A New Neural Network Training Procedure for Improved Interpretability

Anonymous Author(s)
Affiliation
Address
email

Abstract

Saliency methods have been widely used to highlight important input features in model predictions. Most existing methods use backpropagation on a modified gradient function, and thus noisy gradients can result in unfaithful feature attributions. In this paper, we tackle this issue and introduce an *interpretable training* procedure for neural networks to reduce noisy gradients used in predictions while retaining the predictive performance of the model. Our interpretable training procedure iteratively masks features with small and potentially noisy gradients while maximizing the similarity of the model outputs for both masked and unmasked inputs. We apply the interpretable training procedure to various synthetic and real data sets from computer vision, natural language processing, and time series across diverse neural architectures, including Recurrent Neural Networks, Convolutional Networks, and Transformers. We show that using our proposed interpretable training procedure significantly improves model interpretability across various domains while preserving predictive performance through qualitative and quantitative evaluations.

1 Introduction

Deep Neural Networks (DNNs) have been widely applied to a variety of different tasks [30, 25, 40, 35]; yet interpreting complex networks remains a challenge. Reliable explanations are necessary in critical domains like medicine, neuroscience, finance, and autonomous driving [9, 32]. Explanations are also useful for model debugging [56, 33]. As a result, various interpretability methods were developed to understand DNNs [5, 46, 21, 49, 48, 31]. A common approach for understanding model decisions is to identify features in the input that influenced the final classification decision [6, 56, 49, 48, 44, 34, 57]. Such approaches, known as *saliency maps*, often use gradients calculations to assign an *importance* score to individual features, reflecting their influence on the model prediction.

Saliency methods aim to highlight meaningful input features in model predictions to humans; however, the maps produced are often noisy. To improve the faithfulness of the saliency maps, explanations methods that depend on more than one or higher-order gradient calculations were developed. For example, SmoothGrad [48] reduces saliency noise by adding noise to the input multiple times and then taking the average of the resulting saliency maps for each input. Integrated gradients [49], DeepLIFT [44] and Layer-wise Relevance Propagation [5] backpropagate through a modified gradient function [3] while Singla et al. [47] studies the use of higher-order gradients in saliency maps.

In this paper, we take a different approach to improve the interpretability of deep neural networks—instead of developing yet another saliency method, we propose a new *training procedure* that naturally leads to improved model explanations using the current saliency methods. Our proposed training procedure, called *interpretable training*, trains models that produce sparse, meaningful, and less noisy gradients without degrading model performance. This is done by iteratively masking input features with low gradient values (i.e., less important features) and then minimizing a loss function

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

that combines (a) the KL divergence [26] between model outputs from the original and masked inputs, and (b) the appropriate loss function for model prediction. This procedure reduces noise in model gradients without sacrificing prediction performance.

To demonstrate the effectiveness of our proposed interpretable training approach, we consider a variety of classification tasks for images, language, and multivariate time series across diverse neural architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Network (RNNs), and Transformers. In particular, we observe that using interpretable training in image classification tasks leads to a reduction in visual saliency noise and sparser saliency maps, as shown in Figure 2. Interpretable training also improves the comprehensiveness of the produced explanations for sentiment analysis, and fact extraction tasks as shown in Table 1. In multivariate time series classification tasks, we observe up to 40% increase in the precision and recall of saliency maps on a synthetic benchmark when applying interpretable training. Interestingly, we also find that interpretable training reduces the vanishing saliency issue of RNNs [19] as shown in Figure 6. Finally, we note that although we use the vanilla gradient for masking in the interpretable training procedure, we observe significant improvements in the explanations produced after training by several gradient-based saliency methods.

2 Background and Related Work

Interpretability is a rapidly-growing area with several diverse lines of research. One strand of interpretability considers post-hoc explanation methods, aiming to explain why a trained model made a specific prediction for a given input. Post-hoc explanation methods can be divided into gradient-based methods [6, 49, 48, 44, 34, 43] that can be reformulated as computing backpropagation for a modified gradient function and perturbation-based approaches [56, 50, 39, 52] that perturb areas of the input and measure how much this changes the model output. Another line of works aims to measure the reliability of interpretability methods. This can be done by creating standardize benchmarks with interpretability metrics [18, 20, 10, 15, 42, 38] or debugging explanations [1, 23, 14, 2] by identifying test cases where explanations fail. Others [4, 11, 41, 54, 19] focus on modifying neural architectures for better interpretability. Similar to our line of work, Ghaeini et al. [13] and Ross et al. [41] incorporate explanations into the learning process. However, Ghaeini et al. [13] relies on the existence of the ground truth explanations while Ross et al. [41] relies on the availability of annotations about incorrect explanations for a particular input. Our proposed learning approach does not rely on such annotations; since most datasets only have ground truth labels, it may not be practical to assume the availability of positive or negative explanations.

