Saliency Guided Adversarial Training for Learning Generalizable Features with Applications to Medical Imaging Classification System

Xin Li [1] Yao Qiang [1] Chengyin Li [1] Sijia Liu [2] Dongxiao Zhu [1]

Abstract

This work tackles a central machine learning problem of performance degradation on out-of-distribution (OOD) test sets. The problem is particularly salient in medical imaging based diagnosis system that appears to be accurate but fails when tested in new hospitals/datasets. Recent studies indicate the system might learn shortcut and non-relevant features instead of generalizable features, so-called 'good features'. We hypothesize that adversarial training can eliminate shortcut features whereas saliency guided training can filter out non-relevant features; both are nuisance features accounting for the performance degradation on OOD test sets. With that, we formulate a novel model training scheme for the deep neural network to learn good features for classification and/or detection tasks ensuring a consistent generalization performance on OOD test sets. The experimental results qualitatively and quantitatively demonstrate the superior performance of our method using the benchmark CXR image data sets on classification tasks.

1. Introduction

Learning good feature representation that generalizes well to Out-Of-Distribution (OOD) test sets is a central challenge in machine learning. Recently, Deep Neural Network (DNN) has demonstrated impressive performance in classification and objection detection tasks on Independent and Identically Distributed (IID) test sets (Li et al., 2020b). Model regularization techniques, e.g., those based on parameter sparsity and loss function smoothing, used in conjunction with adversarial training, have been proven effective on mitigating *robust overfitting* (Rice et al., 2020) on IID test sets. Nevertheless, the performance degradation on OOD test sets remains a salient problem (Shao et al., 2020). One observation is that the current approach introduces a nearly ideal scenario for DNN to learn spurious shortcuts or non-relevant features (Geirhos et al., 2020) that do not exist in OOD test sets. In medical imaging systems, the problem becomes even more salient due to the significant distribution shift between imaging data sets acquired from different hospitals, populations, and time periods. As a result, the AI imaging system that is seemingly effective on training sets often does not generalize well to new hospitals or data sets (DeGrave et al., 2021). Fortunately, in the relatively closed medical imaging environment, we are not so much concerned about adversarial OOD test sets. Instead, we consider how to leverage adversarial IID data sets for learning good features.

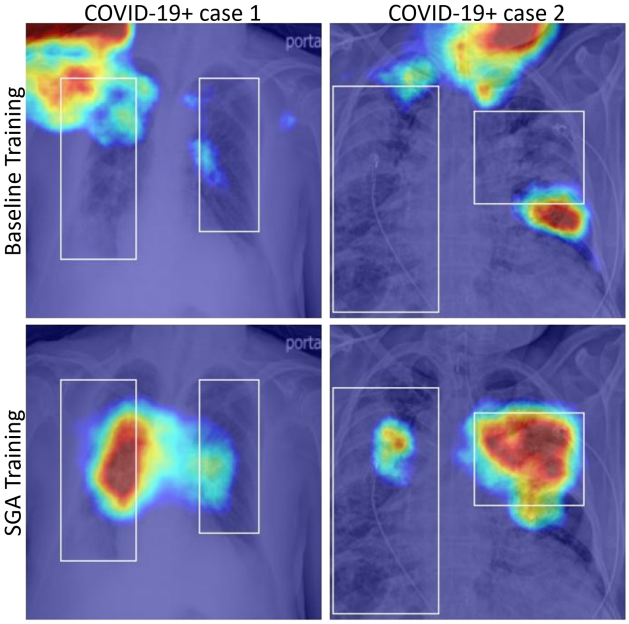

Figure 1. Motivation examples to illustrate the shortcut features (top left) and non-relevant features (top right). Good features are highlighted with high salience in the second rows, overlapping with radiologists' annotations. The heatmap based DNN interpretations are generated by FullGrad.

Here we give two motivation examples to illustrate the above-mentioned problem using salience based model ex-

[1] Wayne State University, Detroit, USA [2] Michigan State University, East Lansing, USA. Correspondence to: Dongxiao Zhu <dzhu@wayne.edu>.

