Self-learn to Explain Siamese Networks Robustly

Chao Chen*, Yifan Shen†, Guixiang Ma‡, Xiangnan Kong§, Srinivas Rangarajan¶, Xi Zhang†, Sihong Xie*
*Computer Science and Engineering Dept, Lehigh University †Laboratory of Trustworthy Distributed Computing and Service (MoE), BUPT,
‡University of Illinois at Chicago, §Worcester Polytechnic Institute, ¶Department of Chemical and Biomolecular Engineering, Lehigh University
chc517@lehigh.edu, shenyifan@bupt.edu.cn, guixiang.ma@intel.com, xkong@wpi.edu, srr516@lehigh.edu, zhangx@bupt.edu.cn, xiesihong1@gmail.com

***Abstract*—Learning to compare two objects are essential in applications, especially when labeled data are scarce and imbalanced. As these applications can involve humans and make high-stake decisions, it is critical to explain the learned models. We aim to study post-hoc explanations of Siamese networks (SN) widely used in learning to compare. We characterize the instability of gradient-based explanations due to the additional compared object in SN, in contrast to architectures with a single input instance. We optimize for global invariance based on unlabeled data using self-learning to promote the stability of local explanations for individual input. The invariance leads to constrained optimization problems that can be solved using gradient descent-ascent (GDA), or KL-divergence regularized unconstrained optimization solved by SGD. We provide convergence proofs when the objective functions are nonconvex due to the Siamese architecture. Results on tabular and graph data from neuroscience and chemical engineering show that our local explanations robustly respects the self-learned invariance while optimizing the explanation faithfulness and simplicity. We further demonstrate the convergence of GDA experimentally.**

I. Introduction

Siamese networks (SN for short in the sequel) are widely used in similarity metric learning [16] and contrastive learning [4] where objects are compared. Different from conventional architectures that take one input instance, an SN maps a pair of instances (the "query" and the "reference") to a similarity score [17]. As SN is widely used in high-stake applications involving individual humans and societal values, it is urgent to provide simple and convincing explanations [9].

Challenges. We focus on post-hoc explanations consisting of a few salient input elements that can closely approximate the prediction made on the original input [19]. In contrast to explaining architectures with one single input [19], [26], explanations for SN should consider both query and reference, and explanations insensitive to either of the inputs [17] are misleading. However, the additional reference can make the explanation over-sensitive to irrelevant perturbations.

A stable explanation should retain the robust features of one input regardless of the other. One example is when SN is used for graph comparison [16]. Neuroscience studies have shown that the global Default Mode Network (DMN) [18] consisting of several brain regions of interest (ROI) involves multiple cognitive and affective functions. Researchers can be interested in an invariant subgraph of the ROIs when using SN to extract the difference between a bipolar patient and various healthy reference controls [28].

Proposed Method. To control superfluous variations due to the compared object, we find global invariant salient features for individual objects using self-supervision. We then formulate optimization problems to adapt the global salient features to explain an SN prediction local to an input pair. The adaptation balances the conformity to the invariance and the local flexibility when comparing a query to different references ("global" means "regardless of the references compared with a query", rather than the universal behaviors of the explained model over all data [15]). Under this framework, we design a gradient descent ascent (GDA) algorithm to solve a constrained optimization problem SNX (SN Explainer), and an unconstrained optimization problem with KL-divergence regularization (SNX-KL) solved by SGD.

The optimization problems can incorporate additional constraints. One-hot encoding is widely used on tabular datasets with categorical attributes [11]. For example, in Fig. 1, three binary features ("minor" features) represent the three values (*young*, *midlife*, and *old*) of the attribute `Age` ("major" features). One-hot encoding leads to the constraints that a major categorical attribute is salient if and only if at least one of the associated minor binary features is salient. Local explanations on the same query but different reference instances can select different binary minor features under the same major features that globally characterize the query. Prior methods explaining SN [22] on images do no have such constraints.

Regarding graph data, explaining the predicted similarity by subgraphs enumeration is NP-hard. Recent graph explanation approaches treat the edges independently, possibly leading to less coherent subgraphs that are not interpretable, as larger connected subgraphs can have biological or chemical significance [26]. We introduce structural constraints to make adjacent edges more likely to be selected into a subgraph as global invariant characterization of each graph.

Contributions. 1) We formulate the explanations of SN as two optimization problems and design algorithms with convergence guarantee theoretically. 2) We demonstrate that self-supervised learning can find meaningful invariant to find more stable local explanations, and our methods outperform the state-of-the-art on six datasets in terms of faithfulness, counterfactual, and conformity. 3) The methods work on both tabular and graph data with convincing case studies.

II. Problem Definition

A. Data with structures

Tabular data is a set of vectors, each with the same list of q categorical *major* features $\mathbf{z} = [z_1, \ldots, z_q]$, such as age and deposits of a credit card applicant [11]. Using one-hot encoding, each major feature z_i is transformed to a set of binary

‡This author now works at Intel Labs.