In this paper, we evaluate our learning procedure with the following saliency methods: **Gradient (GRAD)** [6] is the gradient of the output w.r.t the input. **Integrated Gradients (IG)** [49] calculates a path integral of the model gradient to the input from a non-informative reference point. **DeepLIFT (DL)** [44] compares the activation of each neuron to a reference activation; the relevance is the difference between the two activations. **SmoothGrad (SG)** [48] samples similar input by adding noise to the input, and then takes the average of the resulting sensitivity maps for each sample. **Gradient SHAP (GS)** [34] adds noise to the input, then selects a point along the path between a reference point and input, and computes the gradient of outputs w.r.t those points.

We demonstrate the effectiveness of our training procedure using several neural network architectures: Convolution neural networks (CNNs) including Vgg [45], ResNet [16] and Temporal Convolutional Network (TCN) [36, 27, 7], a CNN that handles sequences; Recurrent neural networks (RNNs) including LSTM [17] and LSTM with Input-Cell Attention [19]; as well as Transformers [53].

3 Notation

First, consider a classification problem on the input data $\{(X_i, y_i)\}_{i=1}^{n}$ such that each $X = [x_1, \ldots, x_N] \in \mathbb{R}^N$ has N features and y is label. Let f_θ denote a neural network parameterized by θ. The standard training of the network involves minimizing the cross-entropy loss $\mathcal{L}$ over the training set as follows:

$$\underset{\theta}{\text{minimize}} \, \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(f_\theta\left(X_i\right), y_i\right)$$

The gradient of the network output $f_\theta(X)$ with respect to the input X is given by $\nabla_X f_\theta(X)$. Let $S(.)$ be a sorting function such that $S_e(Z)$ is the eth smallest element in Z. Hence, $S(\nabla_X f_\theta(X))$ is the sorted gradient. We define the input mask function $M_k(.)$ such that $M_k(S(X), X)$ replaces all x_i where $S(x_i) \in \{S_e(x_i)\}_{e=0}^{k}$ with a mask distribution, i.e. $M_k(S(X), X)$ removes the k lowest features from X based on the order provided by $S(X)$.

For a language input, $X = [x_1, \ldots, x_N]$ where $x_i \in \mathbb{R}^d$ is the feature embedding representing the ith word of the input. In that case, $S(X)$ would sort elements of X based on the sum of feature embeddings of each word x and $M_k(S(X), X)$ would mask the bottom k words according to that sorting. For a multivariate time series input, $X = [x_{1,1}, \ldots, x_{F,1}, \ldots, x_{F,T}] \in \mathbb{R}^{F \times T}$ where T is the number of time steps and F is the number of features per time step. $x_{i,t}$ is the input feature i at time t; sorting and masking would be done at the $x_{i,t}$ level.

For two discrete probability distributions P and Q defined on the same probability space $\mathcal{X}$, the Kullback–Leibler (KL) divergence [26] (or, relative entropy) from Q to P is given as D_{KL}:

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right). \tag{1}$$

4 Interpretable Training

Existing gradient-based methods produce noisy saliency maps as shown in Figure 1. The saliency map noise may be partially due to some uninformative local variations in partial derivatives. Using a standard training procedure based on ERM (expectation risk minimization), the gradient of the model w.r.t. the input (i.e. $\nabla_X f_\theta(X)$) may fluctuate sharply via small input perturbations [48].

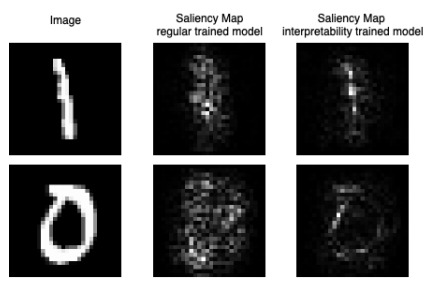

Figure 1: Saliency maps produced by typical training versus our proposed interpretable training.