AdvML Frontiers workshop at 39th International Conference on Machine Learning, Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

planation methods. Salience methods are a main body of explainable machine learning approaches that quantify individual attribution of input features to the output. Exemplar methods include Integrated Gradient (Sundararajan et al., 2017; Pan et al., 2021), Grad-CAM (Selvaraju et al., 2017), Salience Map (Simonyan et al., 2013) and FullGrad (Srinivas & Fleuret, 2019). In Figure 1, the left example shows the shortcut features and the right example shows non-relevant features are used to predict COVID positive cases. Both types of features can harm the generalizability of the DNN models to OOD test sets. The bounding box represents radiologists' annotated pathological features also known as symptom reasoning (Yang et al., 2019), which align well with the high saliency regions highlighted by our saliency guided adversarial training (SGA) scheme, but not the baseline cross-entropy based training.

How can we develop an effective robust training scheme to learn the good features for generalizing to test sets? There are four types of test cases, i.e., IID, Adversarial IID, OOD, and Adversarial OOD. Adversarial test cases are rare in the medical imaging system since it is a relatively closed environment that takes pre-processed clean imaging inputs. Thus a major challenge is that the AI system tends to learn and exploit non-relevant features and/or shortcut features, as opposed to generalizable features from training, leading to downgraded performance on OOD test sets.

Recent studies (Maguolo & Nanni, 2021; Cohen et al., 2020a) demonstrate that CXR classification systems might depend more on nuisance features generated by different medical devices with various manufacturing standards and acquisition parameters. Similar to adversarial perturbation, those nuisance features do not impede human recognition but is obvious to DNN models, particularly when they lay on extremely clean background around the CXR borders (Li & Zhu, 2020). As shown by case 1 (top left in Figure 1), model using those shortcut features would have a poor generalization on OOD test sets. To enhance the OOD generalization, Yi et al. (2021) proves that model trained robust to adversarial perturbation generalizes well on OOD data. Base on their work, we further hypothesize that adversarial training (Madry et al., 2017) can eliminate those shortcut features since adversarial perturbation are also imperceptible and usually considered as the worst case noise. On the other hand, Ismail et al. (2021) demonstrate a saliency guided training encourage the model to learn and assign low gradient values to non-relevant features in model predictions, resulting in a more faithful learning of the intended features. Both have been developed to improve feature learning for better generalizability. Here we propose a novel saliency guided adversarial training for better feature representation learning. The saliency guided component eliminates the non-relevant features by reducing their gradient values, whereas adversarial training enhances the robustness of model against learning shortcut features by adding noise to the most relevant features. Using CXR based experiments, we demonstrate that our SGA training scheme learns generalizable features for improving the test performance on the OOD CXR data sets.

2. Related Work

To enhance model robustness against adversarial ID and OOD examples, various robust training techniques have been proposed, including those training with augmented adversarial examples, aka, adversarial training (Madry et al., 2017), robust regularization (Tack et al., 2021; Chen et al., 2019; Boopathy et al., 2020), and improved loss functions, e.g., (Li et al., 2020b). Here we focus on the related works in robust training.

2.1. Robust training for IID test cases

One line of approaches (Kurakin et al., 2016; Sinha et al., 2017; Zhang et al., 2019; Shafahi et al., 2019) are based on adversarial training (Goodfellow et al., 2014) and achieve effective robustness against different adversarial attacks, where the training dataset is augmented with adversarial examples. Adversarial training has also been shown effective in learning robust features for enhanced robustness (Ilyas et al., 2019; Madry et al., 2017). However, these methods have trade-offs between accuracy and adversarial robustness (Tsipras et al., 2018) and are computationally expensive in adversarial sample generation (Zhang et al., 2019). To reduce the computational burden, Shafahi et al. (Shafahi et al., 2019) propose a training algorithm, which improves the efficiency of adversarial training by updating both model parameters and image perturbation in one backward pass. (Wong et al., 2020) discover that it is possible to train empirically robust models using a much weaker and cheaper FGSM based adversary training combined with random initialization.

Another line of defending strategy against adversaries, other than augmenting the training dataset, is to learn robust feature representations by using model ensembles or altering network architectures (Taghanaki et al., 2019; Mustafa et al., 2019; Tramèr et al., 2017; Liao et al., 2018; Pang et al., 2019; Xu et al., 2017; Meng & Chen, 2017). For example, (Taghanaki et al., 2019) augment DNNs with the radial basis function kernel to further transform features via kernel trick to improve the class separability in feature space and reduce the effect of perturbation. (Mustafa et al., 2019) propose a prototype objective function, together with multi-level deep supervision. Their method ensures the separation in feature space between classes and shows significant improvement of robustness. (Pang et al., 2019) develop a strong ensemble defense strategy by introducing a new regularizer to encourage diversity among models within the ensemble system, which encourage the feature representation from the same class to be close. Although these approaches avoid the high computational cost of adversarial training, they have to modify the network architecture or require an extra training process, limiting the flexibility in adapting to different tasks.