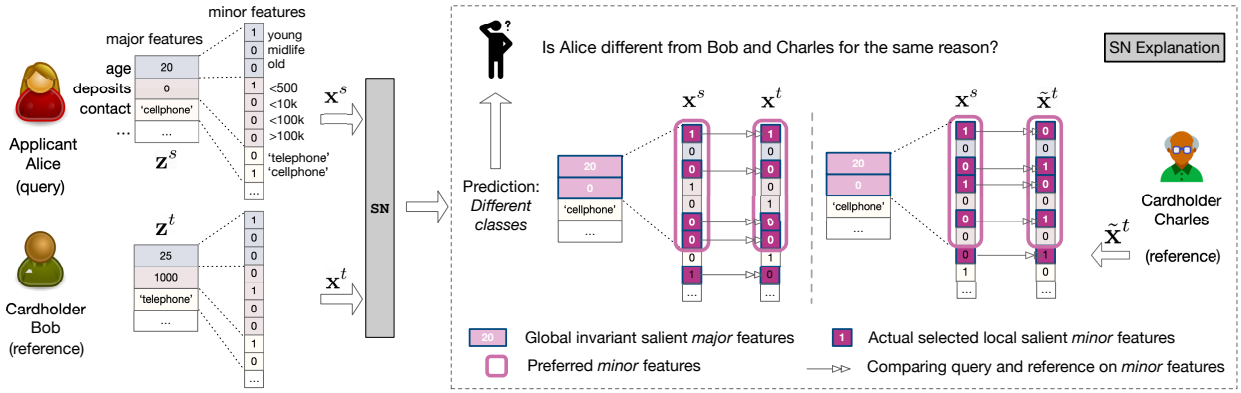

Fig. 1: Explaining Siamese networks with invariant. *Left:* major feature vectors $\mathbf{z}$ can be converted to minor feature vectors $\mathbf{x}$ by one-hot encoding, and the SN predicts the similarity of the query $\mathbf{x}^s$ and reference $\mathbf{x}^t$. *Right:* SN explanation (SNX) for tabular data is enclosed in the dash-lined box. Global invariant salient features for Alice are {`Age, Deposits`}, on both the major and minor levels (light purple values and boxes). The comparisons of Alice with references Bob and Charles lead to different local explanations (dark purple boxes), with two minor features selected beyond global invariant features.

minor features $x_{i,j} \in \{0,1\}$, where $x_{i,j} = 1$ if z_i takes the j-th possible value. As a result, $\sum_j x_{i,j} = 1, \forall i = 1, \ldots, q$. The *minor* feature vector $\mathbf{x} \in \{0,1\}^p$ is the concatenation of all binary minor features, and p is the number of minor features.

Graph data is a set of graphs, and each graph $G = (V, E)$ contains a set of vertices V and edges $E \subset V \times V$. We assume G is undirected and its adjacency matrix A is symmetric. To unify the descriptions of optimization problems, A is flattened to a vector $\mathbf{x}$ of length $(|V|-1)(|V|-2)/2$ due to symmetry.

B. Siamese Networks

An SN accepts a pair of instances, denoted as query $\mathbf{x}^s \in \mathbb{R}^{p_s}$ and reference $\mathbf{x}^t \in \mathbb{R}^{p_t}$ [17], with $p_s = p_t$ for two vectors, and $p_s \neq p_t$ in general for two different graphs. The superscript s or t will be omitted when referring to a single instance in general. The SN consists of a mapping function $emb(\mathbf{x}; \boldsymbol{\theta})$ that maps $\mathbf{x}^s$ and $\mathbf{x}^t$ to a latent space, where a metric measures the similarity between the two embeddings. The SN is then the composite function $f(\mathbf{x}^s, \mathbf{x}^t; \boldsymbol{\theta}) = sim(emb(\mathbf{x}^s; \boldsymbol{\theta}), emb(\mathbf{x}^t; \boldsymbol{\theta}))$. f is trained to maximize the similarity between any two instances of the same class using some loss function ℓ^{SN} as follows,

$$\min_{\boldsymbol{\theta}} \sum_{(s,t)\in\mathcal{T}} \ell^{\text{SN}}(f(\mathbf{x}^s, \mathbf{x}^t; \boldsymbol{\theta}), y_{st}),$$

where $\mathcal{T}$ is the training set containing all query-reference pairs. The label of a pair $(\mathbf{x}^s, \mathbf{x}^t)$ is $y_{st} = \mathbb{1}[y^s = y^t]$.

C. Post-hoc explanation of SN

We assume a trained SN $f(\mathbf{x}^s, \mathbf{x}^t; \boldsymbol{\theta})$ and focus on explaining the SN's predictions on test data. The parameter $\boldsymbol{\theta}$ is fixed and thus omitted from $f(\mathbf{x}^s, \mathbf{x}^t; \boldsymbol{\theta})$ when there is no confusion. Given a pair of query $\mathbf{x}^s \in \{0,1\}^{p_s}$ and reference $\mathbf{x}^t \in \{0,1\}^{p_t}$, let $\mathbf{m}^s \in [0,1]^{p_s}$ and $\mathbf{m}^t \in [0,1]^{p_t}$ be the corresponding multiplicative masks. A large element in a mask indicates that the corresponding feature value contributes more to the SN prediction [26], [20]. The element-wise product $\mathbf{m} \otimes \mathbf{x}$ is a masked instance so that $\mathbf{m}_i \mathbf{x}_i \in [0,1]$ is the importance/saliency of the i-th element of $\mathbf{x}$. Additive perturbations [19], [14] are less interpretable, as the perturbed binary features can be outside of $[0,1]$. A prediction $f(\mathbf{x}^s, \mathbf{x}^t)$ depends on both inputs $(\mathbf{x}^s, \mathbf{x}^t)$, so does its explanation, leading to robustness issues of the gradient-based explanations.

