c-Eval: A Unified Metric to Evaluate Feature-based Explanations via Perturbation

Minh N. Vu*, Truc D. Nguyen*, NhatHai Phan†, Ralucca Gera‡, and My T. Thai*

*University of Florida, Gainesville, Florida, USA
†New Jersey Institute of Technology, Newark, New Jersey, USA
‡Naval Postgraduate School, Monterey, California, USA

Presented by Giulia Ghisolfi and Irene Testa

Outline

Introduction

Motivation

- Many feature based local explainers have been proposed
- Given the lack of ground truth explanations, evaluating the output quality of these explainers is challenging

Related work

- Area Over the Perturbation Curve (AOPC) [1]
  - specifically designed to evaluate heat-maps
  - requires a large number of random perturbations to make a stable evaluation
- log-odds score [2]
  - proposed without a detailed analysis

c-Eval

c-Eval quantifies the **minimum amount of perturbation on non explanatory features that is needed to alter the prediction** of the model.

Formal Definition

Given an input $x \in R^n$, a classifier $f : \mathbb{R}^n \to \mathbb{R}^m$ s.t. $argmax_{1 \leq j \leq m} f(x)_j = l$, a feature based local explainer g_f of the model f s.t. $g_f(x) = e_x \subseteq x$ and a perturbation scheme h_{g_f} of e_x, i.e. $h_{g_f} : \mathbb{R}^n \to \mathbb{R}^n$ s.t.

$$h_{g_f}(x)_i = x_i + \delta_i \text{ where } \begin{cases} \delta_i = 0 & \text{if } x_i \in e_x \\ \delta_i \geq 0 & \text{if } x_i \notin e_x \end{cases}$$

