A Formal Approach to Explainability

Lior Wolf[1,2] **Tomer Galanti**[2] **Tamir Hazan**[3]

[1]Facebook AI Research
[2]The School of Computer Science, Tel Aviv University
[3]Technion

Abstract

We regard explanations as a blending of the input sample and the model's output and offer a few definitions that capture various desired properties of the function that generates these explanations. We study the links between these properties and between explanation-generating functions and intermediate representations of learned models and are able to show, for example, that if the activations of a given layer are consistent with an explanation, then so do all other subsequent layers. In addition, we study the intersection and union of explanations as a way to construct new explanations.

Introduction

Machine learning is often concerned with tacit knowledge, and tacit knowledge leads to black box models. Given a learned model, one cannot "crack it open" in the hope to understand all of the internal nuts and bolts. Explaining the model often relies, instead, on communicating, in a way that is understandable to humans, an internal state of the model during computation.

An explanation process, therefore, has three components: the input, the model's output for that input, which needs to be justified, and an internal state of the model. The explanation itself combines the input and the output into a joint sample that should be understandable by human users. The explaining function (EF) generates these explanations, based on the two inputs, and is intimately tied to the model it explains. We can expect, therefore, that the generated explanations are linked to internal states of the model.

For example, consider a mapping from images to labels of objects. The explanation often takes the visual form of an image, where the predicted object is highlighted and the features related to the label are emphasized, see, e.g., (Zeiler and Fergus 2014). The algorithmic way to explain, is to generate this hybrid image from the internal representation of the black-box model. Another form of explanation is a textual one (Hendricks et al. 2016), and describes features that belong to the recognized class. For example, "[this is an image of a broccoli since] it is green, has a flowering head, and a thick stem with small leaves", where the part in brackets is the label, but not the explanation. This explanation is both a function of the input image (describes what can be seen and where) and the label (contains known properties of broccolis).

We provide a formal framework that captures various desiderata of explanations, among which are: consistency between an internal model's state and the generated explanation, explainability of an internal state, validity of an explanation, and its completeness.

Our main results link various aspects of the properties. For example, a valid explanation has to be complete. We also study the specific case of explaining, using the gradient of the loss, the predictions of multiclass neural networks and show that the explanation is linked to the learned representation. Lastly, we study the intersection and unions of explanations, as a way to create new explanations by combining existing ones.

Settings

We describe a few fundamental concepts in a way that is less formal than what is presented in the subsequent sections. An illustration of the main components of our framework is given in Fig. 1.

What do we want to explain? Given a function $h : \mathcal{X} \to \mathcal{Y}$ from the input domain $\mathcal{X}$ to the output domain $\mathcal{Y}$, we would like to explain the output $h(x)$ for some input $x \in \mathcal{X}$. h is typically a learned model.

What is an explanation? An explanation is a blending of the input and the output. An explanation function (EF for short) $g : \mathcal{X} \times \mathcal{Y} \to G$ maps $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ to $g(x, y)$, which is the explanation for (x, y) in the blended domain G. Hopefully, the elements of domain G are understandable to humans. However, this part is not amendable to formalization.

Consistent representation: Given a function $h : \mathbb{R}^n \to \mathcal{Y}$ of the form $h = c \circ f$, where f is some representation of the input and c a classifier on top of it, we would like to discuss the link between f and an EF g. We say that f is consistent with respect to an EF g, if for all $x_1, x_2 \in \mathcal{X}$, such that: $|g(x_1, h(x_1)) - g(x_2, h(x_2))| \leq \epsilon$, we have: $|f(x_1) - f(x_2)| \leq \beta(\epsilon)$.

Explainable representation: This definition is similar to consistency, with a reversed implication. We say that f is explainable with respect to the EF g if for all $x_1, x_2 \in \mathcal{X}$, such that: $|f(x_1) - f(x_2)| \leq \epsilon$, we have: $|g(x_1, h(x_1)) - g(x_2, h(x_2))| \leq \gamma(\epsilon)$.

$$\begin{tikzcd}
 & & y(x) \\
x \sim D \arrow[urr, "y"] \arrow[r, "f"] \arrow[dr] & f(x) \arrow[r, "c"] & h(x) \arrow[dl] \\
 & g(x, h(x)) &
\end{tikzcd}$$

Figure 1: The main components of our framework. The EF g is a function of input x and the models' label $h(x)$, which approximates the target function y. h is a composition of some representation f and a classifier c. Note that g should generate explanations for a specific h and is not generic.

Equivalence between an EF and a representation: A representation f is equivalent to an EF g, if it is both consistent with it and explainable by it.

Valid explanation: An EF g is valid, if there exists a function t, such that the model's label is predictable from the explanation $t(g(x, h(x))) \approx h(x)$.

Complete explanation: We say that an EF g is complete in the context of a model h, if there is no information left in the input x that is relevant to h, which is independent of the information in $g(x, h(x))$. If we define as $\bar{g}(x, h(x))$ all the information that is the part of x but which has no information on $g(x, h(x))$, then g is complete if there is no function s such that $s(\bar{g}(x, h(x))) \approx h(x)$.

Intersection and Union of EFs: Given a model $h : \mathbb{R}^n \to \mathcal{Y}$ and two EFs g_1, g_2, the intersection between them is a representation $u(x, h(x))$ such that we can write $r_1(g_1(x, h(x))) = (e_1(x, h(x)), u(x, h(x)))$ and $r_2(g_2(x, h(x))) = (e_2(x, h(x)), u(x, h(x)))$, where r_1, r_2 are invertible transformations and e_1 is the part of g_1 that is independent of g_2 (and vice versa for e_2). The union between them is defined as $(e_1(x, h(x)), u(x, h(x)), e_2(x, h(x)))$.

A Formal Model

In this section, we present our formal model of explainability. The sample space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^n$ is the inputs space and $\mathcal{Y}$ is the outputs space. For instance, in binary classification, $\mathcal{Y} = \{\pm 1\}$, in multi-class classification $\mathcal{Y} = \{1, \dots, K\} := [K]$ for some $K \in \mathbb{N}$, and in regression, $\mathcal{Y} = \mathbb{R}$. In addition, there is an unknown target function $y : \mathbb{R}^n \to \mathcal{Y}$ that is being learned and a hypothesis class $\mathcal{H}$ of models $h : \mathbb{R}^n \to \mathcal{Y}$ from which the learning algorithm selects an approximation of the target function y. We denote by D the distribution of data samples in $\mathcal{X}$.