Robustness of SN explanations. Using a simple example SN $f(\mathbf{x}^s, \mathbf{x}^t; \boldsymbol{\theta}) = \sigma(< \boldsymbol{\theta}^\top \mathbf{x}^s, \boldsymbol{\theta}^\top \mathbf{x}^t >)$, we characterize the robustness of gradient-based explanations of SN. Taking the gradient of ℓ^{SN} with respect to the query $\mathbf{x}^s$, we obtain a saliency map over $\mathbf{x}^s$ proportional to $\boldsymbol{\theta}\boldsymbol{\theta}^\top \mathbf{x}^t$. The salient map explains the prediction using the magnitudes of elements in $\boldsymbol{\theta}\boldsymbol{\theta}^\top \mathbf{x}^t$ and depends on the SN parameter $\boldsymbol{\theta}$ and the reference $\mathbf{x}^t$. The saliency map can be manipulated to any pre-defined target explanation $\tilde{\mathbf{m}}^s$, by perturbing the reference $\mathbf{x}^t$:

$$\min_{\boldsymbol{\delta}\in\mathbf{R}^{p_t}} \|(\boldsymbol{\theta}\boldsymbol{\theta}^\top)(\mathbf{x}^t + \boldsymbol{\delta}) - \tilde{\mathbf{m}}^s\|_2^2 \quad \text{s.t.} < \mathbf{x}^s, (\boldsymbol{\theta}\boldsymbol{\theta}^\top)(\boldsymbol{\delta}) >= 0.$$

The objective pushes the saliency map to the target mask $\tilde{\mathbf{m}}^s$ [5], [8], while the equality constraint specifies that the SN prediction is not changed and any vector $\boldsymbol{\delta}$ orthogonal to $\mathbf{x}^s$ will work. Since the one-hot encoding or a sparse graph can result in a large number of zeros in $\mathbf{x}^s$, there are many such orthogonal vectors.

Desiderata. We aim to find local explanations (i.e., masks, denoted by $\mathbf{m}$ in general), with the following merits:

- *Simplicity* [19] of an explanation is measured by the number of important features or edges according to $\mathbf{m}$, such as $\|\mathbf{m}\|$, the ℓ_1 norm of $\mathbf{m}$.
- *Faithfulness* [19], [14] can be evaluated by feeding the masked instances $(\mathbf{m}^s \otimes \mathbf{x}^s, \mathbf{m}^t \otimes \mathbf{x}^t)$ to the target SN f and measuring the distortion in the original output:

$$\ell(f(\mathbf{x}^s, \mathbf{x}^t), f(\mathbf{m}^s \otimes \mathbf{x}^s, \mathbf{m}^t \otimes \mathbf{x}^t)), \tag{1}$$

  where ℓ is some loss functions, such as the cross-entropy loss. A low faithfulness loss indicates that the masks can select salient features to preserve the SN output $f(\mathbf{x}^s, \mathbf{x}^t)$.
- *Counterfactuals* [23] is the complement $(1 - \mathbf{m})$ of $\mathbf{m}$ and can show "what" would the predictions be "if" keeping the non-salient features. We define the following *counterfactual loss* to measure how much $\mathbf{m}^s$ and $\mathbf{m}^t$ miss salient features:

$$\ell(f(\mathbf{x}^s, \mathbf{x}^t), f((1 - \mathbf{m}^s) \otimes \mathbf{x}^s, (1 - \mathbf{m}^t) \otimes \mathbf{x}^t)). \tag{2}$$

- *Conformity* measures how much a local explanation overlaps the global salient features. Conforming to an invariant leads to more robustness against variations in the reference $\mathbf{x}^t$.

III. Self-learn to explain robustly

A. A general optimization formulation

The variables to be optimized are the two masks $\mathbf{m}^s \in [0,1]^{p_s}$ and $\mathbf{m}^t \in [0,1]^{p_t}$ output by the sigmoid function over $\mathbf{x}^s$ and $\mathbf{x}^t$, respectively. To balance faithfulness and simplicity, we have:

$$\min_{\mathbf{m}^s,\mathbf{m}^t} \ell\left(f(\mathbf{x}^s,\mathbf{x}^t), f(\mathbf{m}^s\otimes\mathbf{x}^s, \mathbf{m}^t\otimes\mathbf{x}^t)\right) + \gamma(\|a(\mathbf{m}^s)\| + \|a(\mathbf{m}^t)\|), \tag{3}$$

where γ is a hyperparameter to fine-tune the relative importance of the two goals. $\ell(\cdot,\cdot)$ as defined by Eq. (1) promotes faithfulness and $\|\cdot\|$ is the ℓ_1-norm that promotes simplicity. Note that simplicity can be structural (such as joint sparsity [21]) and the auxiliary function $a(\mathbf{m})$ maps from an unstructured mask to another vector, upon which structural sparsity constraints can be imposed on Section III-B and III-C.

Stage 1. Saliency maps may lack robustness and we use a global invariant to regulate $\mathbf{m}^s$ for more robustness against varying $\mathbf{x}^t$. If there is domain knowledge regarding which features/edges in a query $\mathbf{x}$ are salient, we can set the binary values in the global mask $\mathbf{M}\in[0,1]^p$ for $\mathbf{x}$ accordingly. There is no such knowledge in a more general case, and we propose to extract $\mathbf{M}$ as an invariant to encode global salient elements in $\mathbf{x}$, regardless of references, using self-supervision learning (SSL). SSL [27], [10] train a predictive model h by contrasting $\mathbf{x}$ and its transformation $T(\mathbf{x})$, where $\mathbf{x}$ can be a graph [27], [10] or an image [4]. The objective function in SSL is:

$$h^* = \arg\min_h \mathcal{L}_{self}(T,\mathbf{x},h) = \arg\min_h \ell(h(\mathbf{x}), h(T(\mathbf{x}))),$$

