Auto Encoding Explanatory Examples with Stochastic Paths

Ojeda, C.[§], Sánchez, R.[†‡] Biesner, D. [†ʔ], Cvejoski, K.[†ʔ], Schuecker, J.[ʔ], Bauckhage, C.[ʔ], Georgiev, B.[†]

[†]Competence Center Machine Learning Rhine-Ruhr [‡] University of Bonn [ʔ]Fraunhofer Center for Machine Learning and Fraunhofer IAIS [§]Berlin Center for Machine Learning and TU Berlin

Abstract

In this paper we ask for the main factors that determine a classifier's decision making process and uncover such factors by studying latent codes produced by auto-encoding frameworks. To deliver an explanation of a classifier's behaviour, we propose a method that provides series of examples highlighting semantic differences between the classifier's decisions. These examples are generated through interpolations in latent space. We introduce and formalize the notion of a semantic stochastic path, as a suitable stochastic process defined in feature (data) space via latent code interpolations. We then introduce the concept of semantic Lagrangians as a way to incorporate the desired classifier's behaviour and find that the solution of the associated variational problem allows for highlighting differences in the classifier decision. Very importantly, within our framework the classifier is used as a black-box, and only its evaluation is required.

What is an Explanation?

Recent work [1], however, indicates that saliency maps explanations can be misleading since their results are at times independent of the model, and therefore do not provide explanations for its decisions. The failure to correctly provide explanations by some of these methods lies in their sensibility to feature space changes. We are concerned with the question: *can one find semantic differences which characterize a classifier's decision?*

To explain we mean *to provide textual or visual artifacts that provide qualitative understanding of the relationship between the data points and the model prediction*. Attempts to clarify such a broad notion of explanation require the answers to questions such as:

- What were the main factors in a decision?
- Would changing a certain factor have changed the decision?

By training an auto-encoder one can find a latent code which describes a particular data point. This code will serve as the factors. Our role here is to provide a connection between these latent codes and the classifier's decision. Changes on the code should change the classification decision in a user-defined way.

Explaining Through Examples: A Plaintiff Scenario

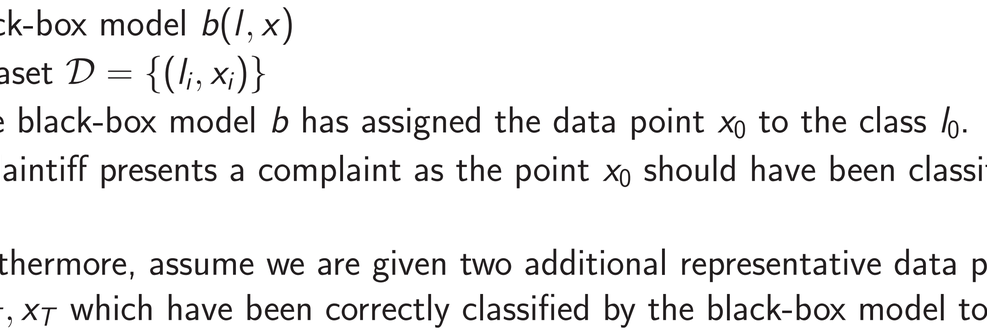
- black-box model $b(l, x)$
- dataset $\mathcal{D} = \{(l_i, x_i)\}$
- The black-box model b has assigned the data point x_0 to the class l_0.
- a plaintiff presents a complaint as the point x_0 should have been classified as l_t.
- Furthermore, assume we are given two additional representative data points x_{-T}, x_T which have been correctly classified by the black-box model to the classes $l_{-T} = l_t, l_T = l_0$

We propose that an explanation why x_0 was misclassified can be articulated through an *example set*
$\mathcal{E} = \{x_{-T}, \ldots, x_0, \ldots, x_T\}$, where $x_t \sim P_\theta(X|Z = z_t)$.
Here $P_\theta(X|Z = z_t)$ is a given decoder distribution and the index t runs over *semantic changes*.

Stochastic Semantic Processes and Corresponding Paths

How to change the codes? In what follows, we first focus on *linear* latent interpolations, i.e.

$$z(t) := t\, z_0 + (1 - t) z_T, \tag{1}$$



In other words, for every pair of points x_0 and x_T in feature space, and its corresponding code samples $z_0 \sim Q_\phi(Z|X = x_0)$ and $z_T \sim Q_\phi(Z|X = x_T)$, the decoder $P_\theta(X|Z)$ induces a measure over the space of paths $\{x(t)|x(0) = x_0, x(T) = x_T\}$.

