Explanation Consistency Training: Facilitating Consistency-Based Semi-Supervised Learning with Interpretability

Tao Han,[1] **Wei-Wei Tu,**[2] **Yu-Feng Li**[1*]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2]4Paradigm Inc., Beijing, China
hant@lamda.nju.edu.cn, tuww.cn@gmail.com, liyf@lamda.nju.edu.cn

Abstract

Unlabeled data exploitation and interpretability are usually both required in reality. They, however, are conducted independently, and very few works try to connect the two. For unlabeled data exploitation, state-of-the-art semi-supervised learning (SSL) results have been achieved via encouraging the consistency of model output on data perturbation, that is, *consistency assumption*. However, it remains hard for users to understand how particular decisions are made by state-of-the-art SSL models. To this end, in this paper we first disclose that the consistency assumption is closely related to causality invariance, where causality invariance lies in the main reason why the consistency assumption is valid. We then propose ECT (Explanation Consistency Training) which encourages a consistent reason of model decision under data perturbation. ECT employs model explanation as a surrogate of the causality of model output, which is able to bridge state-of-the-art interpretability to SSL models and alleviate the high complexity of causality. We realize ECT-SM for vision and ECT-ATT for NLP tasks. Experimental results on real-world data sets validate the highly competitive performance and better explanation of the proposed algorithms.

Introduction

Conventional machine learning assumes that a large number of labeled data are readily available for training, and a black-box nature of complex models is readily excellent for deployment. However, in many real tasks such as image understanding or disease diagnosis, obtaining ample labeled examples is difficult since labeling comes at costly human resources, and it is desirable for users to understand how particular decisions are made by these models. Unlabeled data exploitation and interpretability are usually both required in reality. They, however, conduct independently and very few works try to connect the two, although much progress has been made in these two aspects recently (Miyato et al. 2018; Berthelot et al. 2019; Lipton 2018; Etmann et al. 2019). In this work, we try to consider the use of interpretability to facilitate unlabeled data exploitation.

As a major paradigm of unlabeled data exploitation, semi-supervised learning (SSL) has been extensively studied and made much progress. One popular SSL paradigm was based on *smooth assumption* (Zhu and Goldberg 2009), i.e., similar input data should own similar output labels. Recently, deep SSL methods extend such an idea, by constraining the outputs according to *consistency assumption* (Laine and Aila 2016; Tarvainen and Valpola 2017; Miyato et al. 2018), that is, data perturbations or augmentations should own similar output labels. Consistency assumption has been shown state-of-the-art performance in various mainstream SSL tasks such as computer vision and NLP (Miyato et al. 2018).

Many SSL studies have been proposed to interesting data perturbations or augmentations (Miyato et al. 2018; Berthelot et al. 2019) under *consistency assumption*. However, it remains unclear why *consistency assumption* is effective, and it is still hard for users to understand how particular decisions are made by state-of-the-art deep SSL models.

In this paper, we first realize that consistency assumption is closely related to *causality invariance* (Cartwright 2003). Specifically, the consistency assumption implies that the causality to the class labels will not be destroyed by data perturbation or augmentation. On the other hand, once the data perturbation destroys or violates the causality invariance, the consistency assumption may no longer be effective. In this case, consistency assumption behaves as a rough surrogate to causality invariance, and causality invariance plays a key role to the success of SSL models.

However, direct learning of causality is infeasible, because of the high complexity of causal inference, especially for big data and deep neural network models with high capacities (Spirtes 2010). Fortunately, recent studies discovered that model explanation may provide effective clues and excellent support to the causal relationships (Lipton 2018; Moraffah et al. 2020). In this paper, we regard model explanation as a better surrogate to causality invariance, and propose a new SSL model ECT (Explanation Consistency Training). ECT encourages a consistent reason for the model decision under unlabeled data perturbation instead of output invariance in the consistency assumption. ECT is able to bridge state-of-the-art model interpretations to SSL models. Figure 1 illustrates the motivation. Our basic idea is that a better surrogate for causality invariance may lead to better performance and interpretability of SSL. We realize ECT-SM for vision tasks and ECT-ATT for NLP tasks.

Our contributions mainly include the followings:

*Yu-Feng Li is the corresponding author.

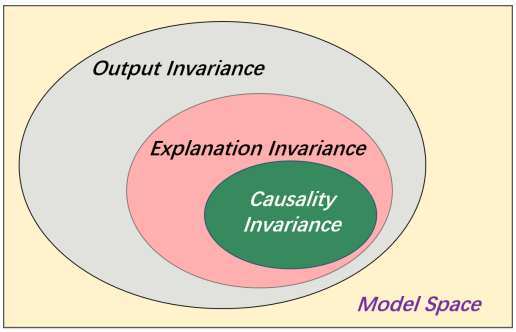

Figure 1: The Venn diagram illustrates the scopes of output invariance (consistency assumption), explanation invariance (explanation consistency) and the causality invariance. Better surrogate for causality invariance may lead to better performance and interpretability.

- A new explanation on the effectiveness of consistency assumption via causality invariance.
- A new SSL method ECT with the use of explanation consistency training.
- Experimental results on real-world data sets validate the highly competitive performance and better explanation of the proposed algorithms.

Explanation Consistency Training

In this section, we first briefly introduce state-of-the-art SSL models based on the consistency assumption, and then present the framework of ECT, followed by its deployment on mainstream tasks with optimization.