where $T(\cdot)$ is a given transformation. For example, $T(\cdot)$ can be random walk masking out irrelevant parts of a graph [10], and $T(\mathbf{x}) = \mathbf{M}\otimes\mathbf{x}$. SSL uses a fixed T function to learn h, while we are interested in learning a T, which is a global mask for $\mathbf{x}$ regardless of different references compared:

$$\min_{\mathbf{M}} \quad \ell\left(f(\mathbf{x},\mathbf{x}), f(\mathbf{x},\mathbf{M}\otimes\mathbf{x})\right) + \gamma\|a(\mathbf{M})\|, \tag{4}$$

$$\text{s.t.} \quad g_i(\mathbf{M}) \le 0, i = 1,\ldots,c. \tag{5}$$

$\mathbf{M}$ extracts the salient features of $\mathbf{x}$ to maximally preserve information in $\mathbf{x}$, thus a low faithfulness loss (Eq. (1)).

Stage 2. After finding $\mathbf{M}^s$ and $\mathbf{M}^t$ for $\mathbf{x}^s$ and $\mathbf{x}^t$, respectively, using Eqs. (4)-(5), we fix the global masks $\mathbf{M}^s$ and $\mathbf{M}^t$ and incorporate them as constraints in the following optimization problem (termed "SNX") to find local masks $\mathbf{m}^s$ and $\mathbf{m}^t$:

$$\min_{\mathbf{m}^s,\mathbf{m}^t} \ell\left(f(\mathbf{x}^s,\mathbf{x}^t), f(\mathbf{m}^s\otimes\mathbf{x}^s,\mathbf{m}^t\otimes\mathbf{x}^t)\right) + \gamma\left(\|a(\mathbf{m}^s)\| + \|a(\mathbf{m}^t)\|\right) \tag{6}$$

$$\text{s.t.} \quad g_i(\mathbf{m}) = a(\mathbf{m}^s)_i - a(\mathbf{M}^s)_i \le 0,\ i = 1,\ldots,c_s, \tag{7}$$

$$g_{c_s+i}(\mathbf{m}) = a(\mathbf{m}^t)_i - a(\mathbf{M}^t)_i \le 0,\ i = 1,\ldots,c_t. \tag{8}$$

$a(\cdot)_i$ means the i-th element of $a(\cdot)$, and c_s and c_t are the numbers of the constraints.

B. Optimization problem for tabular data

Stage 1. Without a particular reference $\mathbf{x}^t$, a global mask over a query can at best identify salient major features, such as `Age`, in $\mathbf{z}^s$. We use the auxiliary function to find a global invariant mask $\mathbf{N}\in[0,1]^q$ over the *major* features, where $\mathbf{N}_i = a(\mathbf{M})_i = 1-\prod_j(1-\mathbf{M}_{i,j})$ is the importance of the i-th major feature[1], and $\mathbf{M}_{i,j}$ indicates the global importance of the j-th value of the i-th major categorical feature of the query $\mathbf{x}$. As we already encode the dependencies among minor features in $a(\mathbf{M})$, there is no more constraints in Eq. (5) ($c=0$).

Stage 2. Comparing with $\mathbf{x}^t$, we further identify salient minor features, such "`Age`<25", associated with the salient major features. For tabular data, any two input vectors to SN are aligned, so we optimize a single mask $\mathbf{m} = \mathbf{m}^s = \mathbf{m}^t$ to find salient features for both instances. We use the same auxiliary function for the local masks $\mathbf{n} = a(\mathbf{m})$ such that $\mathbf{n}_i = 1 - \prod_j(1-\mathbf{m}_{i,j})$ in Eq. (6). As we focus on finding masks for the query $\mathbf{x}^s$ with varying references $\mathbf{x}^t$, only Eq. (7) is kept ($c_s = q$, the number of major features). Alternatively, we formulate an unconstrained optimization problem (termed "SNX-KL"):

$$\min_{\mathbf{m}} \ell\left(f(\mathbf{x}^s,\mathbf{x}^t), f(\mathbf{m}\otimes\mathbf{x}^s,\mathbf{m}\otimes\mathbf{x}^t)\right) + \gamma\|\mathbf{n}\| + \beta\mathrm{KL}(\mathbf{n}\|\mathbf{N}), \tag{9}$$

where $\mathrm{KL}(\mathbf{n}\|\mathbf{N}) = \sum_{i=1}^q \mathrm{KL}(\mathbf{n}_i\|\mathbf{N}_i)$ and $\mathrm{KL}(\mathbf{n}_i\|\mathbf{N}_i)$ is the KL-divergence between $\mathbf{n}_i$ and $\mathbf{N}_i$, which encourages $\mathbf{n}_i$ to be smaller than $\mathbf{N}_i$ (according to Section 10.1 in [2]).

C. Optimization problems for graphs

We set $a(\mathbf{M}) = \mathbf{M}$ for masks on graphs. Isolated single-edged subgraphs are not only difficult for domain experts to interpret, but can also disturb the working of GNNs within SN. Thus, the selection of two adjacent edges should be related. We consider such dependencies in Eq. (5) in stage 1,

$$g_{jk}(\mathbf{M}) = \|\mathbf{M}_j - \mathbf{M}_k\| - \epsilon \le 0, j,k \text{ adjacent in } G, \tag{10}$$

where $\mathbf{M}_j\in[0,1]$ is the mask for the j-th edge. The constraint indicates that the selection of the j-th edge can lead to the selection of the k-th edge if they share a node [15], and ϵ controls the co-occurrence of the two edges. After obtaining $\mathbf{M}^s$ and $\mathbf{M}^t$ for each graph, the local masks $\mathbf{m}^s$ and $\mathbf{m}^t$ over $\mathbf{G}^s$ and $\mathbf{G}^t$ are optimized by solving problem Eqs. (6)-(8), using $\mathbf{M}^s$ and $\mathbf{M}^t$ as constants in the constraints. In general, $\mathbf{G}^s$ and $\mathbf{G}^t$ have different numbers of nodes, which are not aligned. Therefore, $\mathbf{m}^s$ and $\mathbf{m}^t$ can lead to different numbers of constraints ($c_s = p^s$ and $c_t = p^t$). Similar to Eq. (9), one can use KL-regularization terms to enforce the constraints.