We consider a family of EFs $\mathcal{G}$, and each EF $g \in \mathcal{G}$ is a mapping $g : \mathbb{R}^n \times \mathcal{Y} \to G$. Here, G is a set of possible explanations. We do not aim to show how to compute an explanation $g(x, h(x))$. Instead, we focus on providing useful terminology to understand the properties of EFs.

Terminology and notations Before we present our main results, we recall a few technical notations. First, throughout this manuscript, we will assume that D is supported by $\mathcal{X}$, which is assumed, for the purpose of simplifying entropy and mutual-information based arguments, to be a discrete set. We also assume that all logarithms are base 2. The image of a function $f : \mathcal{X}_1 \to \mathcal{X}_2$ is denoted by $f(\mathcal{X}_1)$. We denote by, $\ell : \mathcal{Z} \to \mathbb{R}$ a loss function. Typically, in binary classification, we have the zero-one loss, $\ell(y_1, y_2) = \mathbb{1}[y_1 \neq y_2]$ and in regression, we often employ the L1 loss $|y_1 - y_2|_1$ or the L2 loss $|y_1 - y_2|^2$. Here, $\mathbb{1}[b]$ is an indicator of a boolean variable, b, being true, i.e., $\mathbb{1}[\text{true}] = 1$ and $\mathbb{1}[\text{false}] = 0$.

We recall the classical information theoretic notations from (Cover and Thomas 2006): the expectation and probability operators symbols $\mathbb{E}, \mathbb{P}$, the Shannon entropy (discrete or continuous) $H(X) := -\mathbb{E}_X[\log \mathbb{P}[X]]$, the conditional entropy $H(X|Y) := H(X, Y) - H(Y)$ and the (conditional) mutual information (discrete or continuous) $I(X; Y|Z) := H(X|Z) - H(X|Y, Z)$. For a given value $p \in [0, 1]$, we denote, $H(p) = -p \log(p) - (1-p) \log(1-p)$.

Properties of EFs

We provide formal definitions to the various properties mentioned in the Settings Section. A representation of the input is a function $f : \mathcal{X} \to \mathbb{R}^d$ (for some $d > 0$). In most cases, we will assume that f is a sub-architecture of our mapping $h : \mathbb{R}^n \to \mathcal{Y}$. Specifically, we would consider h to be a composite function that is built in layers $h = p_k \circ \cdots \circ p_1$, where each layer p_i is a function $p_i : \mathbb{R}^{n^{i-1}} \to \mathbb{R}^{n^i}$ (for some $k, n^i \in \mathbb{N}$, n^0 being the input dimension n and $i \in \{1, 2 \dots, k\}$). In this case, f would contain the first m layers $f = p_m \circ \cdots \circ p_1$ and c would contain the $k - m$ top layers: $c = p_k \circ \cdots \circ p_{m+1}$.

Definition 1 (Consistent Representation). *Let $h = c \circ f \in \mathcal{H}$ be a model, $g : \mathcal{Z} \to G$ an EF and $\beta : (0, \infty) \to [0, \infty)$. We say that f is a $\beta(\epsilon)$-consistent representation with respect to g, if for any $\epsilon \in (0, \infty)$ and $x_1, x_2 \in \mathcal{X}$, we have:*

$$\begin{aligned} &|g(x_1, h(x_1)) - g(x_2, h(x_2))| \leq \epsilon \\ \Longrightarrow\ &|f(x_1) - f(x_2)| \leq \beta(\epsilon) \end{aligned} \tag{1}$$

Definition 2 (Explainable Representation). *Let $h = c \circ f \in \mathcal{H}$ be a model and $g : \mathcal{Z} \to G$ an EF. For a given function $\gamma : (0, \infty) \to (0, \infty)$, we say that f is a $\gamma(\epsilon)$-explainable representation with respect to g, if for any $\epsilon \in (0, \infty)$ and $x_1, x_2 \in \mathcal{X}$, we have:*

$$\begin{aligned} &|f(x_1) - f(x_2)| \leq \epsilon \\ \Longrightarrow\ &|g(x_1, h(x_1)) - g(x_2, h(x_2))| \leq \gamma(\epsilon) \end{aligned} \tag{2}$$

Additionally, for a given function $\gamma : (0, \infty) \times (0, \infty) \to (0, \infty)$, we say that f is second-order $\gamma(\epsilon_0, \epsilon_1)$-explainable with respect to g, if for any $\epsilon_0, \epsilon_1 \in (0, \infty)$ and $x_1, x_2 \in \mathcal{X}$, we have:

$$\begin{aligned} &|f(x_1) - f(x_2)| \leq \epsilon_0 \text{ and } \left| \frac{\partial f(x_1)}{\partial x_1} - \frac{\partial f(x_2)}{\partial x_2} \right| \leq \epsilon_1 \\ \Longrightarrow\ &|g(x_1, h(x_1)) - g(x_2, h(x_2))| \leq \gamma(\epsilon_0, \epsilon_1) \end{aligned} \tag{3}$$

Definition 3 (Equivalence between a Representation and an EF)**.** *Let $h = c \circ f \in \mathcal{H}$ be a model, $g : \mathcal{Z} \to G$ an EF and $\beta, \gamma : (0, \infty) \to [0, \infty)$. We say that f is $(\beta(\epsilon), \gamma(\epsilon))$-equivalent to g, if it is $\beta(\epsilon)$-consistent and $\gamma(\epsilon)$-explainable with respect to g.*

Definition 4 (Valid EF)**.** *Let $h \in \mathcal{H}$ be a model, $g : \mathcal{Z} \to G$ an EF, $\epsilon_0 > 0$ a fixed constant and $x \sim D$. We say that g is ϵ_0-valid with respect to h, if there is a function $t : G \to \mathcal{Y}$ that satisfies:*

$$\mathbb{E}_x[\ell(t(g(x, h(x))), h(x))] \leq \epsilon_0 \tag{4}$$

Definition 5 (Complete EF)**.** *Let $h \in \mathcal{H}$ be a model, $g : \mathcal{Z} \to G$ an EF and $x \sim D$. Let $\alpha, \epsilon > 0$ be two constants. We say that g is (ϵ, α)-complete with respect to h, if every function $\bar{g} : \mathcal{X} \to \mathbb{R}^d$, such that, $I(g(x, h(x)); \bar{g}(x)) \leq \epsilon$ and function $s : \mathbb{R}^d \to \mathcal{Y}$, we have:*

$$\mathbb{E}_x[\ell(s(\bar{g}(x)), h(x))] \geq \alpha \tag{5}$$

Linking Representations and EFs

The following theorem states that if an internal representation of a layered model h is $\beta(\epsilon)$-consistent with an EF, then, under mild conditions, downstream layers are also consistent with the specified EF.