the c-Eval of the explanation e_x is

$$c_{f,x}(e_x) = \inf c$$

s.t. $\exists h_{g_f}$ satisfying the following conditions

$$argmax_{1 \leq j \leq m} f(h_{g_f})_j \neq l$$

$$\|h_{g_f}(x) - x\|_p \leq c$$

Example

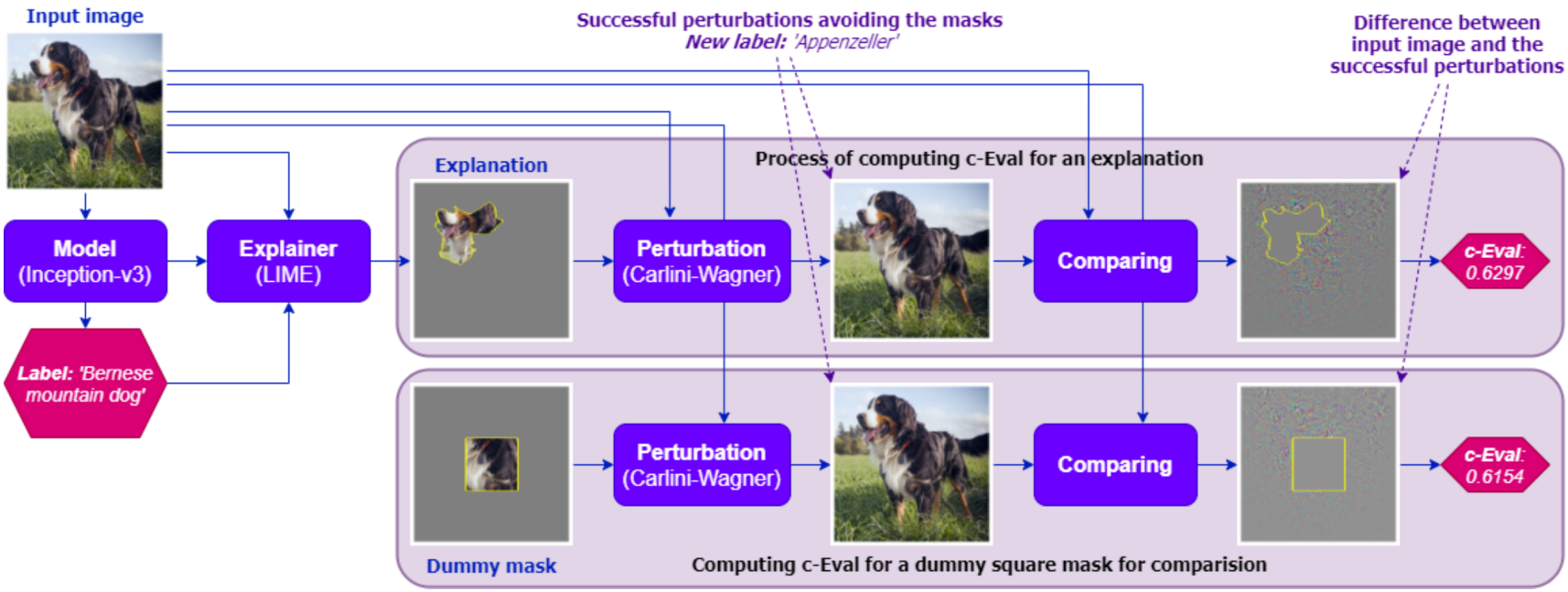

Figure 1: An illustrative example comparing the c-Eval of a LIME explanation with that of a dummy mask.

Remarks

- The higher the c-Eval, the better the explanation
- When $e_x = \emptyset$, h_{g_f} returns the *minimally distorted adversarial example*; while when $e_x = x$ we set $c_{f,x}(e_x) = \inf$
- c-Eval should be applied when the cardinalities of the sets of explanatory features have comparable sizes
- To compare the quality of explanations on different inputs we should consider $C_{f,x}(e_x) = c_{f,x}(e_x)/c_{f,x}(\emptyset)$
- To compute c-Eval the authors use $\|\cdot\|_2$, leaving the study of the optimal norm to future works

Computing c-Eval

Modify existing algorithms to generate adversarial samples of neural networks so that the perturbations act only on non-explanatory features:

- Use the **Carlini Wagner (CW) attack** [3]
  - Compute the **minimal** perturbation on non explanatory features that alters the output of the model
  - Requires **high running times**
- Restrict the definition of c-Eval to a specific class of perturbing schemes $\mathcal{H}$, e.g. those generated by the **Gradient-Sign-Attack (GSA)** [4] and by the **Iterative-Gradient-Attack (IGA)** [5]
  - Compute the **minimal** perturbation on non explanatory features that alters the output of the model **in the class of perturbation schemes** $\mathcal{H}$
  - Requires **lower running times**
  - The computed c-Eval is **correlated** with the c-Eval obtained through the CW attack (Figure 2)

Computing c-Eval

Comparison between GSA, IGA and CW attack

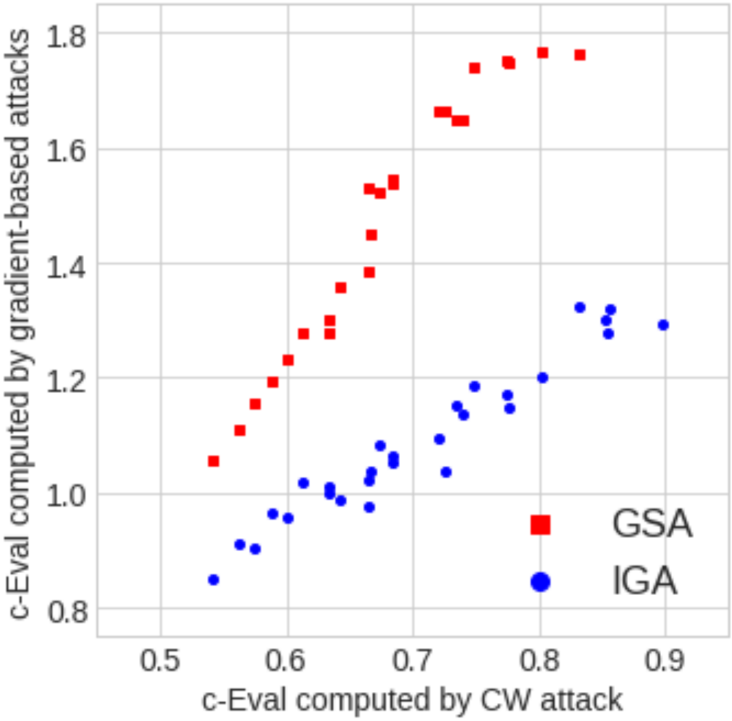

Figure 2: Scatter plot of c-Eval computed by GSA and IGA vs c-Eval computed by CW attack on 30 different images.