If gradient-based explanation methods faithfully interpret the model's predictions, irrelevant features should have gradient values close to zero. Building on this intuition, we introduce *interpretable training*, a procedure to train neural networks such that input gradients computed from trained models provide more faithful measures to downstream (gradient-based) saliency methods. Interpretable training aims to reduce gradient values of irrelevant features without sacrificing model performance. During interpretable training, for every input X, we create a new input $\widetilde{X}$ by masking the features with low gradient values as follows:

$$\widetilde{X} = M_k(S(\nabla_X f_\theta(X)), X)$$

$\widetilde{X}$ is then passed through the network which results in an output $f_\theta\left(\widetilde{X}\right)$. In addition to the classification loss, the interpretable training minimizes the KL divergence between $f_\theta(X)$ and $f_\theta\left(\widetilde{X}\right)$ to ensure the trained model produces similar output probability distributions over labels from both masked and unmasked inputs. The optimization problem of our proposed interpretable training is:

$$\underset{\theta}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \left[\mathcal{L}\Big(f_\theta(X_i), y_i \Big) + \lambda D_{KL}\Big(f_\theta(X_i) \parallel f_\theta\left(\widetilde{X_i}\right) \Big) \right]$$

where λ is a hyperparameter to balance between cross-entropy classification loss and KL divergence terms. Since this loss function is differentiable with respect to θ, it can be optimized using existing gradient-based optimization methods. The KL divergence term encourages the model to produce similar outputs for the original input X and masked input $\widetilde{X}$. For this to happen, the model will need to learn to assign low gradient values to irrelevant features in model predictions. This potentially results in sparse and more faithful gradients as shown in Figure 1.

Masking functions. In images and time series, features with low gradients are replaced with random values within the feature range. In language, the masking function replaces the low salient word with the previous high salient word. The selection of k is dataset-dependent. For example, since most pixels in MNIST are uninformative, a larger k is desired. Detailed hyperparameters used is available in the supplementary.

Limitations (a) Although we use on vanilla gradients during training to mask unimportant features which is not computationally expensive, this training procedure (similar to other training procedures like adversarial training as an example) is more expensive than traditional training. (b) Our training procedure requires two hyperparameters k and λ which might require hyperparameter search (we find that $\lambda = 1$ works well in all our experiments).

Algorithm 1: Interpretable Training

Given: Training samples X, # of features to be masked k, learning rate τ, hyperparameter λ

Initialize f_θ

for $i \leftarrow 1$ **to** *epochs* **do**

  for *minibatch* $B \subset X$ **do**

    Compute the masked input:

    Get sorted index I for the gradient of output with respect to the input.

    $I = S\Big(\nabla_X f_{\theta_i}(X)\Big)$

    Mask bottom k features of the original input.

    $\widetilde{X} = M_k(I, X)$

    Compute the loss function:

    $L_i = \mathcal{L}\Big(f_{\theta_i}(X), y\Big) + \lambda D_{KL}\Big(f_{\theta_i}(X) \parallel f_{\theta_i}\Big(\widetilde{X}\Big)\Big)$

    Use the gradient to update network parameters:

    $f_{\theta_{i+1}} = f_{\theta_i} - \tau \nabla_{\theta_i} L_i$

  end

end

5 Experiments

All experiments have been repeated 5 times; the results reported below are the mean of 5 runs. Hyperparameters used for each experiment, along with the standard error bars, and, computational resources are available in supplementary materials.

5.1 Interpretable Training for Images

In the following section, we compare gradient-based explanations produced by regular training versus interpretable training for MNIST [29] trained on a simple CNN [28], for CIFAR10 [24] trained on ResNet18 [16] and for BIRD [12] trained on Vgg [45]. Further details about the datasets and models are available in the supplementary material.