Theorem 1. *Let $h = p_k \circ \cdots \circ p_1 : \mathbb{R}^n \to \mathcal{Y}$ be a model and $g : \mathcal{Z} \to G$ an EF. Assume that $f_i := p_i \circ \cdots \circ p_1$ is $\beta(\epsilon)$-consistent with respect to g, for some $i \in \{1, \ldots, k\}$. Assume that p_r is a l_r-Lipschitz function for every $r \in \{i+1, \ldots, j\}$. Then, $f_j := p_j \circ \cdots \circ p_1$ is $\hat{\beta}(\epsilon)$-consistent with respect to g, for $\hat{\beta}(\epsilon) := \beta(\epsilon) \cdot \prod_{r=i+1}^{j} l_r$.*

Proof. Assume that f_i is $\beta(\epsilon)$-consistent for some $i \in \{1, \ldots, k\}$. Let $x_1, x_2 \in \mathcal{X}$ be two inputs, such that, $|g(x_1, h(x_1)) - g(x_2, h(x_2))| \leq \epsilon$. Then, for every $j \in \{i, \ldots, k\}$, we have:

$$\begin{aligned} &|f_j(x_1) - f_j(x_2)| \\ =&|p_j \circ \cdots \circ p_{i+1} \circ f_i(x_1) - p_j \circ \cdots \circ p_{i+1} \circ f_i(x_2)| \\ \leq& \prod_{r=i+1}^{j} l_r |f_i(x_1) - f_i(x_2)| \leq \beta(\epsilon) \cdot \prod_{r=i+1}^{j} l_r = \hat{\beta}(\epsilon) \end{aligned} \tag{6}$$

Since each p_r is a l_r-Lipschitz continuous function for every $r \in \{i+1, \ldots, j\}$. □

One implication of this result is that if a layer of a neural network model h is consistent with an explanation g, then h itself is also consistent, i.e., in the case where any of the layers of h is consistent with g, then if $g(x, h(x))$, which is a function of $h(x)$ as well as of x, does not change much when replacing x with x', then $h(x)$ and $h(x')$ are similar.

The following theorem deals with upstream layers: under mild assumptions, if f is an explainable representation, that is obtained as a layer of a neural network model h, then so are the previous layers in this network.

Theorem 2. *Let $h = p_k \circ \cdots \circ p_1 : \mathbb{R}^n \to \mathcal{Y}$ be a model and $g : \mathcal{Z} \to G$ an EF. Assume that $f_i := p_i \circ \cdots \circ p_1$ is $\gamma(\epsilon)$-explainable with respect to g, for some $i \in \{1, \ldots, k\}$. Assume that p_r is a l_r-Lipschitz function for every $r \in \{j+1, \ldots, i\}$. Then, $f_j := p_j \circ \cdots \circ p_1$ is $\hat{\gamma}(\epsilon)$-explainable with respect to g, for $\hat{\gamma}(\epsilon) := \gamma\left(\epsilon \cdot \prod_{r=j+1}^{i} l_r\right)$.*

Proof. Assume that f_i is $\gamma(\epsilon)$-explainable for some $i \in \{1, \ldots, k\}$. Let $x_1, x_2 \in \mathcal{X}$ be two inputs, such that, $|f_j(x_1) - f_j(x_2)| \leq \epsilon$. Then,

$$\begin{aligned} &|f_i(x_1) - f_i(x_2)| \\ =&|p_i \circ \cdots \circ p_{j+1} \circ f_j(x_1) - p_i \circ \cdots \circ p_{j+1} \circ f_j(x_2)| \\ \leq& \prod_{r=j+1}^{i} l_r |f_j(x_1) - f_j(x_2)| \leq \epsilon \cdot \prod_{r=j+1}^{i} l_r \end{aligned} \tag{7}$$

Since each p_r is a l_r-Lipschitz continuous function for every $r \in \{j+1, \ldots, i\}$. Therefore, since f_i is $\gamma(\epsilon)$-explainable with respect to g, we have:

$$|g(x_1, h(x_1)) - g(x_2, h(x_2))| \leq \gamma\left(\epsilon \cdot \prod_{r=j+1}^{i} l_r\right) \tag{8}$$

□

Note that an immediate implication is that if a representation is explainable by g, then so is the input x itself.

A Specific Case Study

We next treat a specific case, which is the conventional multi-class classification approach for deep neural networks, coupled with the iconic image-based explanation that is given by the derivative of the output neuron associated with the predicted label by the input. In this case, the model predicts the label based on an $\arg\max$ of multiple 1D linear projections (m_i, i being the index of the label) of the activations of the penultimate layer $p(x)$ for some input x. The explanation of the prediction $h(x)$ is then given as the matrix derivative of $(m_{h(x)}^\top \cdot p(x))$ by the input x.

The following theorem states that if our model is of the form $h(x) = \arg\max_{i \in \mathcal{Y}} (m_i^\top \cdot p(x))$ and our EF has the form $g(x, h(x)) = \frac{\partial(m_{h(x)}^\top \cdot p(x))}{\partial x}$, where $p = c \circ f$ such that c, f and the derivative of c are Lipschitz continuous functions, then, f is explainable with respect to g.