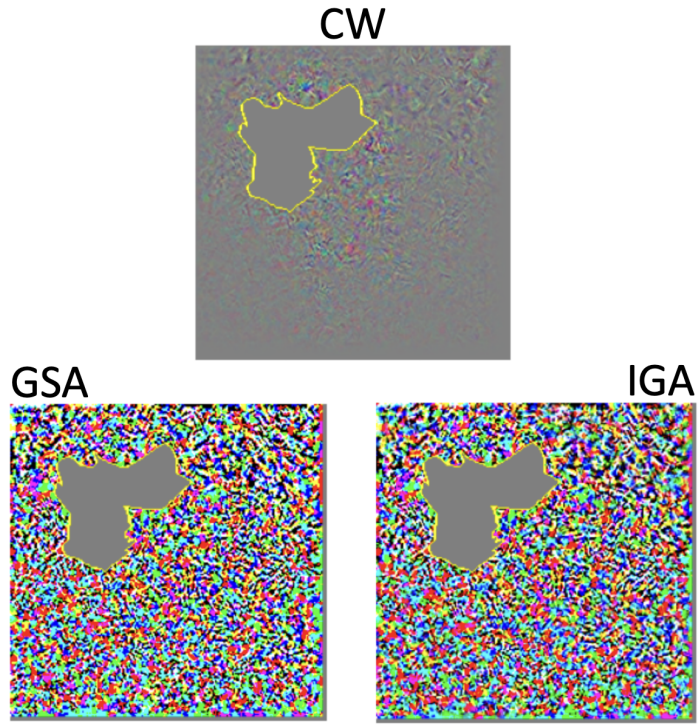

Figure 3: Distortions between different perturbations and the original image of the example in Figure 1.

Relationship with the importance of features

Affine classifiers

c-Eval in affine classifiers

Given an affine classifier $f(x) = W^T x + b$ and an explanation e_x, c-Eval is the solution of the following problem

$$min \ \ ||\delta||_2$$

$$\text{s.t. } \exists j : w_j(x+\delta) + b_j \leq w_{j_0}(x+\delta) + b_{j_0}$$

$$\forall i \in e_i, \delta_i = 0$$

where $j_0 = \underset{j}{arg\,max}\, f(x)$ is the original prediction.

c-Eval determines the **minimum distance** from the **input data** point to the **nearest decision hyperplane** $\mathcal{F}_j = \{x : f_{j_0}(x) = f_j(x)\}$ in the space of the non-explanatory features.