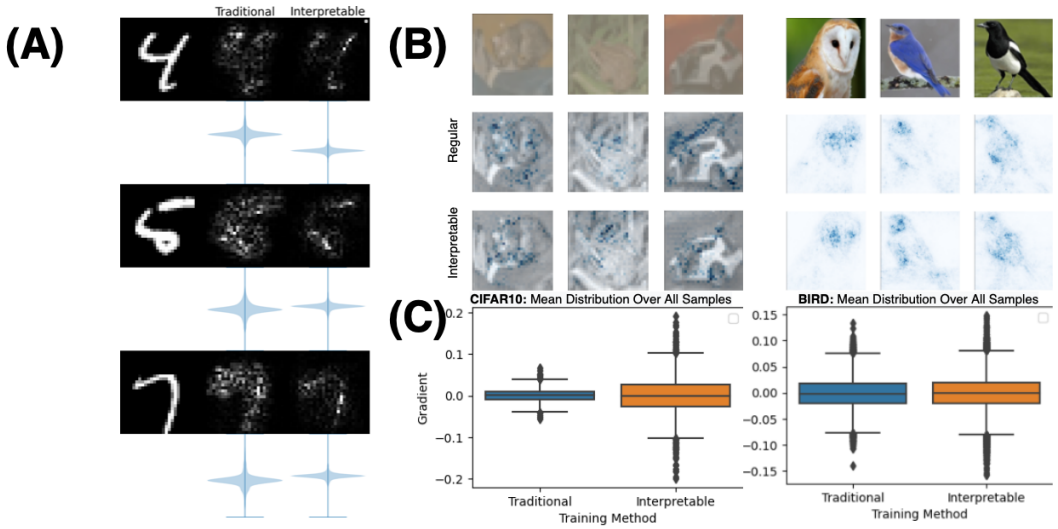

Figure 2: (A) Comparison between different training methods on MNIST along with distributions of gradient values in each sample. (B) Saliency maps for CIFAR10 and BIRD datasets using regular and interpretable training. (C) Distribution of gradient means across examples. Maps produced by interpretable training are more precise: most features have gradient values around zero with large gaps between mean and outliers, indicating the model's ability to differentiate between informative and non-informative features.

Saliency Map Quality for Images

For an image classification problem, in many cases, most features are redundant and not needed by the model to make the prediction. Consider the background of an object in an image; although it covers most of the image, backgrounds are often not totally essential in the classification task. We would expect the background gradient (i.e., most of the features) to be close to zero.

The examples shown in Figure 2 were correctly classified by both models. Gradients are scaled per sample to have values between -1 and 1. In Figure 2 (A) and Figure 2 (B), saliency maps produced by a model trained with interpretable training were more precise than that trained traditionally. Most saliency maps produced by interpretable training highlight the object itself rather than the background across different datasets. The distributions of gradient values per sample in Figure 2 (A) show that most features have small values (near zero) with higher separation of high saliency features away from zero for interpretable training. Similarly, in Figure 2 (C), we find that over the entire dataset, gradient values produced by interpretable training tend to be concentrated around zero with large separation between the mean and outliers (high saliency features), indicating the model's ability to differentiate between informative and non-informative features.

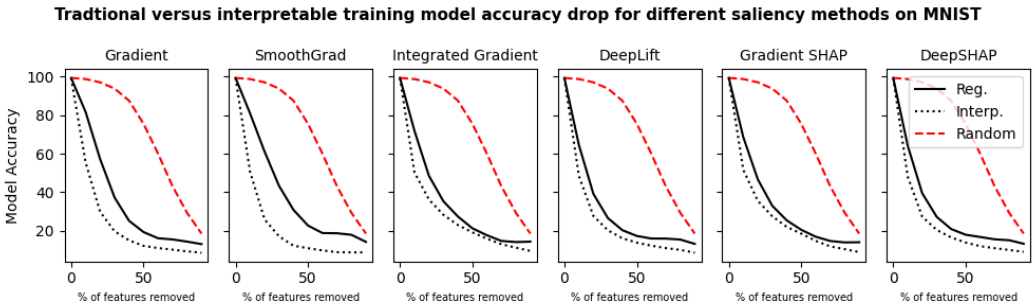

Figure 3: Model accuracy drop when removing features with high saliency using traditional and interpretable training for different gradient-based methods against a random baseline. A steeper drop indicates better performance. We find that regardless of the saliency method used, the performance improves by interpretable training.

Model Accuracy Drop

We compare interpretable and traditional training for different saliency methods with modification-based evaluation [42, 38, 22]: First, features are ranked according to the saliency values. Then, higher-ranked features are recursively eliminated (the original background in MNIST replaces the eliminated features). Finally, the degradation to the trained model accuracy is reported. This is done at different feature percentages. A steeper drop indicates that the removed features affected the model accuracy more. Figure 3 compares model degradation on different gradient-based methods; interpretable training shows a steeper accuracy drop regardless of the saliency method used. Even for simple methods like gradient, interpretable training shows around 80% accuracy drop while only 40% was achieved by traditional training.