Deficiency of Consistency-Based SSL Model

In SSL, we are given a few labeled data $\{X_L, Y_L\} = \{(\mathbf{x}_l, y_l)\}_{l=1}^{L}$ and a large number of unlabeled data $X_u = \{\mathbf{x}_{u+L}\}_{u=1}^{U}$. Usually, we have $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d, y \in \mathcal{Y} = \{1, \cdots, C\}$ where d is the input data dimension and C is the number of output classes. The SSL model $h(\mathbf{x}, \theta) : \{\mathcal{X} : \Theta\} \to \mathcal{Y}$ parameterized by $\theta \in \Theta$ learns from both the labeled and unlabeled data.

The state-of-the-art deep SSL methods are conducted in the teacher-student framework (Qi and Luo 2019). The core idea is to construct a teacher from data perturbation (or augmentation), and use the teacher's outputs to supervise the student model on unlabeled data.

Specifically, let $\ell(\cdot, \cdot) : \{\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}\}$ denote the empirical supervised loss on labeled data. Denote the perturbation function as $\mathcal{A}(\mathbf{x}) : \mathcal{X} \to \mathcal{X}$ (e.g., when $\mathcal{A}$ refers to adversarial augmentations, it turns out to Virtual Adversarial Training (VAT) (Miyato et al. 2018)) and the consistency loss on unlabeled data as $\Omega(\mathbf{x}; \mathcal{A}, \theta) : \mathcal{X} \to \mathbb{R}$. The objective of teacher-student model can be formalized as

$$\min_{\theta \in \Theta} \frac{1}{L} \sum_{l=1}^{L} \ell(h(\mathbf{x}_l; \theta), y_l) + \alpha \frac{1}{U} \sum_{u=1}^{U} \Omega(\mathbf{x}_{u+L}; \theta) \quad (1)$$
$$\text{s.t.} \quad \Omega(\mathbf{x}_u; \mathcal{A}, \theta) = \|h(\mathbf{x}_u; \theta) - h(\mathcal{A}(\mathbf{x}_u); \theta)\|_2^2,$$

$\alpha \in \mathbb{R}$ balances the empirical supervised loss and the consistency loss on unlabeled data. Different choice of $\mathcal{A}(\mathbf{x})$ and Ω in Eq.(1) realizes various state-of-the-art deep SSL models, e.g., Temporal Ensembling (Laine and Aila 2016), VAT (Miyato et al. 2018), MixMatch (Berthelot et al. 2019).

Though consistency assumption has been shown effective in various mainstream tasks, it remains unclear why it is valid, and thus it remains unclear to understand how particular decisions are made by state-of-the-art SSL models.

Proposed ECT Framework

Intuitively, consistency assumption in Eq.(1) is closely related to *causality invariance*. More specifically, the core idea of consistency assumption is that, data perturbations or augmentations should own similar output labels. In other words, the causality to derive the class labels is not affected by data perturbation or augmentation. Taking image classification as an example, the causality of the factors triggering the ground-truth tags of an image will not be changed by the perturbations of some image pixels. This may be true intuitively, but not rigorous as consistency assumption does not explicitly constrain the causality. In fact, on the other hand, once the pixel perturbations destroy or violate the causality to the ground-truth tags, consistency assumption may no longer be effective. As an example illustrated in Figure 2, although the model outputs for two cats still seem to be consistent, the underlying factors revealed by model explanations triggering the ground-truth tags have been changed. This to some extent discloses that consistency assumption is not robust to causality. Ross, Hughes, and Doshi-Velez (2017) proposed that *the model decisions should be right for the right reasons*. A more precise assumption if possible towards underlying causality invariance may be more preferable.

However, it is widely known that direct learning of causality is not practical, due to the high complexity of causal inference, especially for big data and deep neural network models with high capacities (Spirtes 2010). Fortunately, recent studies discovered that model explanation is an excellent surrogate on the causal relationships and provides effective clues (Lipton 2018; Moraffah et al. 2020). For example, a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnosis; interpretable learning models are able to provide clues about the causal relationships between physiologic signals and affective states, etc (Lipton 2018).

We therefore present a new attempt for well-performing and interpretable SSL models by leveraging model explanation as a better surrogate to causality invariance. We propose a new SSL model based on explanation consistency training, in short, ECT. Unlike consistency assumption based on output invariance, ECT encourages a consistent reason for the model decision, under unlabeled data perturbation.

Figure 3 shows the framework of ECT. Given model h and parameter θ, an *explainer* is denoted by $\mathcal{I}(\mathbf{x}; \theta)$, which returns the local explanation of model output $h(\mathbf{x}; \theta)$ for any instance $\mathbf{x}$. Similar to consistency assumption, by measuring explanation consistency $E(\mathbf{x}; \mathcal{I}, \theta)$ with some popular loss,