2.2. Robust training for OOD test cases

When trained on IID examples, DNNs are known to fail against test inputs that lie far away from training distribution, commonly referred to as OOD examples (Hendrycks & Gimpel, 2016). Recent robust training for detecting ODD test cases considers a multi-class dataset as IID (e.g., CIFAR-10) and uses examples from another multi-class dataset as OOD (CIFAR-100) (Liang et al., 2017; Hendrycks & Gimpel, 2016; Wei et al., 2020; Lee et al., 2017). Existing works either train an OOD detector and a classifier sequentially (Sehwag et al., 2019; Li et al., 2020c) or simultaneously (Anonymous, 2021). For example, (Sehwag et al., 2019) employ adversarial training on IID data as well as OOD examples that are close to IID examples to improve learning robust features. These approaches work well for the so-called closed-world detection where OOD examples are either with simpler data modalities (e.g., medical images with large shared backgrounds) or closer to IID examples (CIFAR-10 versus CIFAR-100). Different from IID detection tasks where robust discriminative features are learned from labeled training data, OOD detection needs to learn *high-level*, *task-agnostic* and *semantic* features from the IID dataset to detect diverse OOD inputs at the test time.

More recent OOD detection approaches are self-supervised representation learning using only unlabeled training data, which involves two key steps: 1) learning a good (e.g., compact and semantic) feature representation, and 2) modeling features of ID data without requiring class labels. For example, (Winkens et al., 2020) used contrastive training techniques SimCLR (Chen et al., 2020b) to extract semantic features and proposed confusion log probability to determine whether a test example is a near or far OOD example. Using experiments, they show their approach is scalable to high-dimensional multimodal OOD examples. (Anonymous, 2021) also use contrastive loss based label-free training for self-supervised feature learning followed by OOD detection using Mahalanobis distance.

Another line of label-free feature learning approaches for OOD detection uses flow-based generative models (e.g., VAEs, PixelCNNs, and Glow(Kingma & Dhariwal, 2018), allowing for the exact formulation of the marginal likelihood, to learn task-agnostic and semantic features to address the OOD detection problem. However, even sophisticated neural generative models trained to estimate feature density distribution (e.g., on CIFAR-10 images) can perform poorly on OOD detection, often assigning higher probabilities to OOD test examples than to IID test examples (Nalisnick et al., 2018). Most recent research attempt to learn task-agnostic and semantic features for both IID and OOD images (Zhang et al., 2020; Shao et al., 2020; Nalisnick et al., 2019; Chen et al., 2020a), yet unique challenges exist in learning task-agnostic and semantic representations.

2.3. Saliency guided training for enhancing DNN interpretability

Saliency guided training has recently been shown to reduce noisy gradients used in predictions while retaining the predictive performance of the model. (Ismail et al., 2021) propose a saliency guided training by creating a new input by masking the features with low gradient values (salience) and encouraging the similarity between the new and original outputs. (Uddin et al., 2020) develope a new approach that mixes the patches and labels using the salience density to select patches to dropout as a model regularization. (Chen et al., 2019) propose training objectives in classic robust optimization models to achieve robust Integrated Gradient (IG) attributions and demonstrate comparable prediction robustness (sometimes even better) while consistently improving attribution robustness. With these existing works, the generalizable features are expected to be learned via designing and optimizing a new salience-guided adversarial training objective as described below.

3. Saliency Guided Adversarial Training

Considering a classification problem on the input data $\{(X_i, y_i)\}_{i=1}^{n}$, a deep neural network model f_θ parameterized by θ is trained to predict the target y. The standard training involves minimizing the cross-entropy loss $\mathcal{L}$ over the training set as follows:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_\theta(X_i), y_i). \tag{1}$$

The model parameter θ is updated via one step of gradient descent with the learning rate α:

$$\theta \leftarrow \theta - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \mathcal{L}(f_\theta(X_i), y_i), \tag{2}$$

on a mini-batch of m samples $\{(X_i, y_i)\}_{i=1}^{m}$. We denote the gradient of the model output $f_\theta(X)$ with respect to the input X as $\nabla_X f_\theta(X)$.