This experiment can only be performed on MINST since the uninformative feature distribution is known (black background), while this is not the case in other datasets that we have considered. Although such modification-based evaluation methods have been applied to other datasets, [42, 38, 22]; Hooker et al. [18] showed that removing features produces samples from a different data distribution violating the underlying IID assumption (i.e. the training and evaluation data come from identical distributions). Hooker et al. [18] proposes ROAR where the model is retrained after feature elimination. However, due to the data redundancy, the retrained model can rely on different features to achieve the same accuracy. Figure 4, shows model accuracy drop on traditionally trained MINST when removing the same features. IID line represents replacing features with the black MNIST background (known uninformative distribution), which acts as the ground truth in this particular dataset. OOD line represents replacing the features with the mean image pixel value as done by [42, 38, 22]; and ROAR shows replacing features with the mean value and retraining the model as proposed by Hooker et al. [18]. Since neither OOD nor ROAR produce results similar to those produced by IID feature replacement, we argue that modification-based evaluation methods may provide unreliable results unless the uninformative IID distribution is known. We leave further exploration of modification-based evaluation methods to future work.

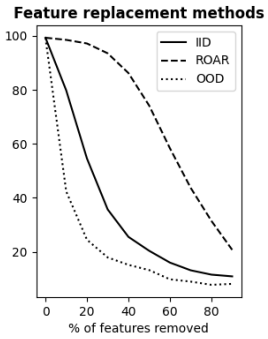

Figure 4: Accuracy drop in different modification-based evaluation masking approaches.

5.2 Interpretable Training for Language

We compare the interpretability of recurrent models trained on language tasks using the ERASER [10] benchmark. ERASER was designed to capture how well an explanation provided by models aligns with human rationales and how faithful these explanations are (i.e., the degree to which explanation influence the predictions). For our purpose, we only focus on the faithfulness of the explanations.

ERASER provides two metrics to measure interpretability. *Comprehensiveness* evaluates if all features needed to make a prediction are selected. To calculate an explanation comprehensiveness, a new input $\widetilde{X}_i$ is created such that $\widetilde{X}_i = X_i - R_i$ where R_i is predicted rationales. Let $f_\theta\left(X_i\right)_j$ be the prediction of model for class j. The model comprehensiveness is calculated as:

$$\text{Comprehensiveness} = f_\theta\left(X_i\right)_j - f_\theta\left(\widetilde{X}_i\right)_j$$

A high score here implies that the explanation removed was influential in the predictions. The second metric is *Sufficiency* that evaluates if the extracted explanations contain enough signal to make a prediction. The following equation gives the explanation sufficiency:

$$\text{Sufficiency} = f_\theta\left(X_i\right)_j - f_\theta\left(R_i\right)_j$$

A lower score implies that the explanations are adequate for a model prediction. The comprehensiveness and sufficiency were calculated at different percentages of features (similar to [10] percentages are $1\%, 5\%, 10\%, 20\%$ and 50%), and Area Over the Perturbation Curve (AOPC) is reported.

We focus on datasets that can be formulated as a classification problem: *Movie Reviews:* [55] positive/negative sentiment classification for movie reviews. *FEVER:* [51] a fact extraction and

verification dataset where the goal is verifying claims from textual sources; each claim can either be supported or refuted. *e-SNLI:* [8] a natural language inference task where sentence pairs are labeled as entailment, contradiction, neutral and, supporting.

Word embeddings are generated from Glove [37]; then passed to a bidirectional LSTM [17] for classification. Table 1 compares the scores produced by gradient for traditional and interpretable training against random assignment baseline. We found that a significant improvement in both comprehensiveness and sufficiently for sentiment analysis task *Movie Reviews* dataset. While for fact extraction task *FEVER* dataset, and natural language inference task *e-SNLI* dataset interpretable training improves comprehensiveness and there is no obvious improvement in sufficiently.