Theorem 3. *Let $\mathcal{Y} = [K]$ and $h : \mathbb{R}^n \to \mathcal{Y}$ a model of the form, $h(x) = \arg\max_{i \in \mathcal{Y}} m_i^\top \cdot p(x)$, where $p : \mathbb{R}^n \to \mathbb{R}^d$ and $m_i \in \mathbb{R}^d$, for $i \in [K]$. Let $g(x, h(x)) = \frac{\partial(m_{h(x)}^\top \cdot p(x))}{\partial x}$ be an EF. Assume that for all $i \in [K]$, $p = c \circ f$, such that: c, $\frac{\partial c(x)}{\partial x}$, $\frac{\partial p(x)}{\partial x}$ and f are Lipschitz continuous functions. Additionally, assume that: $\forall i \neq j \in [K], x \in \mathcal{X} : m_i^\top \neq m_j^\top$ and $\forall x \in \mathcal{X} : |p(x)| \geq \Delta$, for some constant $\Delta > 0$. Then, f is second-order $\mathcal{O}(\epsilon_0 + \epsilon_1)$-explainable with respect to g.*

Proof. Assume that for all $i \in [K]$:

$$|f(x_1) - f(x_2)| \leq \epsilon_0 \text{ and } \left|\frac{f(x_1)}{\partial x_1} - \frac{f(x_2)}{\partial x_2}\right| \leq \epsilon_1 \tag{9}$$

Then, since each c is a Lipschitz continuous function, there is a constant $l_1, \dots, l_K > 0$, such that for all $i \in [K]$ and $x_1, x_2 \in \mathbb{R}^n$:

$$\begin{aligned} &|m_i^\top \cdot p(x_1) - m_i^\top \cdot p(x_2)| \\ =&|m_i^\top| \cdot |p(x_1) - p(x_2)| \leq l \cdot |m_i^\top| \cdot \epsilon_0 \end{aligned} \tag{10}$$

For any small enough $\epsilon_0 > 0$, we have:

$$l \cdot \max_{i \in [K]} |m_i^\top| \cdot \epsilon_0 < \min_{i \neq j} |m_i^\top - m_j^\top| \cdot \Delta/2 \tag{11}$$

Since $\forall i \neq j \in [K], x \in \mathcal{X} : |p(x)| \geq \Delta$, we have:

$$|m_i^\top \cdot p(x) - m_j^\top \cdot p(x)| \geq \min_{i \neq j} |m_i^\top - m_j^\top| \cdot \Delta > 0 \tag{12}$$

In this case, if $h(x_1) = i$, then, for all $j \in [K]$, such that $j \neq i$, we have:

$$\begin{aligned} &m_i^\top \cdot p(x_2) - m_j^\top \cdot p(x_2) \\ \geq& m_i^\top \cdot p(x_1) - |m_i^\top \cdot p(x_1) - m_i^\top \cdot p(x_2)| \\ &- m_j^\top \cdot p(x_2) - |m_j^\top \cdot p(x_1) - m_j^\top \cdot p(x_2)| \\ \geq& \min_{i \neq j} |m_i^\top - m_j^\top| \cdot \Delta - 2l \cdot \max_{i \in [K]} |m_i^\top| \cdot \epsilon_0 > 0 \end{aligned} \tag{13}$$

Therefore, we conclude that: $h(x_1) = h(x_2) = i$. Thus,

$$\begin{aligned} &|g(x_1, h(x_1)) - g(x_2, h(x_2))| \\ =&\Big|\frac{\partial(m_{h(x_1)}^\top \cdot p(x_1))}{\partial x_1} - \frac{\partial(m_{h(x_2)}^\top \cdot p(x_2))}{\partial x_2}\Big| \\ =&\Big|\frac{\partial(m_i^\top \cdot p(x_1))}{\partial x_1} - \frac{\partial(m_i^\top \cdot p(x_2))}{\partial x_2}\Big| \\ =&|m_i^\top| \cdot \Big|\frac{\partial p(x_1)}{\partial x_1} - \frac{\partial p(x_2)}{\partial x_2}\Big| \\ =&\Big|\frac{\partial c(f(x_1))}{\partial f(x_1)} \cdot \frac{\partial f(x_1)}{\partial x_1} - \frac{\partial c(f(x_2))}{\partial f(x_2)} \cdot \frac{\partial f(x_2)}{\partial x_2}\Big| \end{aligned} \tag{14}$$

Since c and f are Lipschitz continuous functions, we have:

$$\begin{aligned} &|g(x_1, h(x_1)) - g(x_2, h(x_2))| \\ =&\mathcal{O}\left(\Big|\frac{\partial f(x_1)}{\partial x_1} - \frac{\partial f(x_2)}{\partial x_2}\Big|\right) \\ &+ \mathcal{O}\left(\Big|\frac{\partial c(f(x_1))}{\partial f(x_1)} - \frac{\partial c(f(x_2))}{\partial f(x_2)}\Big|\right) \end{aligned} \tag{15}$$

Since $\frac{\partial c(u)}{\partial u}$ is also a Lipschitz continuous function, we have:

$$\begin{aligned} &|g(x_1, h(x_1)) - g(x_2, h(x_2))| \\ =&\mathcal{O}(\epsilon_1 + |f_i(x_1) - f_i(x_2)|) = \mathcal{O}(\epsilon_0 + \epsilon_1) \end{aligned} \tag{16}$$

□

Validity and Completeness

The next result shows that if an EF is valid, then it is also complete. The intuition behind this result is, if we are able to recover $h(x)$ from $\bar{g}(x)$ and from $g(x, h(x))$, then, $\bar{g}(x)$ and $g(x, h(x))$ cannot be independent of each other.

Theorem 4 (Valid $\implies$ Complete)**.** *Let $h : \mathbb{R}^n \to \mathcal{Y}$ be a model, $g : \mathcal{Z} \to G$ an ϵ_0-valid EF for some constant $\epsilon_0 \in (0, 0.5)$ and $x \sim D$. Assume that $\mathcal{Y} = \{\pm 1\}$ and denote, $p := \mathbb{P}[h(x) = 1]$. Then, g is (ϵ, α)-complete with respect to h, with $\alpha := \frac{\sqrt{1 + H(p)(H(p) - \epsilon - 2\sqrt{\epsilon_0})} - 1}{H(p)}$ and any $\epsilon > 0$ that satisfies, $H(p) > \epsilon + 2\sqrt{\epsilon_0}$. In particular, if $p = 1/2$, we have: $\alpha = \sqrt{2 - \epsilon - 2\sqrt{\epsilon_0}} - 1$.*

Proof. Let $\bar{g} : \mathcal{X} \to \mathbb{R}^d$ be a function, such that, $I(\bar{g}(x); g(x, h(x))) \leq \epsilon$. Since $g(x, h(x))$ is ϵ_0-valid, there is a function $t : G \to \mathcal{Y}$, that satisfies:

$$\begin{aligned} &\mathbb{P}[t(g(x, h(x))) \neq h(x)] \\ =&\mathbb{E}_x[\ell(t(g(x, h(x))), h(x))] \leq \epsilon_0 < 1/2 \end{aligned} \tag{17}$$