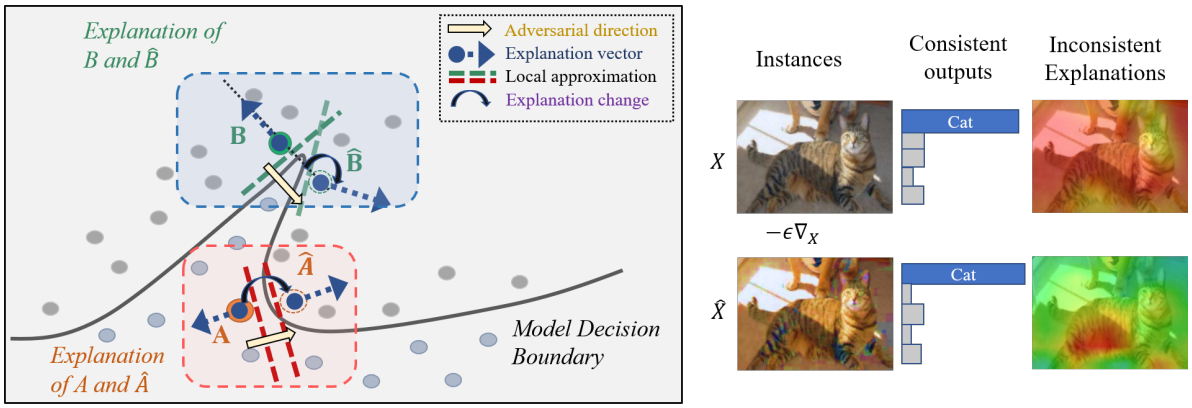

Figure 2: The illustrated two cases of how model outputs and explanations (vectors of Saliency Map) change under adversarial perturbation. In the case of A, both model explanations and outputs change, and output consistency methods handle well. However, in the case of B, model output remains unchanged while the decision reasons change dramatically. The right part provides an illustrative case of a cat, where consistent model outputs come from inconsistent gradient-based explanations.

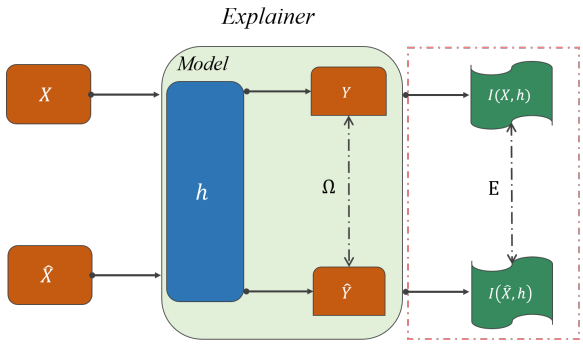

Figure 3: The ECT framework. The *explainer* is a flexible container of various SSL models.

e.g., mean square error (MSE), ECT is then formulated as:

$$\min_{\theta\in\Theta} \frac{1}{L}\sum_{l=1}^{L}\ell(h(\mathbf{x}_l;\theta),y_l)+\lambda\frac{1}{U}\sum_{u=1}^{U}E(\mathbf{x}_{u+L};\mathcal{I},\theta) \quad (2)$$
$$\text{s.t.} \quad E(\mathbf{x}_u;\mathcal{I},\theta)=\|\mathcal{I}(\mathbf{x}_u;\theta)-\mathcal{I}(\mathcal{A}(\mathbf{x}_u);\theta)\|_2^2.$$

where $\lambda\in\mathbb{R}$ balances the empirical supervised loss and the explanation consistency loss on unlabeled data.

Notice that *explainer* $\mathcal{I}$ can be updated accordingly with the development of the interpretable machine learning community. In the following, we first compare explanation consistency with output consistency, and then introduce two explainers on two mainstream tasks, i.e., computer vision and natural language process, respectively.

Explanation Consistency vs Output Consistency

To analyze the properties of explanation consistency, we introduce the notion of local *Lipschitz* condition proposed in (Alvarez-Melis and Jaakkola 2018).

Definition 1 *$h(\mathbf{x};\theta)$ is locally **difference-bounded** by $\mathcal{I}$, if for every $\mathbf{x}_0$ there exist $\delta>0$ and $L\in\mathbb{R}$ such that $\|\mathbf{x}-\mathbf{x}_0\|<\delta$ implies $\|h(\mathbf{x};\theta)-h(\mathbf{x}_0;\theta)\|\le L\|\mathcal{I}(\mathbf{x})-\mathcal{I}(\mathbf{x}_0)\|$.*

This condition resembles the local Lipschitz continuity that similar explanations trigger similar outputs and it allows the constant L (and δ) to depend on $\mathbf{x}_0$. Therefore, explanation consistency would necessarily be output consistency, once the local Lipschitz condition holds for $h(\mathbf{x};\theta)$.

In practice, as for human understandings friendly, many interpretable machine learning methods derive explanations via constructing a generalized linear map between inputs and outputs (Lundberg and Lee 2017), such as,

$$h^c(\mathbf{x};\theta)=\mathcal{I}(\mathbf{x};\theta)^T\mathbf{x}+b, \quad (3)$$

where c is the interested class output and b is the interception item. It is easy to find that $h^c(\mathbf{x};\theta)$ fits the local Lipschitz condition and therefore, explanation consistency generally implies output consistency, but the opposite is not true.

Mainstream Tasks and Optimization

In the following, we introduce two easy-to-optimize techniques as the *explainer* for two mainstream tasks, i.e., computer vision and natural language process, respectively.

