Exercise 10

Network Interpretability and Visualization

Reliable and Interpretable Artificial Intelligence
ETH Zurich

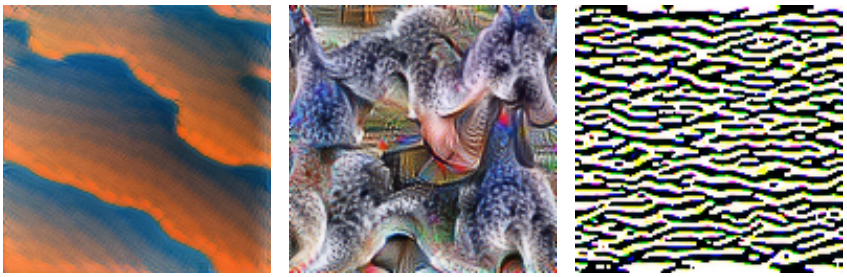

Figure 1: Images strongly activating convolutional filters in three layers created by optimization. Order permuted.

Problem 1 (Feature Visualization)**.** Figure 1 shows three images created by optimization (with Lucid[1]). These images strongly activate convolutional filters in different layers in a convolutional neural network. Order the images according to their depth in the neural network. Provide an argument to justify your answer.

Problem 2 (Shapley Values)**.** Consider the classifier $g : \mathbb{R}^3 \to \mathbb{R}^2$ and data points $\boldsymbol{x}$ and $\boldsymbol{x}'$.

$$\boldsymbol{x} := \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \qquad \boldsymbol{x}' := \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}$$

$$g(\boldsymbol{x}) := \boldsymbol{B}\,\mathrm{ReLU}(\boldsymbol{A}\boldsymbol{x})$$

$$A := \begin{pmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad B := \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

1. Compute the Shapley Values for all components of both inputs and both target classes. Set non-included features to zero. Discuss the results. (Consider implementing repetitive calculation in python.)

[1] Googles framework for neural network visualization. `https://github.com/tensorflow/lucid`

2. Argue why Shapley Values are impractical (applied directly; without approximations) for large images.

Problem 3 (Robust Features vs Non-Robust Features)**.** (Adapted from [1])

Consider a distribution $\mathcal{D}$ of data points $(\boldsymbol{x}, y) \sim \mathcal{D}$ where

$$y \overset{\text{u.a.r.}}{\sim} \{-1, 1\}, \quad \boldsymbol{x} \in \mathbb{R}^{d+1}, \quad \boldsymbol{x}_1 = \begin{cases} +y & \text{w.p. } p \\ -y & \text{w.p. } 1-p \end{cases}, \quad \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{d+1} \sim \mathcal{N}(\eta y, 1).$$

Here, $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. Further consider the family of linear classifiers $f_i(\boldsymbol{x}) := \text{sign}((\boldsymbol{w}^i)^T \boldsymbol{x})$. We use two different instantiation with $\boldsymbol{w}^1 := (0, \frac{1}{d}, \ldots, \frac{1}{d})^T \in \mathbb{R}^{d+1}$ and $\boldsymbol{w}^2 := (1, 0, \ldots, 0)^T \in \mathbb{R}^{d+1}$. In this example we will use $\eta = \frac{3}{\sqrt{d}}$ and $p = 0.9975$. Thus, we refer to the first component of $\boldsymbol{x}_1$ as a robust feature (as it is strongly associated with y), and $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_{d+1}$ as weak or non-robust features as for large d (e.g., $d \geq 100$) there is only weak correlation y and the individual features.

1. Compute the expected accuracy of f_1, f_2. That is, determine value of $\underset{(\boldsymbol{x},y)\sim\mathcal{D}}{\mathcal{P}}[f(\boldsymbol{x}) = y]$.

2. Consider the data point $y = 1$, $\boldsymbol{x} = (1, 0.1, \ldots 0.1)^T$. Attempt to find an adversarial example $\boldsymbol{x}'$ with $\|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq 0.15$, such that it gets misclassified ($f(\boldsymbol{x}') = -1$). For both, f_1 and f_2, find such an example or describe why it does not exist.

References

[1] Dimitris Tsipras et al. "Robustness May Be at Odds with Accuracy". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=SyxAb30cY7.