By $I(X; f(Y)) \leq I(X; Y)$, for every function f, we have:

$$I(\bar{g}(x); t(g(x, h(x)))) \leq I(\bar{g}(x); g(x, h(x))) \leq \epsilon \tag{18}$$

By Lem. 3 in the Appendix,

$$\begin{aligned} &|I(\bar{g}(x); h(x)) - I(\bar{g}(x); t(g(x, h(x))))| \\ \leq& H(\mathbb{P}[t(g(x, h(x))) \neq h(x)]) \end{aligned} \tag{19}$$

Therefore, by Lem. 4, we have:

$$\begin{aligned} I(\bar{g}(x); h(x)) &\leq \epsilon + H(\mathbb{P}[t(g(x, h(x))) \neq h(x)]) \\ &\leq \epsilon + 2\sqrt{\epsilon_0} \end{aligned} \tag{20}$$

Next, we assume that $H(q) > \epsilon + 2\sqrt{\epsilon_0}$, where $p = \mathbb{P}[h(x) = 1]$. Let $\alpha := \frac{\sqrt{1 + H(p)(H(p) - \epsilon - 2\sqrt{\epsilon_0})} - 1}{H(p)}$ and assume by way of contradiction that there is a function $s : \mathbb{R}^d \to \mathcal{Y}$, that satisfies: $\mathbb{E}_x[\ell(s(\bar{g}(x)), h(x))] < \alpha$. Then, by Lem. 2 in the Appendix, we have:

$$\begin{aligned} I(\bar{g}(x); h(x)) &> (1 - \alpha)H(p) - H(\alpha) \\ &\geq (1 - \alpha)H(p) - 2\sqrt{\alpha} \end{aligned} \tag{21}$$

We conclude that:

$$(1 - \alpha)H(p) - 2\sqrt{\alpha} < \epsilon + 2\sqrt{\epsilon_0} \tag{22}$$

finally, by the quadratic formula, we arrive at a contradiction for $\alpha = \frac{\sqrt{1 + H(p)(H(p) - \epsilon - 2\sqrt{\epsilon_0})} - 1}{H(p)}$. Therefore, we conclude that, $g(x, h(x))$ is ϵ-complete. □

EF Operators

We next study the arithmetic of explanations. The practical utility of this is left for future research. However, we can imagine that by combining elementary explanations to complex ones and by intersecting these complex explanations, one can algorithmically construct explanations.

Definition 6 (Intersection and Union of Random Variables)**.** *Let $x \sim D$ and $f_1 : \mathcal{X} \to \mathcal{X}_1$ and $f_2 : \mathcal{X} \to \mathcal{X}_2$ are two functions. We say that the random variables $f_1(x)$ and $f_2(x)$ ϵ-intersect, if there are two invertible functions $r_1 : \mathcal{X}_1 \to \mathcal{V}_1$ and $r_2 : \mathcal{X}_2 \to \mathcal{V}_2$, such that, $r_1(f_1(x)) = (e_1(x), u(x))$ and $r_2(f_2(x)) = (e_2(x), u(x))$, where $I(e_i(x); f_j(x)) \leq \epsilon$ (for any $i \neq j \in \{1, 2\}$). We call the random variable $u(x)$, the ϵ-intersection of $f_1(x)$ and $f_2(x)$. In addition, we call $(e_1(x), u(x), e_2(x))$ the ϵ-union of $f_1(x)$ and $f_2(x)$.*

By Lem. 6 in the appendix, the intersection and union of two random variables $f_1(x)$ and $f_2(x)$ are unique, up to invertible transformations.

The following results show that the intersection of two EFs, one of which is valid and the other complete, is a valid EF.

Theorem 5. *Let $h : \mathbb{R}^n \rightarrow \mathcal{Y}$ be a model, $g_1, g_2 : \mathcal{Z} \rightarrow G$ two EFs and $\epsilon, \epsilon_0, \alpha > 0$ three constants. Assume that $\mathcal{Y} = \{\pm 1\}$, $g_1(x, h(x))$ and $g_2(x, h(x))$ ϵ-intersect and denote by $u(x, h(x))$ the ϵ-intersection of them. Assume that g_1 is ϵ_0-valid (w.r.t h) and g_2 is (ϵ, α)-complete (w.r.t h). Then, u is ϵ_1-valid (w.r.t h), for $\epsilon_1 := 1 - \frac{2^{-\epsilon_0 - 2\sqrt{\epsilon_0} - H(h(x))}}{1-\alpha}$.*

Proof. Let $r_1 : G \rightarrow \mathcal{V}_1$ and $r_2 : G \rightarrow \mathcal{V}_2$ be two invertible functions, such that, $r_1(g_1(x, h(x))) = (e_1(x, h(x)), u(x, h(x)))$ and $r_2(g_2(x, h(x))) = (e_2(x, h(x)), u(x, h(x)))$, where, $e_i(x, h(x)) \perp\!\!\!\perp g_j(x, h(x))$ (for any $i \neq j \in \{1, 2\}$). By the chain rule property of mutual information,

$$\begin{aligned} &I(e_1(x, h(x)), u(x, h(x)); h(x)) \\ =&I(e_1(x, h(x)); h(x)) \\ &+ I(u(x, h(x)); h(x) | e_1(x, h(x))) \\ \leq&I(e_1(x, h(x)); h(x)) + I(u(x, h(x)); h(x)) \end{aligned} \tag{23}$$

Therefore, we have:

$$\begin{aligned} &I(u(x, h(x)); h(x)) \\ \geq&I(e_1(x, h(x)), u(x, h(x)); h(x)) \\ &- I(e_1(x, h(x)); h(x)) \\ =&I(r_1(e_1(x, h(x)), u(x, h(x))); h(x)) \\ &- I(e_1(x, h(x)); h(x)) \\ =&I(g_1(x, h(x)); h(x)) - I(e_1(x, h(x)); h(x)) \end{aligned} \tag{24}$$

Since g_1 is ϵ_0-valid, there is a function, $t_1 : G \rightarrow \mathcal{Y}$, such that:

$$\begin{aligned} &\mathbb{P}_{x\sim D}[t_1(g_1(x, h(x))) \neq h(x)] \\ =&\mathbb{E}_{x\sim D}[\ell(t_1(g_1(x, h(x))), h(x))] \leq \epsilon_0 < 1/2 \end{aligned} \tag{25}$$