Saliency Map as Explanation In vision tasks, Saliency Map (SM) (Simonyan, Vedaldi, and Zisserman 2014) is a common used and effective gradient-based explanation method. By adopting SM as the *explainer*, i.e., $\mathcal{I}(\mathbf{x};\theta)=\nabla_{\mathbf{x}}h^c(\mathbf{x};\theta)$, where $c=\arg\max_{c\in C}h^c(\mathbf{x};\theta)$ corresponding to the class index with the maximum output. Using first-order Talyor expansion on $h^c(\mathbf{x};\theta)$ at point $\mathbf{x}_0$ as follows:

$$\begin{aligned}h^c(\mathbf{x};\theta)&\approx h^c(\mathbf{x}_0;\theta)+\nabla_{\mathbf{x}}h^c(\mathbf{x}_0;\theta)^T(\mathbf{x}-\mathbf{x}_0)\\&=\nabla_{\mathbf{x}}h^c(\mathbf{x}_0;\theta)^T\mathbf{x}+b,\end{aligned} \quad (4)$$

and we can see that saliency map satisfies the form of Eq.(3), therefore consistent saliency map implies consistent output according to Definition 1.

Then we have a new SSL model termed ECT-SM as

$$\min_{\theta\in\Theta} \quad \frac{1}{L}\sum_{l=1}^{L}\ell(h(\mathbf{x}_l;\theta),y_l) \tag{5}$$
$$+\lambda\frac{1}{U}\sum_{u=1}^{U}\|\nabla_{\mathbf{x}}h^c(\mathbf{x}_{u+L};\theta)-\nabla_{\mathbf{x}}h^c(\mathcal{A}(\mathbf{x}_{u+L});\theta)\|_2^2,$$

and the regularizer will be

$$\begin{aligned}&\|\nabla_{\mathbf{x}}h^c(\mathbf{x}_{u+L};\theta)\|_2^2+\|\nabla_{\mathbf{x}}h^c(\mathcal{A}(\mathbf{x}_{u+L});\theta)\|_2^2\\&-2\,\langle\nabla_{\mathbf{x}}h^c(\mathbf{x}_{u+L};\theta),\,\nabla_{\mathbf{x}}h^c(\mathcal{A}(\mathbf{x}_{u+L});\theta)\rangle.\end{aligned} \tag{6}$$

ECT-SM bounds the input gradient to achieve a low energy state, which is endowed with inherent robustness to adversarial attacks as discussed in (Ross and Doshi-Velez 2018). Due to the loss on gradient-based explanation, the solving of ECT-SM is second-order optimization. To reduce the computation cost, we reuse the computed saliency map $\nabla_{\mathbf{x}}h^c(\mathbf{x};\theta)$ in the augmentation function $\mathcal{A}(\mathbf{x}_u) = x_u - \epsilon\nabla_{\mathbf{x}}h^c(\mathbf{x}_u;\theta)$. It is noteworthy that, the explanation will take the same operation to maintain spatial alignment, once some scaling or rotation prior is applied in the augmentation function.

Attention as Explanation Attention is a ubiquitous component in NLP tasks that is naturally interpretable. Denote the attention score vector of input $\mathbf{x}$ as $a_{\mathbf{x}}\in\mathbb{R}^d$, the embedding matrix to aggregate as $\Phi(\mathbf{x})\in\mathbb{R}^{d\times k}$ and the forward network be F, the model output is written as:

$$h(\mathbf{x};\theta) = F(\Phi(\mathbf{x})^T a_{\mathbf{x}}). \tag{7}$$

If F is a linear fully connected layer, then attention consistency also implies ouput consistency according to Eq.(3). Putting the attention into the *explainer* realizes a new SSL model ECT-ATT, where the regularizer is written as

$$E(\mathbf{x}_u;\mathcal{I},\theta) = \|a_{\mathbf{x}_u} - a_{\mathcal{A}(\mathbf{x}_u)}\|_2^2, \tag{8}$$

Hybrid Consistency SSL Model It is natural to further exploit the advantages of explanation consistency and output consistency simultaneously in a whole framework, by putting the two consistency regularizations together,

$$\begin{aligned}\min_{\theta\in\Theta}\quad &\frac{1}{L}\sum_{l=1}^{L}\ell(h(\mathbf{x}_l;\theta),y_l)\\&+\frac{1}{U}\sum_{u=1}^{U}\left(\alpha\Omega(\mathbf{x}_{u+L};\theta)+\lambda E(\mathbf{x}_{u+L};\mathcal{I},\theta)\right).\end{aligned} \tag{9}$$

We call such an SSL model ECT-hybrid.

Related Work

Current consistency-based SSL is mostly realized with a teacher-student training framework (Qi and Luo 2019). The output consistency SSL methods can be roughly divided into spatial consistency and time consistency. Spatial consistency methods focus on the smoothness of output at a given training moment. Ladder Network (Rasmus et al. 2015) and Π-Model (Laine and Aila 2016) construct teacher via adding noise to network layers, and the clean student is noise-free. Virtual Adversarial Training (VAT) (Miyato et al. 2018) employs an adversarial teacher to get a robust student model. MixMatch (Berthelot et al. 2019) combines several techniques to build a unified SSL model. Time consistency methods pursue consistent outputs during a training sequence. Temporal Ensembling (Laine and Aila 2016) and Mean Teacher (Tarvainen and Valpola 2017) ensemble the models at different training stages as a better teacher model. TC-SSL (Zhou, Wang, and Bilmes 2020) utilizes self-supervision to select unlabeled instances possessing time consistent pseudo-labels as teachers.