Relationship with the importance of features

2D affine classifiers

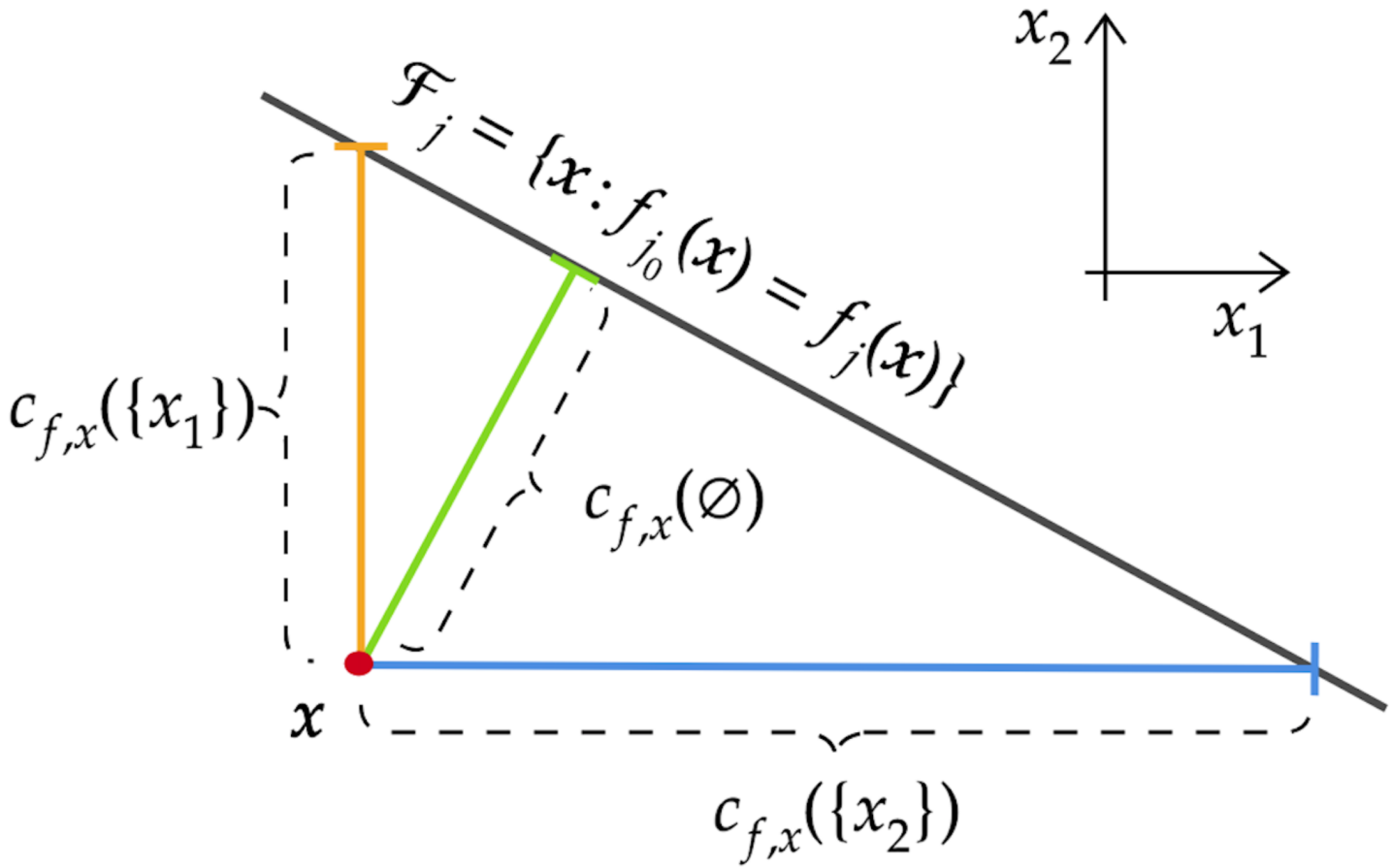

Figure 4: c-Eval in 2D affine classifier.

$c_{f,x}(\{x_1\}) < c_{f,x}(\{x_2\})$ implies that x_2 is more import to the prediction than x_1.
c-Eval reflects the **importance of features** in the explanation.

Relationship with the importance of features

Non-affine classifiers

Hypothesis: Well-known **image classifiers** might be **nearly affine** in a wide-range of local predictions