Therefore, by Lem. 2 and Lem. 4 in the Appendix, we have:

$$\begin{aligned} &I(g_1(x, h(x)); h(x)) \\ \geq&(1 - \epsilon_0)H(h(x)) - H(1 - \epsilon_0) \\ \geq&(1 - \epsilon_0)H(h(x)) - 2\sqrt{\epsilon_0} \end{aligned} \tag{26}$$

By the definition of $e_1(x, h(x))$, we have:

$$I(e_1(x, h(x)); g_2(x, h(x))) \leq \epsilon \tag{27}$$

Therefore, since g_2 is (ϵ, α)-complete, for every function s with outputs in $\mathcal{Y}$, we have:

$$\mathbb{E}_{x\sim D}[\ell(s(e_1(x, h(x))), h(x))] \geq \alpha \tag{28}$$

Therefore, by Lem. 5 in the Appendix,

$$I(e_1(x, h(x)); h(x)) \leq \log(1 - \alpha) + H(h(x)) \tag{29}$$

We conclude that:

$$\begin{aligned} &I(u(x, h(x)); h(x)) \\ \geq&(1 - \epsilon_0)H(h(x)) - 2\sqrt{\epsilon_0} \\ &- (\log(1 - \alpha) + H(h(x))) \\ \geq&\log\left(\frac{1}{1 - \alpha}\right) - \epsilon_0 \cdot H(h(x)) - 2\sqrt{\epsilon_0} \\ \geq&\log\left(\frac{1}{1 - \alpha}\right) - \epsilon_0 - 2\sqrt{\epsilon_0} \end{aligned} \tag{30}$$

Finally, by Lem. 5 in the Appendix, there is a function t_2 with outputs in $\mathcal{Y}$, such that:

$$\begin{aligned} &\mathbb{E}_{x\sim D}[\ell(t_2(u(x, h(x))), h(x))] \\ \leq&1 - \frac{2^{-\epsilon_0 - 2\sqrt{\epsilon_0} - H(h(x))}}{1 - \alpha} \end{aligned} \tag{31}$$

□

Similar results hold for the union of two EFs: if at least one of which is valid, the union is a valid EF, and a similar result for at least one complete EF.

Lemma 1. *Let $h : \mathbb{R}^n \rightarrow \mathcal{Y}$ be a model, $g_1, g_2 : \mathcal{Z} \rightarrow G$ two EFs and $\epsilon, \epsilon_0, \alpha > 0$ three constants. Assume that $\mathcal{Y} = \{\pm 1\}$, $g_1(x, h(x))$ and $g_2(x, h(x))$ ϵ-intersect and denote by $\hat{g}(x, h(x))$ the ϵ-union of them. If g_1 (or g_2) is ϵ_0-valid (w.r.t h), then, $\hat{g}$ is ϵ_0-valid as well. Additionally, if g_1 (or g_2) is (ϵ_1, α)-complete (w.r.t h), $\hat{g}$ is also (ϵ_1, α)-complete.*

Proof. First, by the definition of ϵ-union, there is a representation, $\hat{g}(x, h(x)) = (e_1(x, h(x)), u(x, h(x)), e_2(x, h(x)))$, such that, there is an invertible function r, that satisfies: $r(e_1(x, h(x)), u(x, h(x))) = g_1(x, h(x))$.

We would like to prove that if g_1 is ϵ_0-valid, then, $\hat{g}$ is also ϵ_0-valid. Since, g_1 is ϵ_0-valid, there is a function $t : G \rightarrow \mathcal{Y}$, such that:

$$\mathbb{E}_{x\sim D}[\ell(t(g_1(x, h(x))), h(x))] \leq \epsilon_0 \tag{32}$$

In addition, by the definition of $\hat{g}$, we have a representation: $\hat{g}(x, h(x)) = (e_1(x, h(x)), u(x, h(x)), e_2(x, h(x)))$, such that, there is an invertible function r, that satisfies: $r(e_1(x, h(x)), u(x, h(x))) = g_1(x, h(x))$. Therefore, we define, $r'(\hat{g}(x, h(x))) = g_1(x, h(x))$ and obtain,

$$\mathbb{E}_{x\sim D}[\ell(t \circ r'(\hat{g}(x, h(x))), h(x))] \leq \epsilon_0 \tag{33}$$

Hence, $\hat{g}$ is also ϵ_0-valid.

Next, we prove that if g_1 is (ϵ_1, α)-complete, then, $\hat{g}$ is also (ϵ_1, α)-complete. Let $\bar{g}(x)$ be a function that satisfies: $I(\bar{g}(x); \hat{g}(x, h(x))) \leq \epsilon_1$. In particular, there is a representation

$$\begin{aligned} &I(\hat{g}(x); \hat{g}(x, h(x))) \\ =&I(\hat{g}(x); e_1(x, h(x)), u(x, h(x)), e_2(x, h(x))) \\ \geq&I(\hat{g}(x); e_1(x, h(x)), u(x, h(x))) \\ =&I(\hat{g}(x); r(e_1(x, h(x)), u(x, h(x)))) \\ =&I(\hat{g}(x); g_1(x, h(x))) \end{aligned} \tag{34}$$

Therefore, $I(\hat{g}(x); g_1(x, h(x))) \leq \epsilon_1$. Since, g_1 is (ϵ_1, α)-complete, for any function s, we have:

$$\mathbb{E}_{x\sim D}[\ell(s(\bar{g}(x)), h(x))] \geq \alpha \tag{35}$$

In particular, we conclude that $\hat{g}$ is also (ϵ_1, α)-complete. □

Discussion

In this work, we have studied the properties of EFs g. We do not propose new ways to obtain such g, which is an active research topic with an increasing interest. Our focus is on blending functions, which mix the input and the output. We view this is a basic property of a wide class of existing and future types of explanations.

The challenge in formalizing EFs using conventional machine learning tools, is that these are not learned from data (they are designed by the practitioners). Therefore, one cannot use the usual convergence-based results. The claims that can be made are based on the mutual information between the model and the EF, the structure of the EF as a two-input function, and the validity requirement, which entails a specific recursive formula $h \approx t(g(x, h(x)))$.

There are three levels of abstractions, which are often referred to as explanations. One is the concrete explanation itself, which for us is an object in domain G, which is the target domain of g. The second one is the function that generates such explanations. We call these EFs. The third level is the algorithm that provides the EF g given a model h. Our analysis focuses on the EF level and it is important to note that g is not general to all h, but is given and analyzed in the context of a specific h.