Interpretable Machine Learning (IML) concentrates on bridging the gap between complex models and human understanding to improve the safety and reliability of machine learning (Lipton 2018). Post-hoc explanations reveal the decision process or dominated elements of a given model and input. For agnostic black-box models, LIME (Ribeiro, Singh, and Guestrin 2016) approximates the model output of a given instance with a local sparse linear model, where the model coefficients realize the explanation. SHAP (Lundberg and Lee 2017) provides a unified linear method to approximate feature contributions from game theory. For white-box models, explanations are specially designed according to the model structure. Performing first-order Taylor expansion at the input data point (Bach et al. 2015) is effective for differentiable models such as CNN and LSTM. Saliency Map (SM) (Simonyan, Vedaldi, and Zisserman 2014) computes the gradient of the model output regarding the input image to visualize the contribution or sensitivity of each pixel. More gradient-based tools are designed to enhance visual effects (Springenberg et al. 2015; Selvaraju et al. 2017), but Nie, Zhang, and Patel (2018) proves that GuidedBP is essentially doing (partial) image recovery, which is unrelated to the network decisions.

Recently, there have been few works applying IML to improve models. Etmann et al. (2019) discovers the connection between saliency maps and model robustness. Ross and Doshi-Velez (2018) regularizes the norm of input gradients to get a robust and interpretable model. GradMask (Viviano et al. 2019) uses image masks to supervise saliency maps to ignore distracting features. Wang et al. (2020) proposes to learn from explanations with neural module execution tree. All the above efforts are devoted to supervised learning, while the attempts on SSL have not been thoroughly studied yet.

Experiments

Firstly, we study ECT on the benchmark datasets to evaluate the effectiveness. Then we compare the visual effect and measure the robustness of explanations. Finally, we study ECT on biased data sets. To fairly compare each method, we control the other implementations to be the same and set the following α, λ to realize different models.

- Pure SL with labeled data: $\alpha = 0, \lambda = 0$.
- SSL with output consistency (VAT) : $\alpha > 0, \lambda = 0$.
- SSL with explanation consistency: $\alpha = 0, \lambda > 0$.
- SSL with hybrid consistency: $\alpha > 0, \lambda > 0$.

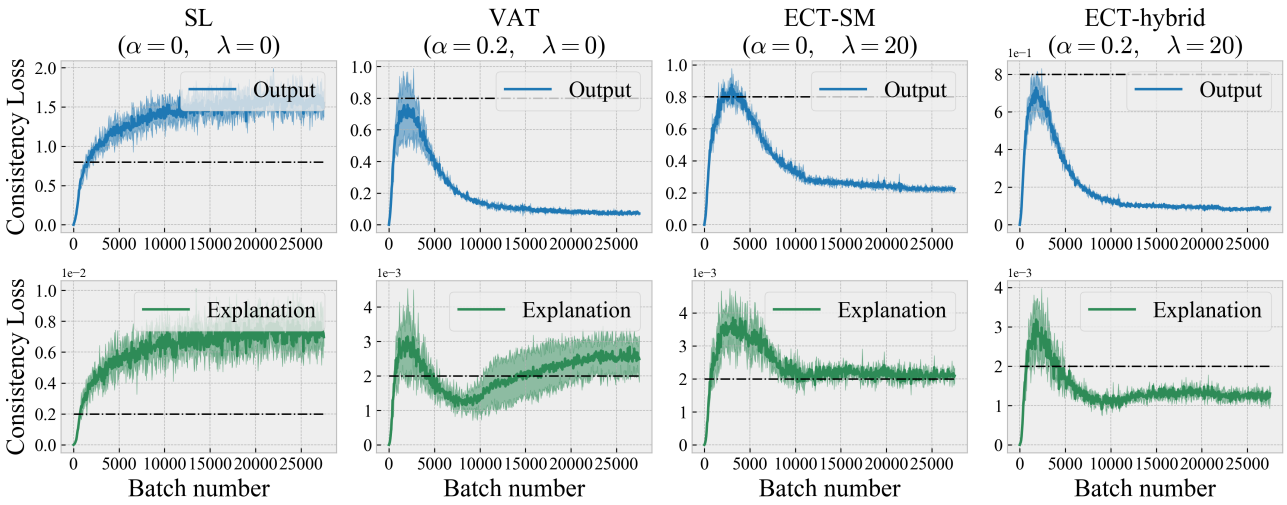

Figure 4: Tranining procedures on the Fashion-MNIST dataset with $\alpha \in \{0, 0.2\}, \lambda \in \{0, 20\}$. Each column shows the training loss of output consistency (above) and explanation consistency (below) of a corresponding model. For comparison, the black dashed lines mark the same level of y-axis for each loss.

	Fashion-MNIST		
#labels	100	200	1000
Baseline (SL)	68.90±1.97	73.50±1.17	81.10±0.79
VAT	71.89±1.41	75.46±0.86	82.58±0.73
ECT-SM	72.37±1.34	**76.23±0.68**	82.92±0.59
ECT-hybrid	**72.45±1.58**	75.65±0.59	**82.99±0.61**
	MNIST		
#labels	100	200	1000
Baseline (SL)	87.74±1.46	91.63±0.55	96.07±0.23
VAT	88.61±0.80	92.15±0.76	96.64±0.28
ECT-SM	**89.39±1.24**	92.43±0.92	96.38±0.15
ECT-hybrid	88.97±1.61	**92.69±0.82**	**96.80±0.16**

Table 1: The test ACC on the Fashion-MNIST and MNIST.