	Acc.	Comprehensiveness ↑	Sufficiency ↓
Movies			
Glove+LSTM Gradient	0.890	0.200	0.042
Glove+LSTM Interpretable + Gradient	0.885	**0.240**	**0.013**
Glove+LSTM Random	0.890	0.056	0.294
FEVER			
Glove+LSTM Gradient	0.721	0.007	0.012
Glove+LSTM Interpretable + Gradient	0.727	**0.008**	0.011
Glove+LSTM Random	0.721	0.001	**0.003**
e-SNLI			
Glove+LSTM Gradient	0.904	0.117	0.420
Glove+LSTM Interpretable + Gradient	0.904	**0.126**	0.387
Glove+LSTM Random	0.904	0.058	**0.366**

Table 1: Eraser benchmark scores: *Acc.* is model accuracy. *Comprehensiveness* and *sufficiency* are in terms of AOPC. 'Random' is a baseline when words are assigned random scores.

5.3 Interpretable Training for Time Series

We evaluated interpretable training on a multivariate time series, both quality on multivariate time series MNIST and quantitatively through synthetic data.

Saliency Maps Quality for Multivariate Time Series

We compare the saliency maps produced on MNIST treated as a multivariate time series where one image axis is time. Figure 5 shows the saliency maps produced by different *(neural architecture, saliency method)* pairs when different training procedures were used. There is a visible improvement in saliency quality across different networks when interpretable training is used.

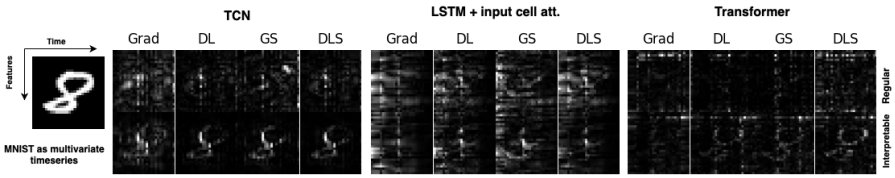

Figure 5: Saliency maps produced for *(neural architecture, saliency method)* pairs. Traditional training was used for networks in the 1st row, while interpretable training was used for the 2nd row. There is an improvement in the quality of saliency maps when interpretable training is used.

5.3.1 Quantitative Analysis on Synthetic Data

We evaluated interpretable training on a multivariate time series benchmark proposed by Ismail et al. [20]. The benchmark consists of 10 synthetic datasets, each examining different design aspects in typical time series datasets. Informative features are highlighted by the addition of a constant μ to the positive class and subtraction of μ from the negative class. Following Ismail et al. [20], we compare 4 neural architectures: LSTM [17], LSTM with Input-Cell Attention [19], Temporal Convolutional Network (TCN) [27] and, Transformers [53]. Additional details about the dataset and architectures are provided in the supplementary material.

Quantitatively measuring the interpretability of a *(neural architecture, saliency method)* pair involves applying the saliency method, ranking features according to the saliency values, replacing high salient features with uninformative features from the original distribution at different percentages. Finally, we measure the model accuracy drop, weighted precision, and recall.

Overall we observes that applying interpretable training significantly improves the results across all metrics for different *(neural architecture, saliency method)* pairs. Detailed results for each dataset is available in the supplementary material.

Interpretable training reduces vanishing saliency of recurrent neural networks

Ismail et al. [19] showed that saliency maps in RNNs vanish over time, biasing detection of salient features only to later time steps. This section investigates if adding the gradient to the loss function reduces the vanishing saliency issue in RNNs. Repeating experiments done by Ismail et al. [19], three synthetic datasets were generated as shown Figure 6 (A). The specific features and the time intervals (boxes) on which they are considered important are varied between datasets to test the model's ability to capture importance at different time intervals. We trained an LSTM with traditional and interpretable training procedures. the precision curve (AUP) and the area under the recall curve (AUR). The curves are calculated by the precision/recall values at different levels of degradation. Higher AUP and AUR suggest better performance. Results are shown in Figure 6 (B).

A traditionally trained LSTM shows clear bias in detecting features in the later time steps; AUP and AUR increase as informative features move to later time steps. When interpretable training is used, LSTM was able to identify informative features regardless of their locations in time.

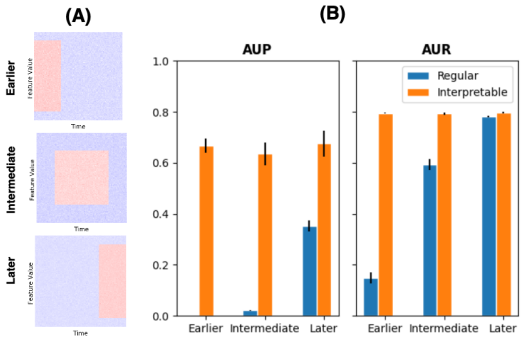

Figure 6: (A) Samples from 3 different simulated datasets informative features are located at the earlier, intermediate, and later time steps. (B) AUP and AUR were produced by LSTM by traditional and interpretable training procedures. Traditionally trained LSTM shows clear bias in detecting features in the later time steps. When interpretable training is used, there is no time bias.