Related Work

The examples that we have provided on available work on explainable solutions, are a fraction of the growing literature on the subject. See (Guidotti et al. 2018) for a survey. Our work covers what is referred to in this survey as the *outcome explanation problem*. It is interesting to contrast the definition of this term, given as Def 4.2 in that survey, to our terminology. Their definition assumes that the explanation is viewed through the lens of a local model c_l, which is constructed by some process f from the black box model (h in our terminology b in theirs) at a specific location x. The explanation itself $E(c_l, x)$ maps this local model and the input x to a human interpretable domain.

The example given is of a decision tree, with decision rules that are based on single attribute values (coordinates of x), that approximated the black box model in a given neighborhood of x. The explanation is given by the sequence of decisions along the path in this decision tree taken for sample x. The well known LIME approach (Ribeiro, Singh, and Guestrin 2016) also fits this definition well. In this approach, random samples are created in the vicinity of x, by perturbing this sample, and are weighed by their distance from x, when learning the local model c_l.

Our framework does not discuss the process $f(h, x)$. The two frameworks are compatible in the sense that $g(x, h(x))$ can be written as $E(f(h, x), x)$, since our g is a function of h (recall that g is specific for a given h), and since g could be a function that is based on local approximations of h. However, our framework emphasizes the blending properties of the explanation domain, while their definitions emphasize locality and local proxies of h by simple functions that are easy to explain, such as decision trees or linear functions.

The notion of locality is deferred in our model to the notions of consistency and explainability. However, these exist between intermediate representations and the EF, and Lipschitz continuity type properties and does not necessarily imply an actual approximation.

Recently, (Alvarez-Melis and Jaakkola 2018) have suggested a framework to learn models that are explainable by design. The basic structure is of a model that, similar to linear functions, is monotonic and additive in each of a set of learned attributes, and on learning attributes that are meaningful. The explanation itself takes the form of presenting the contribution of each attribute, while explaining the attributes using prototypes. While our framework focuses on explaining general models h and not learning self-explainable models, it is interesting to compare their stated desiderata with ours.

The specified desiderata on that work are:

1. Fidelity: the explanation of x should present the relevant information. This is captured by our validity property (relevancy to the label), as well as by the completeness property.

2. Diversity: the attributes should be disentangled and there should not be too many of them. This is a property on the explanation domain G, which in their work is also used for the representation of the network's penultimate layer. We consider a broader class of explanations, and our analysis of representations refers to f that can be any layer of the network h.

3. Grounding: the attributes of the explanations should be immediately interpretable to humans. In their model, the interpretation is done through prototype samples. A prototype based G is compatible with our framework. However, we cannot formalize the notion of interpretability.

Conclusions

The basic concepts of explanations in AI are elusive for several reasons. First, as mentioned, they need to be interpretable by humans, and human understanding has not been fully modeled. Second, there are multiple approaches in the literature. Third, tacit knowledge, by definition, cannot be fully laid down as a set of rules.

We build a formal framework for explainable AI, by considering, as a first principle, that outcome explanations blend the input and the prediction. Then, we link representations, which we typically take as intermediate activations of neural network models, to these explanations. The interrelationships between the explanations, the models, and the representations are potent enough to lead to several theoretical results.

One result is that desirable links between explanations and layers of a neural network cannot be specific to this layer, but also manifest to other layers. Another is that a valid explanation must also be complete. A third result studies explainability in the context of a concrete explanation of the predictions of multiclass neural networks. Lastly, we show results on the union and intersection of explanations.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant ERC CoG 725974). The contribution of Tomer Galanti is part of Ph.D. thesis research conducted at Tel Aviv University.

References

[Alvarez-Melis and Jaakkola 2018] Alvarez-Melis, D., and Jaakkola, T. S. 2018. Towards robust interpretability with self-explaining neural networks. In *NIPS*.

[Cover and Thomas 2006] Cover, T. M., and Thomas, J. A. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience.

[D'Aurizio 2015] D'Aurizio, J. 2015. An upper bound of binary entropy. `https://math.stackexchange.com/users/44121/jack-daurizio`.

[Feder and Merhav 1994] Feder, M., and Merhav, N. 1994. Relations between entropy and error probability. *IEEE Trans. Information Theory* 40:259–266.

[Guidotti et al. 2018] Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51(5):93:1–93:42.

[Hendricks et al. 2016] Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *European Conference on Computer Vision*, 3–19. Springer.

[Kozachinski 2018] Kozachinski, S. 2018. An upper bound on the difference between two similar mutual informations. `https://math.stackexchange.com/questions/2964570/a-bound-on-ixy-in-terms-of-ixz-for-y-and-z-that-are-similar/2964661`.

[Regev 2013] Regev, O. 2013. Entropy-based bounds on dimension reduction in l^1. *Israeli Journal of Mathematics*.

[Ribeiro, Singh, and Guestrin 2016] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

[Zeiler and Fergus 2014] Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833.

Useful Lemmas

For completeness, we provide some useful lemmas that are being employed in the proofs on the theorems in our paper.

Lemma 2. *Let X and Y be two random variables. Assume that there is a function F, such that $\mathbb{P}[F(Y) = X] \geq q \geq 1/2$. Then, $I(X;Y) \geq qH(X) - H(q)$.*

Proof. The lemma is a modification of Claim 2.1 in (Regev 2013).

Lemma 3. *Let X, Y and Z be three random variables, where Y and Z are binary. We have:*

$$|I(X;Y) - I(X;Z)| \leq H(\mathbb{P}[Y \neq Z]) \tag{36}$$

Proof. See. (Kozachinski 2018).

Lemma 4. *Let $p \in [0,1]$. Then,*

$$H(p) \leq 2\sqrt{p(1-p)} \tag{37}$$

Proof. See (D'Aurizio 2015).

Lemma 5. *Let X and Y be two discrete random variables taking values from $\mathcal{S}_1$ and $\mathcal{S}_2$ (resp.). Then, there is a function $t : \mathcal{S}_2 \to \mathcal{S}_2$, such that:*

$$\mathbb{P}_{X,Y}[X = t(Y)] \leq 1 - 2^{I(X;Y) - H(X)} \tag{38}$$

Proof. See (Feder and Merhav 1994).