D. Optimization algorithm and convergence

SNX-KL defined in Eq. (9) is unconstrained optimization and SGD can be used. Alternatively, global explanations on graph data (Eq. (4)-(5)) and local explanations on both tabular and graph data (Eq. (6)-(8)) are constrained optimization

[1]We tried alternatives, such as $\mathbf{N}_i = \sum_k \mathbf{M}_{i,k}$ and $\mathbf{N}_i = \prod_k \mathbf{M}_{i,k}$. They cannot focus on minor features for a significant major feature, or lead to numerical underflow issues.

Algorithm 1 SNX: Siamese Network Explanation with GDA

Input: a target SN model f, a query instance $\mathbf{x}^s$ and reference instance $\mathbf{x}^t$, *optional* human-defined constraints in $\mathbf{M}^s$ and $\mathbf{M}^t$, learning rate η_1, η_2 for $\mathbf{m}$ and $\boldsymbol{\lambda}$.
Output: local masks $\mathbf{m}^s$ for $\mathbf{x}^s$, and $\mathbf{m}^t$ for $\mathbf{x}^t$.
Init: $\boldsymbol{\lambda} = [1/c, \ldots, 1/c] \in \mathbb{R}^c$.
if $\mathbf{M}^s$ and $\mathbf{M}^t$ not given **then**
    Extract global masks $\mathbf{M}^s$ and $\mathbf{M}^t$ by Eqs. (4)-(5) ▷SSL
end if
Pretrain $\mathbf{m}^s, \mathbf{m}^t$ without constraints using Eq. (6).
Solve the full constrained optimization problems Eqs. (6)-(8) using GDA to find local masks $\mathbf{m}^s$ and $\mathbf{m}^t$.

problems. We adopt the gradient descent-ascent (GDA) algorithm [13] (Algorithm 1) to allow violations of the constraints. Take optimizing local masks as an example, the mask to optimize is primal variables $\mathbf{m} = \mathbf{m}^s = \mathbf{m}^t$ for tabular data and $\mathbf{m} = [\mathbf{m}^s; \mathbf{m}^t]$ for graph data. The objective function $g_0(\mathbf{m})$ is that defined in Eq. (4) or Eq. (6), and the inequality constraints $g_i(\mathbf{m}), \forall i \in \{1, \ldots, c\}$ are those defined in Eq. (5) or Eqs. (7)-(8), and c is the number of inequality constraints. We introduce the non-negative dual Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^c_+$ and construct the Lagrangian function

$$\mathcal{L}(\mathbf{m}, \boldsymbol{\lambda}) = g_0(\mathbf{m}) + \textstyle\sum_{i=1}^{c} \lambda_i g_i(\mathbf{m}), \tag{11}$$

Then gradient descent is applied to $\mathbf{m}$ and gradient ascent is applied to $\boldsymbol{\lambda}$ with learning rates η_1 and η_2:

$$\mathbf{m} \leftarrow \mathbf{m} - \eta_1 \tfrac{\partial \mathcal{L}}{\partial \mathbf{m}}, \qquad \boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \eta_2 \tfrac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}}. \tag{12}$$

Between the two updates, we use the latest $\mathbf{m}$ to evaluate the partial derivatives with respect to $\boldsymbol{\lambda}$. Also, the $\boldsymbol{\lambda}$ vector is normalized to a probability distribution before starting the next iteration. The time complexity of each optimization iteration is the sum of that of training the SN using back-propagation and that of evaluating the c constraints.

Convergence. The Lagrangian is nonconvex in the primal variables and linear in the dual variables. The GDA algorithm is convergent based on Theorem 4.4 of the work [13], given that the Lagrangian satisfies their Assumption 4.2:

1) *$\mathcal{L}$ is l-smooth and $\mathcal{L}(\cdot, \boldsymbol{\lambda})$ is L-Lipschitz for each $\boldsymbol{\lambda}$ and $\mathcal{L}(\mathbf{m}, \cdot)$ is concave for each $\mathbf{m}$.*
2) *The domain of $\boldsymbol{\lambda}$ is convex and bounded.*

We verify these assumptions in the full version [3].

IV. Experiments

A. Experimental settings

Datasets. We conduct experiments on four tabular datasets and two graph datasets. More details about experimental settings and more comprehensive results for *evaluation of global masks*, *conformity*, *convergence of GDA*, *sensitivity analysis*, and *case studies* are provided in the full version [3].

Metrics. We evaluate the faithfulness (FA, Eq. (1)) and counterfactual loss (CF, Eq. (2)). The conformity is also calculated by averaging the Jaccard similarity between global mask $\mathbf{M}$ and local mask $\mathbf{m}$ $J = \frac{|a(\mathbf{M}) \cap a(\mathbf{m})|}{|a(\mathbf{M}) \cup a(\mathbf{m})|}$ for both $\mathbf{x}^s$ and $\mathbf{x}^t$.

