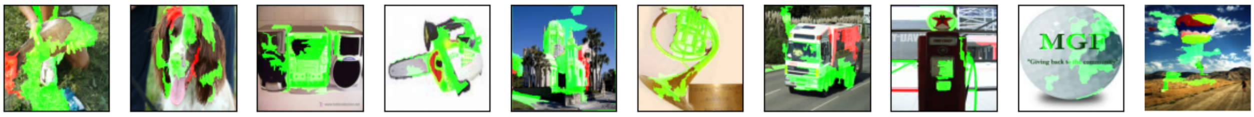
可信机器学习 HW1 Report

1 LIME

1.1 实现过程

直接使用 `lime.lime_image.LimeImageExplainer` 建立 LIME 模型, 并使用 `explain_instance` 进行实例化.

1.2 实现效果



2 Saliency Map

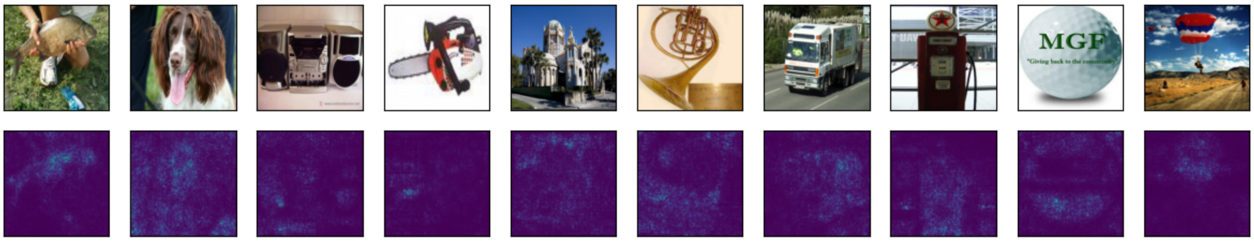
2.1 实现过程

Step 1. 在 `eval` 状态下, 计算输入图像 X 的预测结果 y_{pred}.

Step 2. 计算损失 $\mathcal{L} = \mathrm{CrossEntropy}(\mathrm{y_{pred}}, \mathrm{y})$.

Step 3. 计算 $\left|\dfrac{\partial \mathcal{L}}{\partial X}\right|$ 并做 min-max 归一化, 即可得到 Saliency Map.

2.2 实现效果



3 Smooth Grad

3.1 实现过程

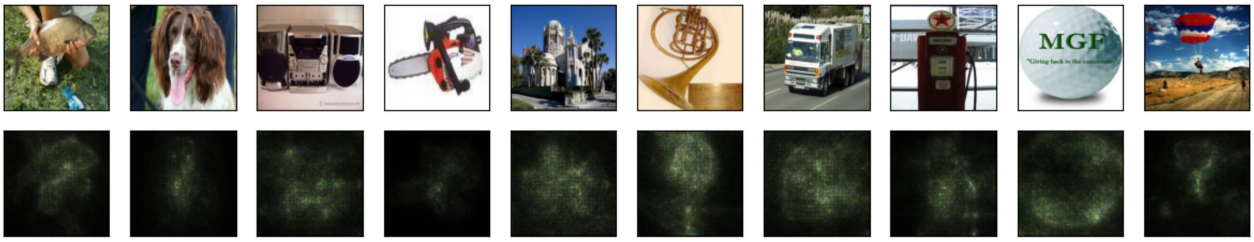
Step 1. 选取噪声标准差 $\sigma = \dfrac{\sigma_0}{x_{\max} - x_{\min}}$

Step 2. 随机生成 n 个满足正态分布 $\mathcal{N}(0, \sigma^2)$ 的噪声 $\text{noise}_1, \cdots, \text{noise}_n$. 计算 $X + \text{noise}_i$ 的损失 $\mathcal{L}_i$.

Step 3. 计算 $\overline{\nabla X} = \dfrac{1}{n}\sum_{i=1}^{n} \left| \dfrac{\partial \mathcal{L}_i}{\partial (X + \text{noise}_i)} \right|$ 即可得到 Smooth Grad.

实际代码中对其做 min-max 归一化以优化可视化效果.

3.2 实现效果



4 Grad-CAM

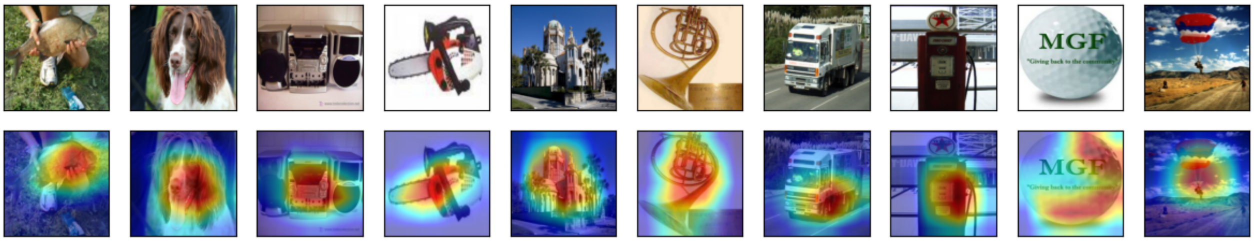
4.1 实现过程

Step 1. 利用 `grad` 求出权重, 对 `activation` 进行加权求和得到 `cam`.

Step 2. 对 `cam` 进行 ReLU 激活和 min-max 归一化.

Step 3. 将所得的 `cam` 覆盖到原图像并进行可视化.

4.2 实现效果



5 Paper Reading

选择的论文为Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.

5.1 Brief Summary

Grad-CAM is a technique for producing visual explanations by localizing discriminative image regions. It uses the gradients of any target concept with respect to intermediate features determined by a CNN, to understand and visualize which parts of the image have the greatest influence on the final result. It is proved to be interpretable to humans and faithful to the model through human studies and experiment. Therefore, Grad-CAM is a strong tool for analysis and explanation in multiple tasks.

5.2 Strength

(a) Generalization. Grad-CAM can be applied to any CNN-based models. (b) Simple. Grad-CAM is easy to implement as we only needs to take derivative, do a weighted sum and apply ReLU activation. (c) Multi-use. Apart from creating visualized explanations, Grad-CAM can be used to analyze failure models and identify bias in the dataset. (d) Interpretable. The result of Grad-CAM has a correlation with human attention maps.

5.3 Weakness

(a) Grad-CAM lacks the ability to show fine-grained importance. (b) If an object occurs more than once in an image, Grad-CAM fails to properly localize them. (c) The localization often correspond to parts of an object rather than the entire one.