Lemma 6 (Intersection Equivalence)**.** *Let $x \sim D$ and $f_1 : X \to X_1$ and $f_2 : X \to X_2$ are two functions. In addition, let $u_1(x)$ and $u_2(x)$ be two ϵ-intersections of $f_1(x)$ and $f_2(x)$, i.e., there are two pairs of invertible functions $r^i_1 : X_1 \to \mathcal{V}_1$ and $r^i_2 : X_2 \to \mathcal{V}_2$, such that, $r^i_1(f_1(x)) = (e^i_1(x), u_i(x))$ and $r^i_2(f_2(x)) = (e^i_2(x), u_i(x))$, where, $I(e^i_j(x); f_k(x)) \leq \epsilon$ (for any $i \in \{1,2\}$ and $k \neq j \in \{1,2\}$). Then, there are functions s_1, s_2 and d_1, d_2, such that, for all $i \neq j \in \{1,2\}$, we have:*

$$\mathbb{E}_{x\sim D}[\ell(s_i(u_i(x)), u_j(x))] \leq 1 - 2^{-\epsilon} \tag{39}$$

and also,

$$\mathbb{E}_{x\sim D}[\ell(d_i(e^i_1(x)), e^j_1(x))] \leq 1 - 2^{-\epsilon} \tag{40}$$

In particular, if $\epsilon = 0$, $s_1(u_1(x)) = u_2(x)$, s_1 is invertible, such that $s_1^{-1} = s_2$ and $d_1(e^1_1(x)) = e^2_1(x)$, d_1 is invertible and $d_1^{-1} = d_2$.

Proof. First, we would like to show that $I(e^1_1(x); u_2(x)) \leq \epsilon$. Assume by contradiction that this is not the case. We consider that, $u_2(x)$ can be represented as a function of $f_2(x)$, since $u_2(x)$ consists of the last coordinate of $r^2_2(f_2(x))$. Therefore, since r^2_2 is invertible,

$$\begin{aligned} I(e^1_1(x); f_2(x)) &= I(e^1_1(x); r^2_2(f_2(x))) \\ &\geq I(e^1_1(x); u_2(x)) > \epsilon \end{aligned} \tag{41}$$

In contradiction to the assumption that $I(e^1_1(x); f_2(x)) \leq \epsilon$. By the same argument, we also have, $I(e^1_1(x); u_1(x)) \leq \epsilon$. By the chain rule property of mutual information,

$$\begin{aligned} I(e^1_1(x), u_1(x); u_2(x)) =& I(u_1(x); u_2(x)) \\ &+ I(e^1_1(x); u_2(x) | u_1(x)) \\ \leq& I(u_1(x); u_2(x)) \\ &+ I(e^1_1(x); u_2(x)) \end{aligned} \tag{42}$$

Therefore, since r^1_1 and r^2_1 are invertible functions,

$$\begin{aligned} I(u_1(x); u_2(x)) \geq& I(e^1_1(x), u_1(x); u_2(x)) \\ &- I(e^1_1(x); u_2(x)) \\ \geq& I(e^1_1(x), u_1(x); u_2(x)) - \epsilon \\ =& I(r^1_1(e^1_1(x), u_1(x)); u_2(x)) - \epsilon \\ =& I(f_1(x); u_2(x)) - \epsilon \\ =& I(r^2_1(f_1(x)); u_2(x)) - \epsilon \\ =& I(e^2_1(x), u_2(x); u_2(x)) - \epsilon \\ \geq& I(u_2(x); u_2(x)) - \epsilon \\ =& H(u_2(x)) - \epsilon \end{aligned} \tag{43}$$

Again, by the chain rule property of mutual information,

$$\begin{aligned} I(e^1_1(x), u_1(x); e^2_1(x)) =& I(e^1_1(x); e^2_1(x)) \\ &+ I(e^2_1(x); u_1(x)|e^1_1(x)) \\ \leq& I(e^2_1(x); e^1_1(x)) \\ &+ I(e^1_1(x); u_1(x)) \end{aligned} \tag{44}$$

Therefore, since r^1_1 and r^2_1 are invertible functions,

$$\begin{aligned} I(e^1_1(x); e^2_1(x)) \geq& I(e^1_1(x), u_1(x); e^2_1(x)) \\ &- I(e^1_1(x); u_1(x)) \\ \geq& I(e^1_1(x), u_1(x); e^2_1(x)) - \epsilon \\ =& I(r^1_1(e^1_1(x), u_1(x)); e^2_1(x)) - \epsilon \\ =& I(f_1(x); e^2_1(x)) - \epsilon \\ =& I(r^2_1(f_1(x)); e^2_1(x)) - \epsilon \\ =& I(e^2_1(x), u_2(x); e^2_1(x)) - \epsilon \\ \geq& I(e^2_1(x); e^2_1(x)) - \epsilon \\ =& H(e^2_1(x)) - \epsilon \end{aligned} \tag{45}$$

Thus, we conclude that $I(u_1(x); u_2(x)) \geq H(u_2(x)) - \epsilon$ and that $I(e^1_1(x); e^2_1(x)) \geq H(e^2_1(x)) - \epsilon$. In a similar manner, we can show the other directions as well, $I(u_1(x); u_2(x)) \geq H(u_1(x)) - \epsilon$ and $I(e^1_1(x); e^2_1(x)) \geq H(e^1_1(x)) - \epsilon$. Therefore, by Lem. 5 in the Appendix, there are functions s_1, s_2 and d_1, d_2, such that, for all $i \neq j \in \{1, 2\}$, we have:

$$\begin{aligned} &\mathbb{E}_{x\sim D}[\ell(s_i(u_i(x)), u_j(x))] \\ =& \mathbb{P}_{u_1(x),u_2(x)}[s_i(u_i(x)) = u_j(x)] \\ \leq& 1 - 2^{I(u_1(x);u_2(x)) - H(u_j(x))} \\ \leq& 1 - 2^{H(u_j(x)) - \epsilon - H(u_j(x))} = 1 - 2^{-\epsilon} \end{aligned} \tag{46}$$

and also,

$$\mathbb{E}_{x\sim D}[\ell(d_i(e^i_1(x)), e^j_1(x))] \leq 1 - 2^{-\epsilon} \tag{47}$$

Finally, if $\epsilon = 0$, for every $x \in \mathcal{X}$, we have: $s_1(u_1(x)) = u_2(x)$, $s_2(u_2(x)) = u_1(x)$, $d_1(e^1_1(x)) = e^1_1(x)$ and $d_2(e^2_1(x)) = e^2_1(x)$. Therefore, $s^{-1}_1 = s_2$ and $d^{-1}_1 = d_2$. □