相似类别）：presoftmax对于相似的类别for or against的像素区域大致**相似**，但softmax就完全不一样

Experiment

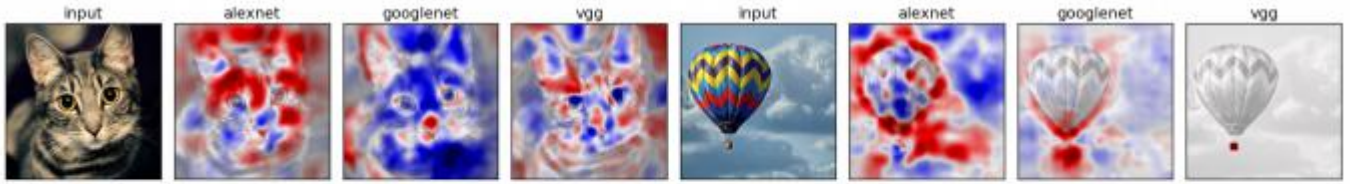


Figure 8: **Comparison of the prediction visualization of different DCNN architectures.** For two input images, we show the results of the prediction difference analysis when using different neural networks - the AlexNet, GoogLeNet and VGG network.

AlexNet：更关注背景信息

VGG：气球图片中篮子更为重要；top2 class为降落伞，网络可能通过检测一个方形的篮子，学会了不把气球和降落伞混淆

Conclusion

Conclusion

- **feature for or against output**
- **遍历移除特征，并观察预测值变化，从而计算relevance score**
- **改进：考虑特征之间的相关性（条件采样代替边缘分布）；同时移除多个特征（降低网络对单一像素的敏感性**
- **计算效率不高**

THANKS!

LIVE AND LEARN