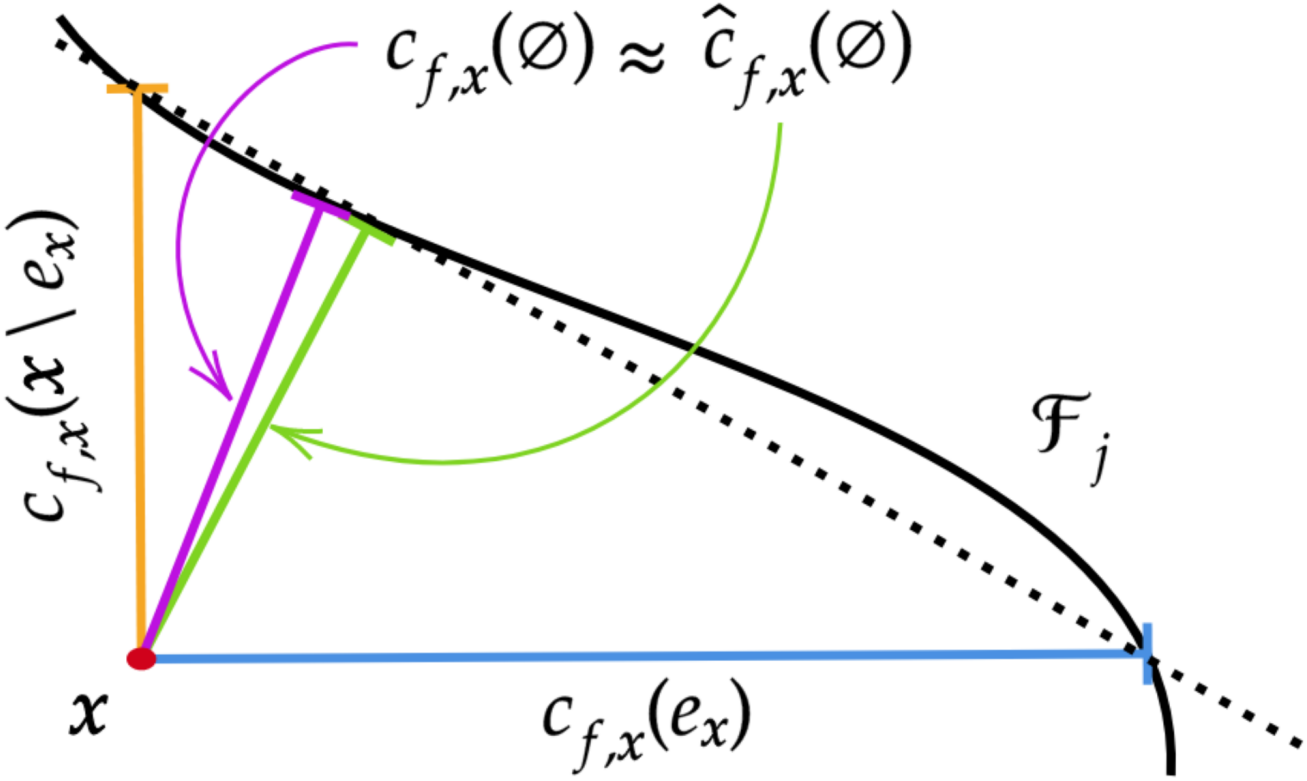

Figure 5: c-Eval in nearly affine classifier

c-Eval and the importance of features

Non-affine classifiers

Empirical validation of the hypothesis:

- Generation of an explanation with Gcam
- Computation of $c_{f,x}(e_x)$, $c_{f,x}(x \setminus e_x)$, $c_{f,x}(\emptyset)$ using CW attack

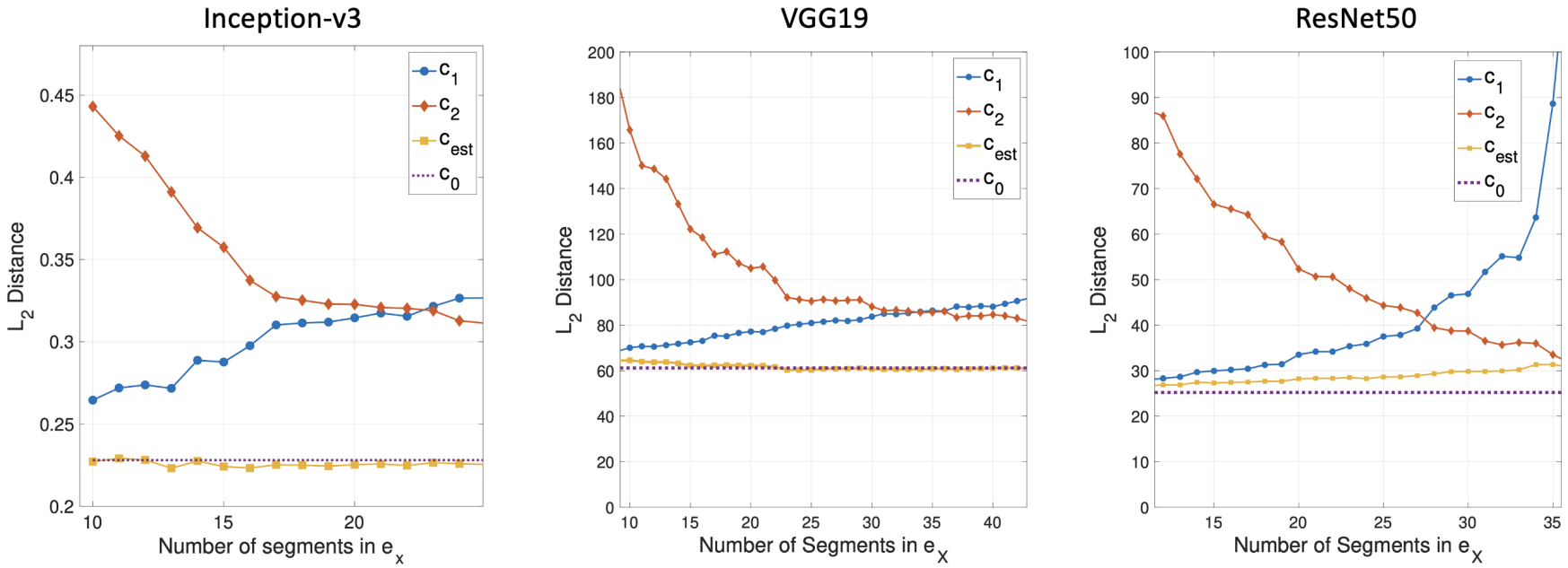

Figure 6: $c_{f,x}(e_x)$, $c_{f,x}(x \setminus e_x)$, $\hat{c}_{f,x}(\emptyset)$ and $c_{f,x}(\emptyset)$ (corresponding to c_1, c_2, c_{est} and c_0 in the legend) when varying the number of explanatory segments.