$$dP_{t_0,\ldots,t_n}(x(t)) := \int_{\mathcal{Z}} \int_{\mathcal{Z}} \left(\prod_{i=1}^{n} p_\theta(x_i|z(t_i)) \right) \times q_\phi(z_0|x_0) q_\phi(z_T|x_T) dz_0 dz_T, \tag{2}$$

Principle of Least Semantic Action

Thus, to design auto-encoding mappings P_θ, Q_ϕ accordingly, we propose an optimization problem of the form

$$\min_{\theta,\phi} S_{P_\theta, Q_\phi}[X_t], \tag{3}$$

where X_t is a stochastic semantic process and S_{P_θ,Q_ϕ} is an appropriately selected functional that extracts certain features of the black-box model $b(l, x)$.
For a given stochastic semantic process X_t, and given initial and final feature "states" x_0 and x_T, we introduce the following function, named the *model-b semantic Lagrangian*

$$\mathcal{L} : [0, 1] \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad (t, x_0, x_T) \mapsto \mathcal{L}[X_t, x_0, x_T], \tag{4}$$

which gives rise to the *semantic model action*:

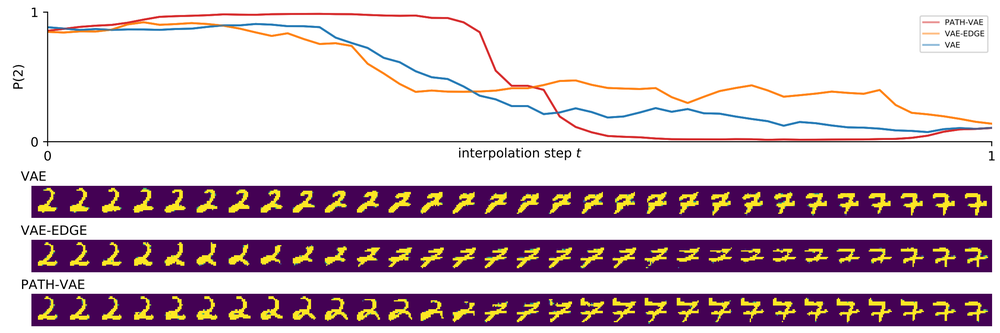
$$S[X_t] := \int_0^T \mathcal{L}[X_t, x_0, x_T] dt. \tag{5}$$

Our problem, viz. to find encoding mappings P_θ, Q_ϕ which yield explainable semantic paths with respect to a black-box model, is then a constrain optimization problem whose total objective function we write as

$$L(\theta, \phi) := L_{\text{VAE}}(\theta, \phi) + \lambda\, \mathbb{E}_{dP[x(t)]} S[x(t)], \tag{6}$$

Lagrangians

- Minimum Hesitant Path $\mathcal{L}_1(x(t), x_0, x_T) := -\left(b(l_T, x(t)) - b(l_0, x(t))\right)^2$
- Minimum Transformation Path $\mathcal{L}_2(x(t), x_0, x_t) := \|\nabla \mathcal{B}(l_T|x(t)) - \alpha \dot{x}(t)\|^2$
- Fix Lenght Path
  $\mathcal{L}_3(x(t), x_0, x_T) = \|\dot{x}(t)\|_g$



Probability Paths for the litigation case $l_0 = 2, l_T = 7$. Y axis corresponds to classification probability and x axis corresponds to interpolation index. Interpolation images for a specific paths are presented below the x axis.

Comparison to other models

Interpolation saliency Map as:

$$S(x_0) = 1/T \int \delta B(x|x_0) \delta x dP[x(t)] = \\ = 1/T \int \left(B(l_T|x(t)) - B(l_0|x(t))\right)(x(t) - x_0) dP[x(t)] \tag{7}$$

We obtained approximations of this integral by using a discrete approximation as performed for the Action.
For a given image x and its corresponding saliency map s, the masking is accomplished by changing the pixels of x which have a saliency value bigger than the τ percentile set of values of the map s itself. We then quantify the change in the odds probability, per number of pixel changed (in percentage values)

$$\log P(l_0|x) = \log P(l_0|x) - \log(1 - P(l_0|x)), \tag{8}$$

In short, a good saliency map will achieve the biggest change in the log odds, with the least amount of pixel changed.

Conclusion

- In the present work we provide a novel framework to explain black-box classifiers through examples obtained from deep generative models.
- We train the auto-encoder, not only by guaranteeing reconstruction quality, but by imposing conditions on its interpolations.
- Beyond the specific problem of generating explanatory examples, our work formalizes the notion of a stochastic process induced in feature space by latent code interpolations, as well as quantitative characterization of the interpolation through the semantic Lagrangian's and actions.

References

[1] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *Advances in Neural Information Processing Systems*, pp. 9525–9536, 2018.