Performance on Benchmarks

Datasets We test model performances on the benchmark datasets of MNIST (LeCun, Cortes, and Burges 2010) and Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017). MNIST contains gray images of hand-written digits from 0 to 9. Fashion-MNIST is a dataset of Zalando's article images, including {*T-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot*}. Both of them are 10-class classification tasks on images, consisting of a training set of 60,000 examples and a test set of 10,000 examples. We sample labeled data evenly from each class, and the rest in the training set is used as unlabeled data. Each experiment is repeated for 5 times with different labeled data numbers of {100, 200, 1,000} and 1,000 data for validation.

We build a convolutional neural network using BatchNorm and ReLu functions, followed by a 128-unit fully connected layer. The optimizer is SGD with a decayed learning rate, $1 \times e^{-4}$ weight decay, and the momentum is 0.9. All the models are trained for 50 epochs on unlabeled data, so it is close to 30,000 batch iterations. The perturbation extend is set to $\epsilon = 2.0$, and the value of α, λ are set according to the magnitude of their losses without heavy tuning. All the experiments are conducted with Pytorch[1].

Training Curves and Results We plot the training procedure of 5 times random split experiments on Fashion-MNIST in Figure 4 to show the effect of different regularizers. As expected, without any consistency regularization in supervised learning, the two losses keep increasing. An interesting phenomenon is that the explanation loss of VAT shows a trend of up and down fluctuations. We think it reflects how VAT affects learning when exploiting unlabeled data. In the beginning, the network learns consistent decision patterns, and then VAT progressively enforces patterns to grow more and more complicated to keep output consistency. From this perspective, the output consistency method fails explanation consistency due to the high model capacity such that it can easily learn inconsistent patterns including biases. The proposed ECT-SM with only explanation consistency promotes both losses to converge, and ECT-hybrid gets more smooth curves during optimization. The ablation outcomes verify our motivation and support our theoretical result that ECT-SM implies output consistency.

Without heavy-tuning, the test accuracy in Table 1 shows that the proposed ECT-SM and ECT-hybrid have competitive performance. Therefore, explanation consistency is more generic and powerful towards consistent SSL. In the next section, we will show that the proposed ECT endows better explanations as well.

Visual Interpretability and Robustness

Alvarez-Melis and Jaakkola (2018) has discussed that explicitness, faithfulness and stability of explanations are cru-

[1]https://pytorch.org

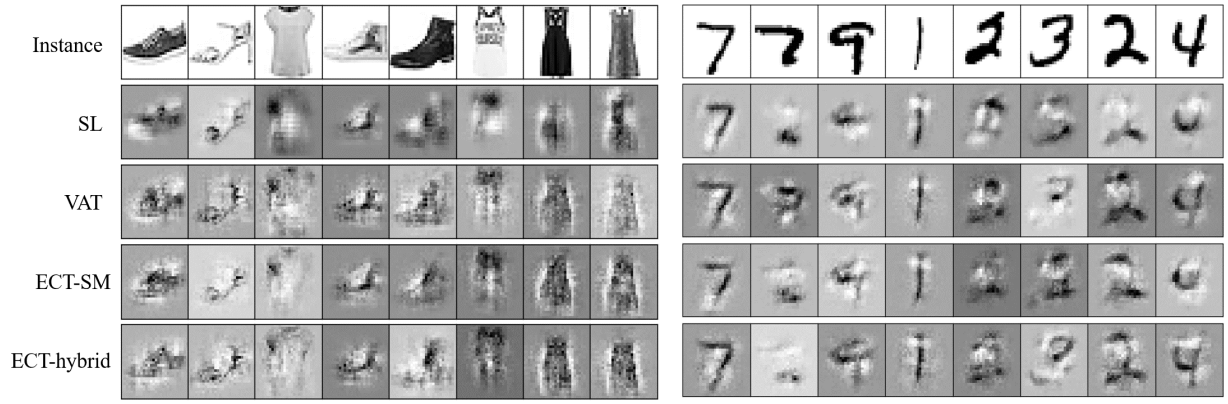

Figure 5: Visual effect of explantions (Saliency Map) from different models. The left eight images come from the Fashion MNIST test set and the right come from MNIST test set.

cial properties for meaningful explanations. In this section, we show that constraining SSL for a consistent reason sheds light on the robustness of interpretability.

Visual Effect We randomly sample 8 instances from two vision datasets and visualize the saliency map explanations as grey images in Figure 5. The saliency maps of the proposed ECT-SM and ECT-hybrid highlight more informative details of the images than that of supervised learning. It looks like there exists much "pepper noise" on the VAT saliency map and this reflects that the explanations of VAT are not very consistent and smooth. The proposed explanation consistency not only earns more details than supervised learning but also achieves more smooth explanations. Meanwhile, the norm of explanation vectors from ECT is much smaller due to the low energy L2-regularization, which benefits the explanation robustness and stability.