Baselines and variants. There are several options to obtain local masks $\mathbf{m}^s$ for the queries. The following three methods use global masks as local masks and are agnostic to references.

- Pick-all: set all elements in the global masks to one.
- DES (tabular data only) [25]: an unsupervised feature selection that generates pseudo labels for all pairs of instances using kNN to supervise the learning of a feature selector.
- SNX-global: solve Eq. (4) for $\mathbf{M}^s$ which is treated as a local mask without stage 2 local mask optimization.

The following baselines disregard constraints by global masks.

- Saliency maps (SM) [1]: take the gradient of ℓ^{SN} w.r.t. input $\mathbf{x}^s$ and retain features with the largest gradient magnitudes.
- SNX-unconstrained (SNX-UC): minimize the objective Eq. (6) without the constraints Eqs. (7)-(8).
- PGExplainer (PGExp) [15]: train a shareable generator for all graphs by maximizing the MI.
- GNNExplainer (GNNExp) [26]: learn soft masks for edges by maximizing the mutual information (MI).

The following three baselines extract local masks with variants of global masks. They apply to tabular data only.

- SNX-DES: optimize $\mathbf{m}$ by Eqs. (6)-(8), with global mask $\mathbf{M}^s$ found by DES used in the constraints Eqs. (7)-(8).
- SNX-inter / SNX-union: use the element-wise minimum $\mathbf{M} = \min(\mathbf{M}^s, \mathbf{M}^t)$ (or maximum $\mathbf{M} = \max(\mathbf{M}^s, \mathbf{M}^t)$) to simulate the intersection (or union) of $\mathbf{M}^s$ and $\mathbf{M}^t$.

B. Quantitative evaluation on tabular datasets

In Table I, we compare the faithfulness (FA) and counterfactual losses (CF) of local masks found by various methods. We answer the following questions.

Does the two-stage optimization find better local masks? Overall, the best-performing methods in local mask faithfulness are in the last four columns representing variants of the SNX using GDA, with different global masks as constraints. In terms of counterfactual loss, the optimal local masks outperform the remaining methods, except on the Credit dataset (SNX-UC has no constraint and can include more salient features).

Why do the other methods underperform SNX and its variants?

- DES has the worst FA except on COMPAS, where SM is the worst. That is because DES is agnostic about the SN architecture and does not consider the reference when finding $\mathbf{M}^s$ to mask both $\mathbf{x}^s$ and $\mathbf{x}^t$. DES does not perform well in CF, indicating that it fails to include the most salient features.
- SNX-global uses $\mathbf{M}^s$ found in Stage 1 and takes the target SN into account. It has good performance in FA. While it is worse than DES in the CF on 3 out of 4 datasets, meaning that it can miss even more salient features.
- SNX-UC performs Stage 2 optimization to find local masks without constraints. It is the runner-up in FA on Adult and Bank and is the best in CF on Credit. However, no constraint leads to less conformity to global invariant masks (Fig. 2).
- SNX-KL is very similar to SNX-UC, except that the soft constraints are implemented as KL-divergence penalty terms. They have similar FA, but SNX-KL significantly underperforms SNX-UC in CF, indicating that SNX-KL selects just enough salient features but can possibly excludes other salient ones.

TABLE I: Performance of local masks. The best methods except Pick-all on each dataset is boldfaced, and the runner-up is highlighted by $*$. $\circ$ indicates significantly better performance according to t-tests. Column Pick-all provides lower (underlined) and upper bounds (overlined) of faithfulness (the counterfactual (CF) of selecting all features is equivalent to the faithfulness (FA) of selecting no feature)).

Performance (mean with std. in parentheses) of each algorithm in terms of *faithfulness*. Lower is better (↓).

Method	Pick-all	DES	SNX-global	SM	SNX-UC	SNX-KL	SNX-DES	SNX-inter	SNX-union	SNX
Adult	0.65 (0.14)	1.67 (2.46)	0.71 (0.18)	0.81 (0.72)	0.68 (0.14) $*$	0.69 (0.20)	**0.66** (0.14)	**0.66** (0.14)	**0.66** (0.14)	**0.66** (0.14) $\circ$
Bank	0.62 (0.16)	0.82 (0.85)	0.72 (0.22)	0.71 (0.24)	0.65 (0.16) $*$	0.66 (0.16)	**0.63** (0.16)	**0.63** (0.16)	**0.63** (0.16)	**0.63** (0.16) $\circ$
Credit	0.64 (0.15)	1.15 (1.73)	0.70 (0.19)	1.13 (1.35)	0.72 (0.19)	0.74 (0.21)	0.69 (0.18) $*$	**0.68** (0.16)	**0.68** (0.16)	**0.68** (0.16) $\circ$
COMPAS	0.62 (0.19)	0.76 (0.57)	0.71 (0.65)	0.79 (0.73)	0.64 (0.17)	0.63 (0.19) $*$	**0.62** (0.19)	**0.62** (0.19)	**0.62** (0.19)	**0.62** (0.19) $\circ$

Performance (mean with std. in parentheses) of each algorithm in terms of *counterfactual*. Higher is better. (↑)