c-Eval plot

- Plot of the c-Eval values of an explanation varying the number of explanatory features

  Allows to:
  - Visualize the explainer's behaviour
  - Tune the explainer's parameters

c-Eval plot

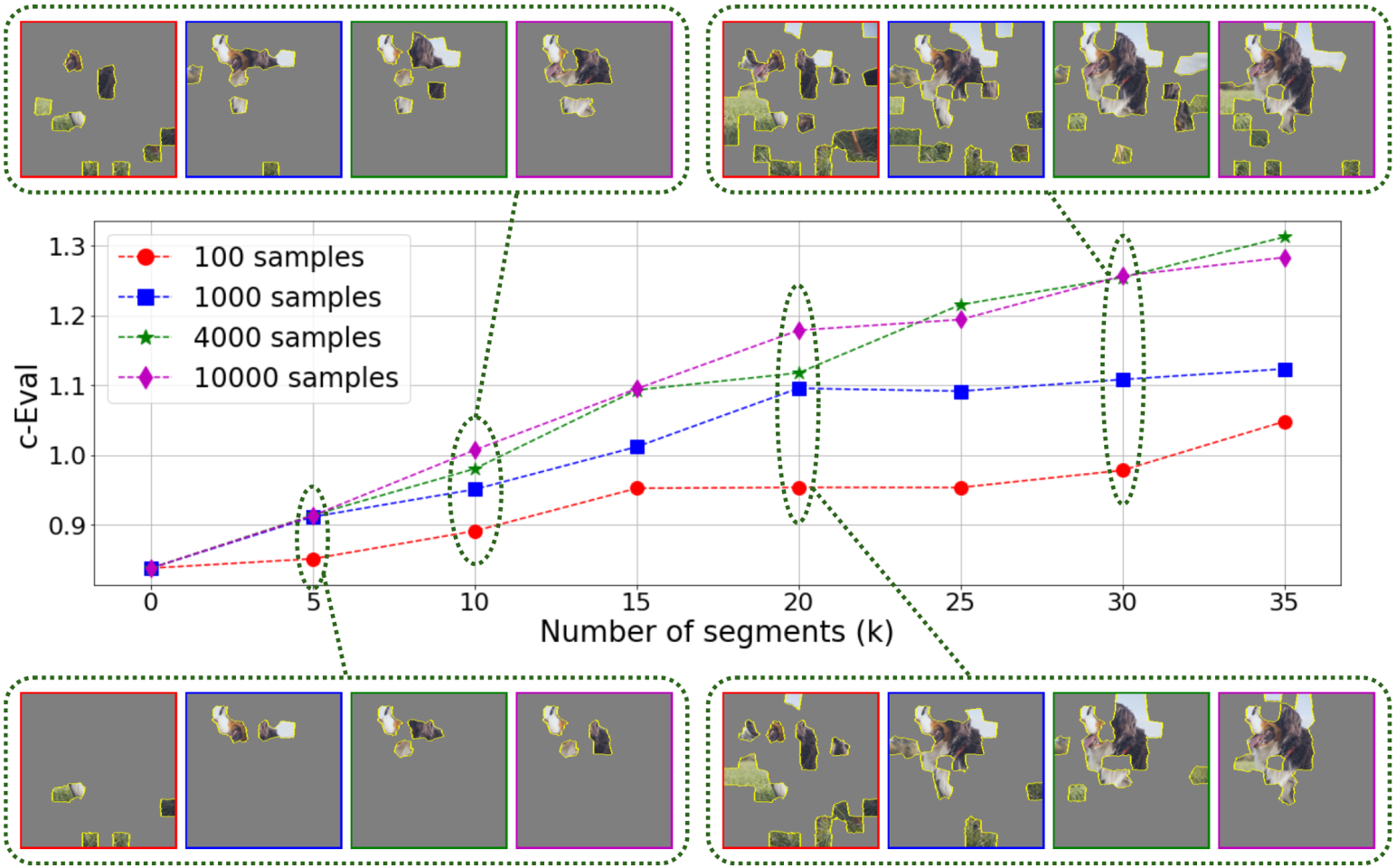

Figure 7: c-Eval plot of LIME with different sample rates.