Since the standard training procedure is based on ERM (expectation risk minimization) using stochastic gradient descent (SGD), the gradient of model w.r.t. the input (i.e., $\nabla_X f_\theta(X)$) may fluctuate sharply via small input perturbations (Smilkov et al., 2017), e.g., adversarial noise. In this way, the model would probably learn some non-relevant features due to some uninformative local variations in partial derivatives. Furthermore, (Geirhos et al., 2020) observes that the traditional training approach introduces a nearly ideal scenario for DNN models to learn some spurious shortcut features, which do not exist in OOD test sets.

Building on these intuitions, we propose saliency guided adversarial (SGA) training, a novel procedure to train the neural network models to learn the good features by suppressing non-relevant and eliminating shortcut features.

During saliency guided adversarial training, we augment the training set by generating a new training sample for each input sample X by masking the features with low gradient values as follows:

$$\tilde{X} = M_k(X, S(\nabla_X f_\theta(X))), \tag{3}$$

where $S(\nabla)$ is a function that sorts the gradient of each feature from X in the ascending sequence. $M_k(X, S(\nabla))$ is an input mask function, which replaces the k lowest features from X with random values within the feature range based on the order provided by $S(\nabla)$, as the non-relevant features usually have gradient values close to zero. k is a tuning parameter, and its selection is based on the amount of nuisance information in a training sample. To further eliminate the shortcut features, we generate an adversarial example for the new sample $\tilde{X}$ as:

$$X' = \tilde{X} + \delta^\star, \tag{4}$$

where $\delta^\star$ is estimated as:

$$\delta^\star = \underset{|\delta|_p \leq \epsilon}{\arg\max} \mathcal{L}(f_\theta(\tilde{X} + \delta), y), \tag{5}$$

and p can be $0, 1, 2, \ldots$ and ∞. In most cases, the perturbation budget ϵ is small so that the perturbations are imperceptible to human eyes. In our case, the adversarial example X' is generated based on the masked input $\tilde{X}$. Thus, X' does not contain the non-relevant features but has some shortcut features compared to the clean input X.

X' is then passed through the model, resulting in an output $f_\theta(X')$. In addition to the classification loss used in the traditional training, saliency guided adversarial training adds another regularization term that minimizes the Kullback–Leibler (KL) divergence between $f_\theta(X)$ and $f_\theta(X')$. This regularization term ensures the model produces similar output probability distributions over labels for the original clean input X and the masked adversarial example X'. For this to happen, the model is ensured to learn the good features that ensure generalization performance on OOD test set.

The optimization problem for our saliency guided adversarial training is:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \left[\mathcal{L}(f_\theta(X_i), y_i) + \lambda D_{KL}(f_\theta(X_i) || f_\theta(X'_i)) \right], \tag{6}$$

where λ is a hyperparameter to leverage the importance of the cross-entropy classification loss and the KL divergence regularization term. Since this loss function is differentiable with respect to θ, it can be optimized using existing gradient-based optimization methods. We show the saliency guided adversarial training procedure in Algorithm 1.

Algorithm 1 Saliency Guided Adversarial Training

```
Require: Training Sample $X$, # of features to be masked $k$, attack order $p$, perturbation budget $\epsilon$, learning rate $\tau$, hyperparameter $\lambda$, initialized model $f_\theta$
for epochs do
  for minibaches do
    Create the masked input:
    1. Get sorted index $I$ for the gradient of output with respect to the input: $I = S(\nabla_X f_\theta(X))$
    2. Mask bottom $k$ features of the original input: $\tilde{X} = M_k(X, I)$
    Generate the adversarial example:
    1. Compute $\delta$: $\delta^\star = \underset{|\delta|_p \leq \epsilon}{\arg\max} \mathcal{L}(f_\theta(\tilde{X} + \delta), y)$,
    2. Generate the adversarial example: $X' = \tilde{X} + \delta^\star$
    Compute the loss:
    $\mathcal{L}_i = \mathcal{L}(f_{\theta_i}(X), y) + \lambda D_{KL}(f_{\theta_i}(X) || f_{\theta_i}(X'))$
    Update $\theta$:
    $\theta_{i+1} = \theta_i - \tau \nabla_{\theta_i} \mathcal{L}_i$
  end for
end for
Return: $f_\theta$
```

4. Experiments and Results

In this section, we first define IID and OOD test sets in medical imaging domain and then explain how we built dataset for the COVID-19 detection task. Our proposed training approach is then evaluated qualitatively and quantitatively. Finally, the ablation analysis is performed to assess the effect of hyper-parameters.