Method	Pick-all	DES	SNX-global	SM	SNX-UC	SNX-KL	SNX-DES	SNX-inter	SNX-union	SNX
Adult	7.50 (2.05)	0.85 (0.49)	0.78 (0.14)	1.02 (1.09)	1.28 (1.57) $*$	0.99 (0.92)	**1.31** (1.62) $\circ$	1.25 (1.53)	1.26 (1.54)	1.26 (1.54)
Bank	6.86 (2.49)	1.26 (1.58)	0.71 (0.11)	1.67 (2.09)	1.80 (2.25)	1.25 (1.47)	2.54 (2.83)	2.78 (2.97) $*$	2.78 (2.97) $*$	**2.79** (2.97) $\circ$
Credit	7.54 (2.26)	0.74 (0.23)	0.83 (0.19)	1.23 (0.82) $*$	**1.24** (1.17) $\circ$	0.89 (0.57)	1.16 (1.10)	0.97 (0.90)	0.97 (0.90)	0.97 (0.88)
COMPAS	7.02 (2.52)	1.86 (2.53)	0.72 (0.09)	1.76 (2.07)	3.65 (3.22)	2.65 (2.74)	4.56 (3.48)	4.93 (3.43)	**5.00** (3.43) $\circ$	4.98 (3.43) $*$

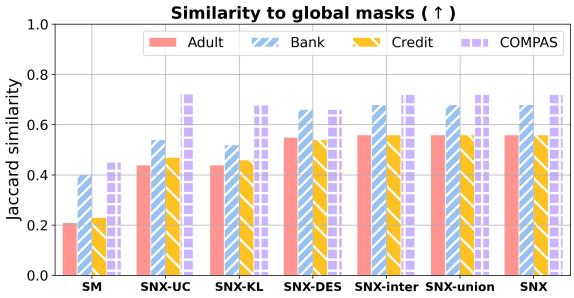

Fig. 2: Conformity of local masks to global ones.

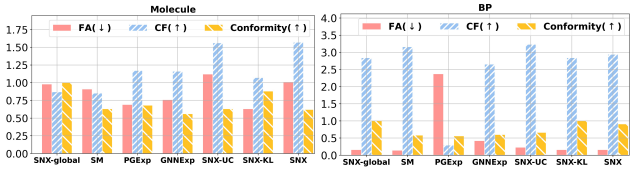

Fig. 3: FA, CF, and conformity of local masks on graph datasets.

How well do the local masks conform to the global masks? In Fig. 2, we report the average of Jaccard similarities between the important major features selected by global masks $\mathbf{N}^s$ and local masks $\mathbf{n}^s$, when comparing $\mathbf{x}^s$ to multiple $\mathbf{x}^t$. The higher similarity indicates the more conformity of local masks to global masks. Except for the COMPAS dataset, SNX with GDA results in the best conformity (highest similarity).

C. Quantitative evaluation on graph datasets

In Fig. 3, we compare local graph masks with respect to FA, CF, and conformity. On the molecule dataset, SNX-KL can achieve the best (lowest) FA and highest conformity ($\sim$90%) to global masks found by SNX-global (whose conformity is 1). SNX with GDA and SNX-UC have the best CF loss, while SNX achieves better FA than SNX-UC. Both PGExp and GNNExp are competitive baselines, whose CF is similar to SNX-KL's but are significantly worse in the other two metrics. SM, PGExp, GNNExp, and SNX-UC, disregard any constraints and have worse conformity ($\sim$ 60%) and worse FA than SNX-KL.

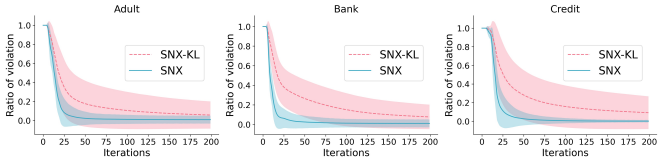

Fig. 4: Ratio of constraints violated by SNX-KL and SNX.

On the BP dataset where graphs are much larger, both SNX-KL and SNX achieve better FA and conformity, and thus more robust than other methods. Since SNX-global has already provided very competitive global masks w.r.t. FA and CF, SNX-KL has a similarly good performance when following SNX-global closely, while SNX sacrifices a bit of conformity to bypass some global constraints for better masks w.r.t. FA and CF. On the contrary, SM has the worst conformity (<50%) with only a slightly better FA than SNX-KL, and it validates the analysis in Section II-C that SM has worse robustness as it is too sensitive to various references given a query.

D. More on conformity

Fig. 4 compares the ratio of constraints violated by SNX and SNX-KL during optimization for three datasets. We evaluate Eqs. (7)-(8) using the masks without selecting top edges.

E. Qualitative evaluation

Fig. 5 shows masks learned by SNX-global and four local methods on the BP dataset, respectively. The numbered colored nodes are brain regions of interest (ROIs) related to the human dorsal and ventral systems, and edges connected to these nodes are colored accordingly. These ROIs, such as ventromedial prefrontal cortex (#39), dorsolateral prefrontal cortex (#17), superior parietal lobule (#10), and anterior cingulate cortex (#51), could be highly affected by bipolar disorder according to neuroscience studies [28]. As a result, SNX-global well captures many edges adjoined to these ROIs, particularly the interconnections among the aforementioned four ROIs. Also, SNX and SNX-KL found local masks more conformal to global mask than the baselines.

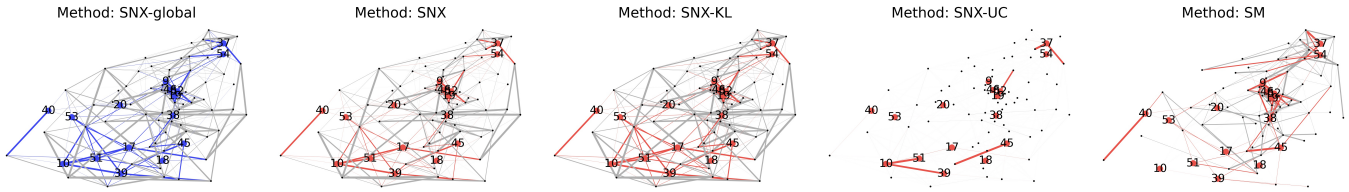

Fig. 5: Explaining brain network comparisons on the BP dataset. Details are in Section IV-E.