Quantitative Robustness To measure and compare the robustness of model explanations, we measure the relative change of explanations regarding perturbations (Alvarez-Melis and Jaakkola 2018), formulated as follows:

$$C(\mathbf{x}) = \frac{\|\mathcal{I}(\mathcal{A}(\mathbf{x});\theta) - \mathcal{I}(\mathbf{x};\theta)\|_2^2}{\|\mathcal{A}(x) - \mathbf{x}\|_2^2}. \tag{10}$$

We follow the metric with adversarial perturbations and smaller $C(\mathbf{x})$ means better robustness and stability of model explanations. Each model is randomly chosen from 5 times experiments and the quantitative results on the Fashion MNIST test datasets are shown in Figure 6. Similar results are found on the MNIST dataset.

The value of $C(\mathbf{x})$ decreases when we increase ϵ because the denominator increases larger. We observe the proposed ECT-SM and ECT-hybrid to consistently and substantially outperform pure SL and VAT in this metric with different ϵ. The better explanation results also inspire us that requiring consistent explanations will force the model to learn more stable features and filter the distracting biases.

Overfitting Prevention

Biases are the distracting factors of true evidence, which fit the training data well while fail to generalize. In the real world, insufficient sampling and the intrinsic prejudice in our society exacerbate biases such as genders, colors, and so on. SSL faces great challenges from both, especially for a teacher-student model, where biases often make the teacher over-confident to mislead its student. In this experiment, we show that striving for consistent reasons is effective to fight against overfitting.

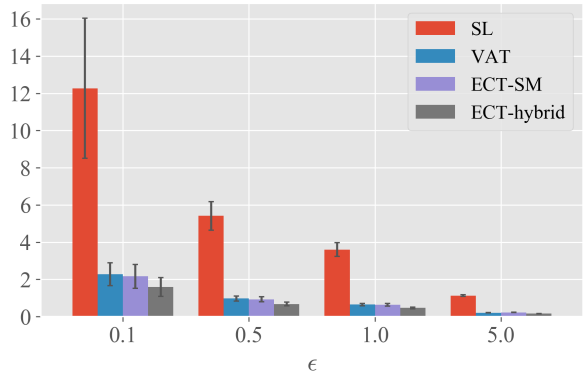

Figure 6: The relative change of saliency map with respect to adversarial perturbations on Fashion MNIST. The ϵ controls the adversarial perturbation weight.

Dataset We use the biographies texts collected by (De-Arteaga et al. 2019), which is collected to study biases against gender-minorities in occupation classification models. And we follow the (Pruthi et al. 2020) to carve out a binary classification task of distinguishing between surgeons and (non-surgeon) physicians, where a majority of surgeons ($> 80\%$) in the dataset are male. To enhance the gender bias, (Pruthi et al. 2020) further downsample minority classes — female surgeons, and male physicians by a factor of ten. We randomly select 100 label texts from 17629 training texts as labeled data and the rest 17529 as unlabeled data, and the test dataset has 5037 texts. To compare the performance with/without biases, we simulate an unbiased test dataset by anonymizing those gender words in the test

Gender words	Replacements
he/she	they
him	them
his/her/hers	their
himself/herself	themself
ms./mrs./mr.	m.

Table 2: The replacement dictionary of biased words reflecting genders. The unbiased dataset is constructed through replacing the left with the right accordingly.

Parameter	Biased test	Unbiased test
Baseline (SL)	92.64±(1.99)	72.68±(3.56)
VAT	74.94±(2.56)	74.31±(2.72)
ECT-ATT	**94.00±(0.99)**	77.76±(3.07)
ECT-hybrid	89.38±(2.82)	**79.94±(2.48)**

Table 3: The ACC of different models on gender biased and unbiased test dataset.

dataset with an replacement dictionary listed in Table 2.

Model Architecture We train a simple embedding attention model (Pruthi et al. 2020), where the attention is directly over word embeddings (128 dimensions). Then the word embeddings are weighted aggregated by attention score, followed by a linear layer and a softmax to perform prediction. The augmentation method is randomly replacing ϵ words of top attention scores with other words in this text, where ϵ is the hyper-parameter. The ECT-ATT is realized based on attention explanations.

Performance We run these models for 5 times and the performances of models on the biased and unbiased test dataset are listed in Table 3. The results show that pure SL is overfitting on the gender biases easily that achieves 92.64% accuracy on the biased test. However, the performance drops to 72.68% on the unbiased test without gender words. The performance of VAT drops dramatically on the biased test but increases not much on the unbiased test. After regularizing the explanation (attention) with ECT-ATT and ECT-hybrid, the performance on the unbiased test gets significantly improved.

Attention Distribution To verify our proposal, we look inside into these models by summing the attention scores on the biased words according to Table 2 called biased attention. Figure 7 shows the results from one run of our experiments. The histogram on the left shows the number of test instances where the corresponding model has the highest score among the four models. The right part is the distribution of biased attention over the two classes.

As we can see, VAT owns the most instances with the highest biased attention, which verifies our intuition that the biases will make the teacher-student model over-confident and cause severe overfitting. The visualization effect of attention-word distribution is drawn in Figure 8. The size

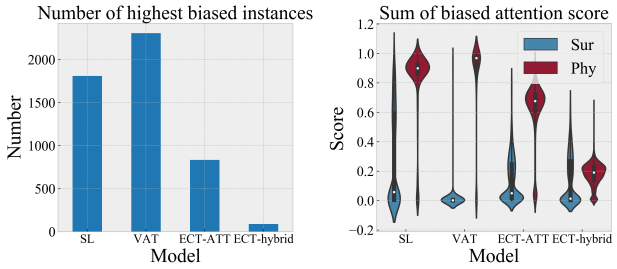

Figure 7: The histogram on the left shows the number of test instances where the corresponding model has the highest score among the four models. The right part is the distribution of biased attention over the two classes.