Experimental results

- c-Eval was used to experimentally evaluate explanations generated by different feature-based local explainers
- To demonstrate the statistical behavior of c-Eval on large number of samples, the ratio between $c_{f,x}(e_x)$ and $c_{f,x}(\emptyset)$ was reported

Experimental results

MNIST dataset

Experiments were conducted with:

- A dataset of 1000 images from MNIST
- 8 different feature-based local explainers
- GSA and IGA to compute c-Eval
- 2 different convolutional neural networks (whose architecture are described in [6] and [2])

Experimental results

MNIST dataset

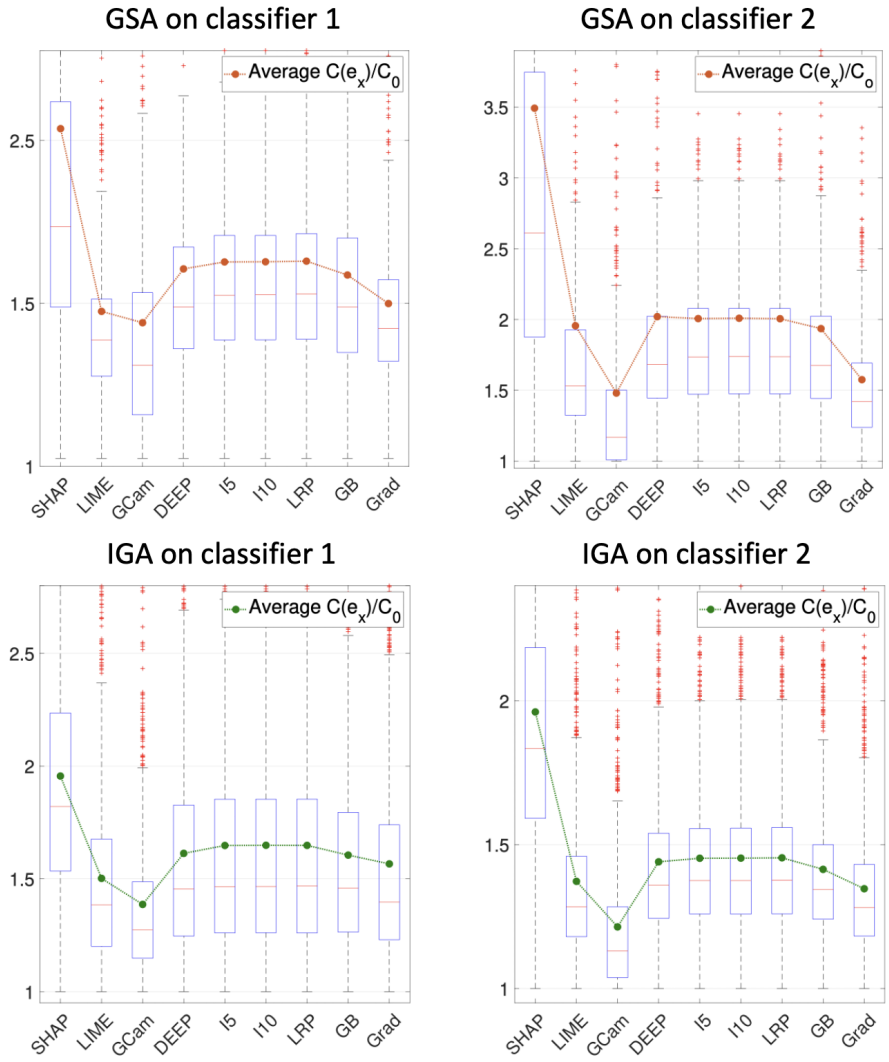

Figure 8: Distributions of the normalized c-Eval.