V. Related work

Explainability have been extensively studied for tabular data [19], images [20], and more recently, graph data [26], [1], [15], [7], [12]. Robustness in explanations is gaining attention [6], [24]. In [6], the goal is to train neural networks for image classification that has robust explanations with malicious data manipulations. In [24], vulnerability of explanations of architecture with a single image is analyzed.

Explaining why two instances are similar or different has been only sparsely researched [7], [12], [22]. [7] extracts the most contrastive parts of the graphs to tell the similarity among those in the same class. [12] extracts contrasting subgraphs to discriminate two different groups of brain networks.

The closest work to this work is [23], where the authors used GDA to solve a constrained optimization problem to find counterfactual explanations for a classifier. Rather, we apply the algorithm to solve a novel problem for SN explanation.

VI. Conclusions

We address the robustness in explaining Siamese networks due to the varying reference objects. We formulate the global invariance in explaining SN as a self-supervised learning problem, and propose to use SGD and GDA to find superior explanations with provably and empirically convergence. Case studies demonstrate the relevance of the found invariants when explaining individual comparisons.

Acknowledgement

Chao and Sihong were supported in part by the National Science Foundation under Grants NSF IIS-1909879, NSF CNS-1931042, and NSF IIS-2008155. Any opinions, findings, conclusions, or recommendations expressed in this document are those of the author(s) and should not be interpreted as the views of any U.S. Government. Yifan and Xi were supported by Natural Science Foundation of China (No.61976026) and 111 Project (B18008).

References

[1] Federico Baldassarre and Hossein Azizpour. Explainability techniques for graph convolutional networks. *ICML workshop*, 2019.

[2] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[3] Chao Chen, Yifan Shen, Guixiang Ma, Xiangnan Kong, Srinivas Rangarajan, Xi Zhang, and Sihong Xie. Self-learn to explain siamese networks robustly. *https://arxiv.org/pdf/2109.07371.pdf*.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 2020.

[5] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *NeurIPS*, 2019.

[6] Ann-Kathrin Dombrowski, Christopher J. Anders, K. Müller, and P. Kessel. Towards robust explanations for deep neural networks. *ArXiv*, abs/2012.10425, 2020.

[7] Lukas Faber, Amin K Moghaddam, and Roger Wattenhofer. Contrastive graph neural network explanation. In *ICML Workshop*, 2020.

[8] Amirata Ghorbani, Abubakar Abid, and James Y Zou. Interpretation of Neural Networks is Fragile. In *AAAI*, 2017.

[9] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38, 2017.

[10] Wei Jin, Tyler Derr, Haochen Liu, Yiqi Wang, Suhang Wang, Zitao Liu, and Jiliang Tang. Self-supervised learning on graphs: Deep insights and new direction. *arXiv preprint arXiv:2006.10141*, 2020.

[11] Peter Kolesar and Janet L Showers. A robust credit screening model using categorical data. *Management Science*, 31(2):123–133, 1985.

[12] Tommaso Lanciano, F. Bonchi, and A. Gionis. Explainable classification of brain networks via contrast subgraphs. *SIGKDD*, 2020.

[13] Tianyi Lin, Chi Jin, and Michael Jordan. On Gradient Descent Ascent for Nonconvex-Concave Minimax Problems. In *ICML*, 2020.

[14] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774, 2017.

[15] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *NeurIPS*, 33, 2020.

[16] Guixiang Ma, Nesreen K Ahmed, Theodore L Willke, Dipanjan Sengupta, Michael W Cole, Nicholas B Turk-Browne, and Philip S Yu. Deep graph similarity learning for brain data analysis. In *CIKM*, 2019.

[17] Bryan A Plummer, Mariya I Vasileva, Vitali Petsiuk, Kate Saenko, and David Forsyth. Why do these match? explaining the behavior of image similarity models. *ECCV*, 2019.

[18] Marcus E Raichle. The brain's default mode network. *Annual review of neuroscience*, 38:433–447, 2015.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *KDD*, 2016.

[20] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[21] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. A sparse-group lasso. *J. Comput. Graph. Stat.*, 2013.

[22] Lev V Utkin, Maxim S Kovalev, and Ernest M Kasimov. An explanation method for siamese neural networks. *arXiv:1911.07702*, 2019.

[23] Sandra Wachter, Brent D Mittelstadt, and Christopher Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *ArXiv*, abs/1711.0, 2017.

[24] Zifan Wang, H Wang, Shakul Ramkumar, Matt Fredrikson, Piotr Mardziel, and A Datta. Smoothed Geometry for Robust Attribution. In *NeurIPS*, 2020.

[25] Xiaokai Wei, Sihong Xie, Bokai Cao, and S Yu Philip. Rethinking unsupervised feature selection: From pseudo labels to pseudo must-links. In *ECML-PKDD*, pages 272–287. Springer, 2017.

[26] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, pages 9244–9255, 2019.

[27] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *NeurIPS*, 2020.

[28] Niccolò Zovetti, Maria Gloria Rossetti, Cinzia Perlini, Eleonora Maggioni, Pietro Bontempi, Marcella Bellani, and Paolo Brambilla. Default mode network activity in bipolar disorder. *Epidemiol. Psychiatr. Sci.*, 29, 2020.