6 Summary and Conclusion

We propose *interpretable training* as a new training procedure that improves the quality of explanations produced by existing gradient-based saliency methods. *Interpretable training* is optimized to reduce gradient values for irrelevant features. This is done by masking input features with low gradients and then minimizing the KL divergence between outputs from the original and masked inputs along with the main loss function. We demonstrated the effectiveness of *Interpretable training* on images, language, and multivariate time series.

Our proposed training method encourages models to sharpen the gradient-based explanations they provide. It does this however without requiring explanations as input. It instead may be cast as a regularization procedure, where regularization is provided by feature sparsity driven by gradient-based feature attribution. This is an alternative approach to cases ground truth explanations are used to force the model to be *right for the right reasons* [41]. However, we found that training model explanations in an unsupervised fashion also improves model faithfulness. This opens an interesting avenue for other unsupervised, perhaps regularization-based, methods to improve the interpretability of prediction models.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

[2] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *International Conference on Learning Representations*, 2018.

[4] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, 2014.

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In *PLoS ONE*, 2015.

[6] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. How to explain individual classification decisions. In *Journal of Machine Learning Research*, 2010.

[7] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[8] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *arXiv preprint arXiv:1812.01193*, 2018.

[9] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *International conference on knowledge discovery and data mining*, 2015.

[10] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

[11] Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

[12] Gerry. 265 bird species, 2021. URL `https://www.kaggle.com/gpiosenka/100-bird-species/`.

[13] Reza Ghaeini, Xiaoli Z Fern, Hamed Shahbazi, and Prasad Tadepalli. Saliency learning: Teaching the model where to pay attention. *arXiv preprint arXiv:1902.08649*, 2019.

[14] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *AAAI Conference on Artificial Intelligence*, 2019.

[15] Daniel Harborne, Richard Tomsett, Supriyo Chakraborty, Prudhvi Gurram ARL, and Alun Preece. Sanity checks for saliency metrics.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural computation*, 1997.

[18] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

[19] Aya Abdelsalam Ismail, Mohamed Gunady, Luiz Pessoa, Hector Corrada Bravo, and Soheil Feizi. Input-cell attention reduces vanishing saliency of recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2019.

[20] Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *arXiv preprint arXiv:2010.13924*, 2020.

[21] Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*, 2016.

[22] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*, 2017.

[23] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 2019.

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[26] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 1951.

[27] Colin Lea, Michael Flynn, Rene Vidal, Austin Reiter, and Gregory Hager. Temporal convolutional networks for action segmentation and detection. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[29] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database, 2010.

[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[31] Alexander Levine, Sahil Singla, and Soheil Feizi. Certifiably robust interpretation in deep learning. *arXiv preprint arXiv:1905.12105*, 2019.

[32] Zachary C Lipton. The mythos of model interpretability. In *Queue*, 2018.

[33] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158, 2012.

[34] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[35] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. In *The New England journal of medicine*, 2016.

[36] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[38] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[40] Michael L Rich. Machine learning, automated suspicion algorithms, and the fourth amendment. In *University of Pennsylvania Law Review*, 2016.

[41] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[42] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 2016.

[43] 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 *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[44] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 2017.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, 2013.

[47] Sahil Singla, Eric Wallace, Shi Feng, and Soheil Feizi. Understanding impacts of high-order loss approximations and features in deep learning interpretation. In *International Conference on Machine Learning*, pages 5848–5856. PMLR, 2019.

[48] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smooth-grad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[49] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

[50] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

[51] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[52] Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series models. *arXiv preprint arXiv:2003.02821*, 2020.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[54] Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *AAAI Conference on Artificial Intelligence*, 2018.

[55] Omar Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 conference on Empirical methods in natural language processing*, pages 31–40, 2008.

[56] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 2014.

[57] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Checklist

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above, along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] **section 4**
   (c) Did you discuss any potential negative social impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] **In the supplemental material**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] **In the supplemental material**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] **In the supplemental material**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL?[N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]