Experimental results

Caltech101 dataset

Experiments were also conducted with:

- A dataset of 700 images from Caltech101
- 4 different feature-based local explainers
- GSA and IGA to compute c-Eval
- VGG19 as classifier

Experimental results

Caltech101 dataset

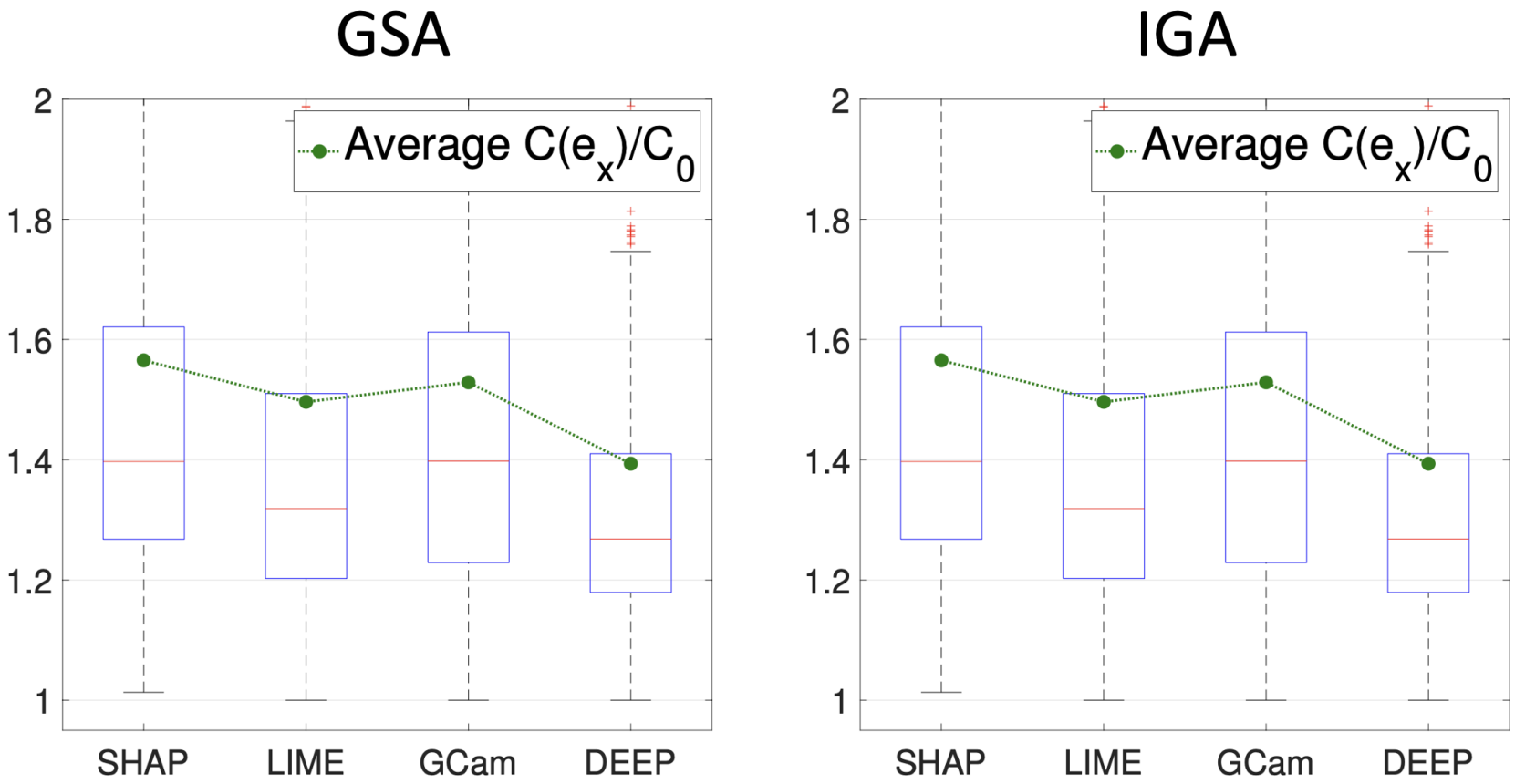

Figure 9: Distributions of the normalized c-Eval.

Conclusion

- c-Eval reflects the importance of features included in the explanation
- Experiments advocates that there is a fundamental difference between perturbation-based explainers and back-propagation explainers
- Additional experiments (not shown here) demonstrate that c-Eval can also be used with adversarial robust models
- c-Eval will offer a much clearer view on predictions made by neural networks

References

[1] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

[2] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*, PMLR, 2017, pp. 3145–3153.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, Ieee, 2017, pp. 39–57.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[5] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, Chapman and Hall/CRC, 2018, pp. 99–112.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.