Auto Encoding Explanatory Examples with Stochastic Paths

César Ojeda, Ramsés J. Sánchez, David Biesner, Kostadin Cvejoski, Jannis Schücker, Christian Bauckhage and Bogdan Georgiev

TU Berlin, Fraunhofer IAIS, University of Bonn

December 10, 2020
Motivation

We are concerned with the question: *can one find semantic differences which characterize a classifier’s decision?*
What is an Explanation?

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?
What do we mean by factors?

Let us denote the feature (data) space by $\mathcal{X}$ and the latent linear space of codes (describing the data) by $\mathcal{Z}$, where usually $\text{dim}(\mathcal{Z}) \ll \text{dim}(\mathcal{X})$.

▶ Decoder $P_\theta(X|Z)$
▶ Encoder $Q_\phi(Z|X)$

$$L_{\text{VAE}} = \mathbb{E}_{P_D(X)} - \mathbb{E}_{Q_\phi(Z|X)} \left[ \log p_\theta(x|z) \right] + D_{\text{KL}} \left( Q_\phi(Z|X), P(Z) \right)$$  \hfill (1)

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.
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 $l_{-T}, l_T$

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

In what follows, we first focus on linear latent interpolations, i.e.

\[ z(t) := t z_0 + (1 - t) z_T, \]  

(a) Paths in feature space with black-box classifier level sets.

(b) Procedure for sampling paths in feature space with auto-encoders: interpolations in latent space and decoding of images.

Figure: Auto-Encoding Examples Setup: Given a misclassified point \( x_0 \) and representatives \( x_{-T}, x_T \), we construct suitable interpolations (stochastic processes) by means of an Auto-Encoder.
An Approach via Explicit Family of Measures

The collection of measures prescribed by induces a corresponding continuous-time stochastic process. Moreover, under appropriate reconstruction assumptions on the auto-encoder mappings $P_\theta, Q_\phi$, the sample paths are interpolations, that is, start and terminate respectively at $x_0, x_T$ almost surely.

\[
dP_{t_0,\ldots,t_n}(x(t)) := \int Z \int Z \left( \prod_{i=1}^{n} p_\theta(x_i|z(t_i)) \right) \times q_\phi(z_0|x_0)q_\phi(z_T|x_T)dz_0dz_T, \tag{3}
\]

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 \{\(x(t)|x(0) = x_0, x(T) = x_T\}\).
Principle of Least Semantic Action

Thus, to design auto-encoding mappings $P_{\theta}, Q_{\phi}$ accordingly, we propose an optimization problem of the form

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

(4)

where $X_t$ is a stochastic semantic process and $S_{P_{\theta}, Q_{\phi}}$ 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

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

(5)

which gives rise to the semantic model action:

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

(6)
Objective Function

Our problem, viz. to find encoding mappings $P_\theta, Q_\phi$ 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

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

(7)

where $L_{\text{VAE}}$ is given by eq. (1), $S[x(t)]$ corresponds to the Lagrangian action and $\lambda$ is an hyper parameter controlling the action’ scale. The average over the paths is taken with respect to the stochastic paths and the corresponding measure $dP[x(t)]$, that is, the path integral

$$\mathbb{E}_{dP[x(t)]} S[(x(t))] = \int \mathcal{L}[x(t), x_0, x_T] dP[x(t)]$$

(8)

$$\approx \frac{1}{nK} \sum_k \sum_t \mathcal{L}[x^k_t, x_0, x_T],$$

(9)
Lagrangians

- **Minimum Hesitant Path**
  \[ \mathcal{L}_1(x(t), x_0, x_T) := - (b(l_T, x(t)) - b(l_0, x(t)))^2 \]

- **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 Length Path**
  \[ \mathcal{L}_3(x(t), x_0, x_T) = \| \dot{x}(t) \|_g \]
The classifier must change decision in a monotonous fashion. These can be enforced in a straightforward way by introducing the constraint:

\[ r_e = \frac{d}{dt} \mathcal{B}(l_T|x(t)) < 0, \quad \forall t \in [0, T]. \quad (10) \]

Note that this constraint requires differentiability of \( x(t) \) - in contrast, the notion of explanatory path is not relying on such. We approximate the differential with finite differences. Our final loss reads:

\[
L(\theta, \phi) := L_{VAE}(\theta, \phi) + \lambda \mathbb{E}_{dP[x(t)]} S_1[x(t)] + \lambda_m r_m + \lambda_e r_e,
\]  

(11)
Example

Figure: 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:

\[ S(x_0) = \frac{1}{T} \int \delta B(x|x_0) \delta x dP[x(t)] = \]

\[ = \frac{1}{T} \int (B(l_T|x(t)) - B(l_0|x(t))) (x(t) - x_0) dP[x(t)] \quad (12) \]

We obtained approximations of this integral by using a discrete approximation as performed for the Action.
Evaluation

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 $\tau$ 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)

$$\log P(l_0|x) = \log P(l_0|x) - \log(1 - P(l_0|x)),$$

(13)

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

Table: Relevance Statistics for Different Models and Comparison Saliency Maps. In parenthesis we include the value of the $\lambda$ regularizers

<table>
<thead>
<tr>
<th>index</th>
<th>max</th>
<th>min</th>
<th>mean</th>
<th>random</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE DL (1.0) FL (5.0)</td>
<td>26.890537</td>
<td>13.248348</td>
<td>13.780094</td>
<td>24.259376</td>
</tr>
<tr>
<td>VAE FL (5.0)</td>
<td>17.528036</td>
<td>12.430488</td>
<td>13.249453</td>
<td>16.323000</td>
</tr>
<tr>
<td>VAE MTL (5.0)</td>
<td>22.425894</td>
<td>1.593850</td>
<td>17.865745</td>
<td>18.878264</td>
</tr>
<tr>
<td>VAE DL (1.0) MTL (5.0)</td>
<td>41.968799</td>
<td>3.516279</td>
<td>25.474598</td>
<td>41.076384</td>
</tr>
<tr>
<td>WAE FL (5.0)</td>
<td>1.348628</td>
<td>7.626165</td>
<td>4.758138</td>
<td>2.598853</td>
</tr>
<tr>
<td>WAE DL (1.0) FL (5.0)</td>
<td>28.650618</td>
<td>19.274864</td>
<td>13.260943</td>
<td>24.217626</td>
</tr>
<tr>
<td>WAE MTL (5.0)</td>
<td>33.308588</td>
<td>1.721710</td>
<td>10.395801</td>
<td>27.469413</td>
</tr>
<tr>
<td>WAE DL (1.0) MTL (5.0)</td>
<td>21.113389</td>
<td>25.378131</td>
<td>6.343344</td>
<td>16.944631</td>
</tr>
<tr>
<td>vanilla</td>
<td>18.799356</td>
<td>12.129845</td>
<td>12.124648</td>
<td>18.617377</td>
</tr>
<tr>
<td>smooth</td>
<td>2.626274</td>
<td>16.802856</td>
<td>10.184854</td>
<td>3.966701</td>
</tr>
<tr>
<td>guided</td>
<td>25.264783</td>
<td>4.653241</td>
<td>2.523255</td>
<td>15.527908</td>
</tr>
<tr>
<td>mask</td>
<td>4.276590</td>
<td>0.248701</td>
<td>3.414551</td>
<td>3.244211</td>
</tr>
</tbody>
</table>
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.