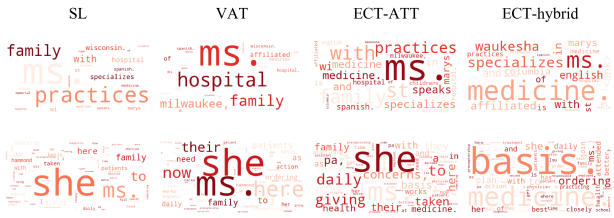

Figure 8: The wordcloud of two instances from different models. The attention of VAT mainly focus on gender words.

and color of words have a positive correlation with their attention scores. The proposed ECT-hybrid gets the best performance as well as the best explanations preventing gender biases via encouraging consistent reasons.

Conclusion

This paper considers the use of interpretability to facilitate semi-supervised learning (SSL). We disclose that the consistency assumption in SSL is closely related to causality invariance, where causality invariance works as the main reason for why the consistency assumption is valid. To this end, we propose ECT to encourage a consistent reason for model decisions under data perturbation. ECT employs model explanation as a surrogate of causality, and thus is able to bridge state-of-the-art interpretability to SSL models. Experimental results validate the highly competitive performance and better explanation of the proposed algorithms. Moreover, consistent explanations may help fight against overfitting and generalize much better on the biased dataset. In the future, we will study the feasibility and efficiency of optimization, especially on non-differentiable functions, for better explanation consistency.

Acknowledgments

This research was supported by the National Key R&D Program of China (2017YFB1001903) and the NSFC (61772262).

References

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, 7786–7795.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* 10: 1–46.

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Advances in Neural Information Processing Systems*, 5050–5060.

Cartwright, N. 2003. Two Theorems on Invariance and Causality. *Philosophy of Science* 70(1): 203–224.

De-Arteaga, M.; Romanov, A.; Wallach, H. M.; Chayes, J. T.; Borgs, C.; Chouldechova, A.; Geyik, S. C.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.

Etmann, C.; Lunz, S.; Maass, P.; and Schönlieb, C. 2019. On the Connection Between Adversarial Robustness and Saliency Map Interpretability. In *Proceedings of the 36th International Conference on Machine Learning*, 1823–1832.

Laine, S.; and Aila, T. 2016. Temporal Ensembling for Semi-Supervised Learning. *CoRR* abs/1610.02242.

LeCun, Y.; Cortes, C.; and Burges, C. J. 2010. The MNIST Database of Handwritten Digits. URL http://yann.lecun.com/exdb/mnist/. Last visited on 2021/03/04.

Lipton, Z. C. 2018. The Mythos of Model Interpretability. *Communications of the ACM* 61(10): 36–43.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, 4765–4774.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(8): 1979–1993.

Moraffah, R.; Karami, M.; Guo, R.; Raglin, A.; and Liu, H. 2020. Causal Interpretability for Machine Learning - Problems, Methods and Evaluation. *ACM SIGKDD Explorations Newsletter* 22(1): 18–33.

Nie, W.; Zhang, Y.; and Patel, A. 2018. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. In *Proceedings of the 35th International Conference on Machine Learning*, 3806–3815.

Pruthi, D.; Gupta, M.; Dhingra, B.; Neubig, G.; and Lipton, Z. C. 2020. Learning to Deceive with Attention-Based Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4782–4793.

Qi, G.; and Luo, J. 2019. Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods. *CoRR* abs/1903.11260.

Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; and Raiko, T. 2015. Semi-Supervised Learning with Ladder Networks. In *Advances in Neural Information Processing Systems*, 3546–3554.

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 International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Ross, A. S.; and Doshi-Velez, F. 2018. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. In *Proceedings of the 32rd AAAI Conference on Artificial Intelligence*, 1660–1669.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2662–2670.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision*, 618–626.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *2nd International Conference on Learning Representations*.

Spirtes, P. 2010. Introduction to Causal Inference. *Journal of Maching Learning Research* 11: 1643–1662.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. A. 2015. Striving for Simplicity: The All Convolutional Net. In *3rd International Conference on Learning Representations*.

Tarvainen, A.; and Valpola, H. 2017. Mean Teachers are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In *Advances in Neural Information Processing Systems*, 1195–1204.

Viviano, J. D.; Simpson, B.; Dutil, F.; Bengio, Y.; and Cohen, J. P. 2019. Underwhelming Generalization Improvements from Feature Attribution. *CoRR* abs/1910.00199.

Wang, Z.; Qin, Y.; Zhou, W.; Yan, J.; Ye, Q.; Neves, L.; Liu, Z.; and Ren, X. 2020. Learning from Explanations with Neural Execution Tree. In *8th International Conference on Learning Representations*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR* abs/1708.07747.

Zhou, T.; Wang, S.; and Bilmes, J. A. 2020. Time-Consistent Self-Supervision for Semi-Supervised Learning. *Proceedings of the 37th International Conference on Machine Learning* .

Zhu, X.; and Goldberg, A. B. 2009. Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3(1): 1–130.