IID and OOD test sets In machine learning, it's typical to randomly divide the available data into a training/validation and test set, with the former being used to select and teach the model to perform a particular task, and the latter being used to check the model's performance. One common assumption is that those two datasets are drawn from the same distribution. In relation to the training dataset, this test set is then referred as IID data (Geirhos et al., 2020). Aside from the IID data, recent studies (Sehwag et al., 2019) evaluate the performance of AI systems on OOD data, which are systematically different from the IID data with a significant distribution shift. For example, in the medical domain, a test set acquired from different hospitals from the training dataset can be treated as OOD data (DeGrave et al., 2021).

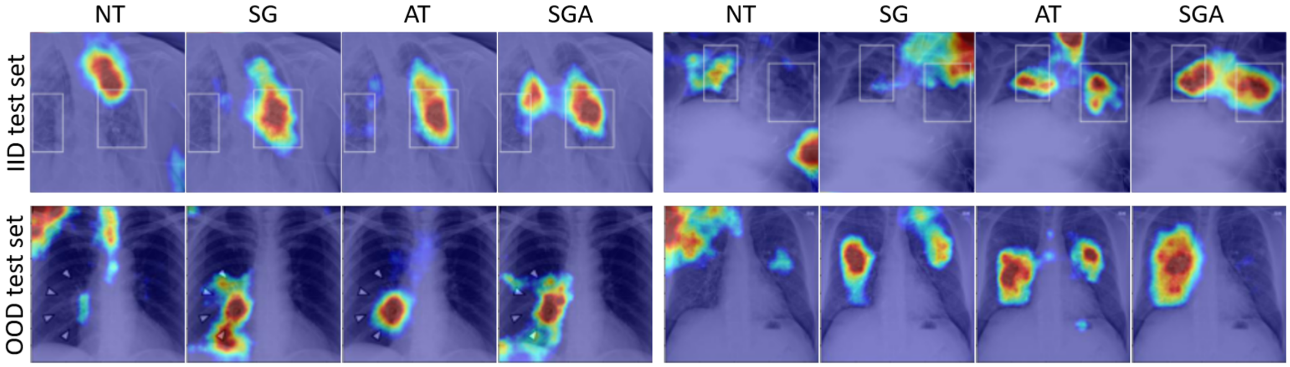

Figure 2. Examples to illustrate the features used by models to detect COVID-19+ in both IID and OOD test set. The bounding boxes in IID test set (top row) represents radiologists' annotated pathological features. Note we do not have similar annotations in OOD test set (bottom row).

COVID-19 dataset We select one benchmark data set and generate another dataset to evaluate our method for COVID-19 detection tasks. Dataset I is from SIIM-FISABIO-RSNA COVID-19 detection competition (Lakhani et al., 2021), which is used as the IID data set for training, validating, and internal testing. The dataset comprises 6334 CXR scans that are labeled by a panel of experienced radiologists with appearance and bounding box of COVID-19 opacities. The Dataset I is split into training, validation, and testing sets by a ratio of $6:2:2$.

Dataset II is used for external test (OOD) only, which consists of COVID-19-positive X-ray in the GitHub-COVID repository (Cohen et al., 2020b) collected from some public figures and other online sources with different geographic origins. Similar to (Li et al., 2020a), we further supplement these figures with COVID-19-negative ('No Findings') X-rays from the ChestX-ray14 dataset (Wang et al., 2017), which originate from a single hospital in the United States. It is important to note that the samples in Dataset I may contain COVID-19-negative CXRs from individuals with unknown pulmonary diseases, whereas the COVID-19-negative samples in Dataset II come from healthy individuals. As a result, the task of detecting positive COVID cases from the OOD test set can be less challenging than from the IID test set because the COVID negative cases in the former can be separated from COVID-19-positive cases more easily, giving rise to an enhanced performance (as opposed to degraded performance) in OOD test set using enhanced SGA training,

Experiment settings We use ResNet-18 (He et al., 2016) pre-trained with ImageNet as the DNN architecture. which is trained with the SGD optimizer for 30 epochs with a batch size of 64. The adversarial samples X' are generated by FGSM for each minibatch with a uniformly sampling perturbation from the interval $[0.01, 0.05]$ during the training process. The hyperparameter k and λ are fine-tuned as 0.1 and 1 respectively. The model that achieve the best performance on the validation set are used for IID and OOD testing. In order to demonstrate the effectiveness of our approach, we perform experiments comparing the performance of our SGA with three baseline methods: natural training (NT) with cross-entropy loss only, FGSM-based adversarial training (AT), and saliency guided training (SG). We show the heat map interpretations generated by FullGrad (Srinivas & Fleuret, 2019), which highlights the most salient regions of each CXR image that contribute mostly to the output, to illustrate the features exploited by the pre-trained models for COVID-19 detection.

Qualitative evaluation Figure 2 shows the heat maps generated from the models trained with four competing training methods on two examples from IID and OOD test sets, respectively. NT generates the worst heat map interpretations on IID test set (top row) since the models seemly just learn some *non-relevant* features (e.g., those corresponding to the backbone shown in the left IID/NT panel) and/or some *shortcut* features (e.g., some special tags lying on the borders shown in the right IID/NT panel). This problem is aggravated on the OOD test set shown in the panels of the bottom row. Although SG and AT achieve slightly better interpretations than NT, SG still can not eliminate the *shortcut* features and AT seems to be plagued by some *non-relevant* features. On the contrary, the models trained with our SGA trend to use the COVID-19 pathological features (within the lungs inside the annotated bounding boxes) to detect COVID-19 in both IID and OOD test sets. This figure indicates that the regularization term added to our SGA training objective successfully learns good features by preventing the model from extracting *shortcut* and *non-relevant* features.

Quantitative evaluation In addition to the qualitative examples presented above, we also conduct quantitative experiments to validate our SGA method and compare with three baselines using the area under the curve (AUROC) for the imbalanced binary classification task. Figure 3 demonstrates that SGA has the best average performance (0.81) on both IID and OOD test sets compared to other baselines: NT (0.78), SG (0.79), AT (0.79). A more important consideration is the performance drop from IID to OOD test set, which indicates whether the model uses *shortcut* and/or *non-relevant* features to make predictions. It is striking to note that the performances of AT and SGA increase from IID to OOD test set, which indicates (1) both training schemes learn and leverage good features and (2) the COVID prediction task on the OOD test set (Dataset II) is less challenging since the COVID-19-negative cases are free of lung diseases and thus comparatively more contrasting. The model trained with NT has a significant performance drop of $8.5\%(0.82 \rightarrow 0.75, 0.07/0.82)$, indicating that the model trained with NT uses shortcut and/or non-relevant features to make the prediction. This is consistent with what we demonstrate in the qualitative evaluation section.

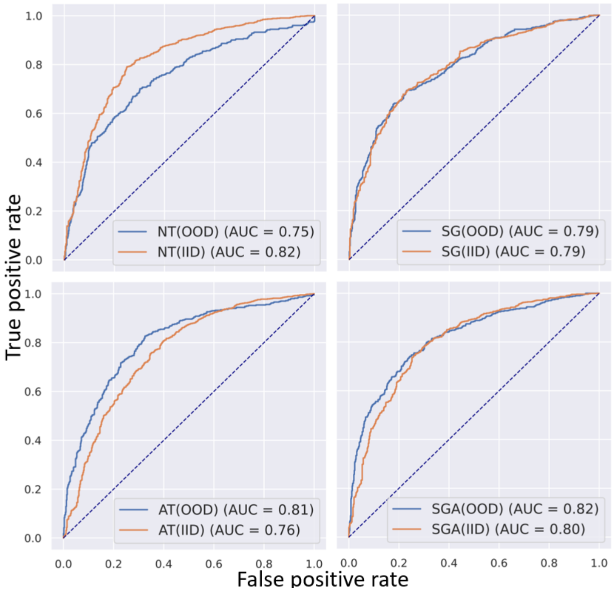

Figure 3. Model evaluations with receiver operating characteristic (ROC) curves, which show the performances on both internal test set (IID) and an external test set (OOD). The difference between IID and OOD test set performance is the performance degradation. Only the model trained with NT has a performance drop of $8.5\%(0.07/0.82)$ between IID and OOD test sets.

Hyper-parameter k As previously indicated, assuming gradient-based explanation approaches interpret the model's predictions accurately, non-relevant features should have small gradient values. Based on this insight, we remove the k lowest features from the input picture X so as to encourage the model to learn good features. Note that the gradient value generated by saliency Map can be negative. Higher negative values indicate that its absence contributes to an increased score of the class. These regions might be other objects in the background that cause the model to make incorrect predictions, and masking those region enables the model to concentrate on the foreground. As a result, in contrast to the original paper (Simonyan et al., 2013), which used absolute gradient values to demonstrate the significance of features, we directly ordered the features and eliminated the lowest k of features during training.

k	0.00	0.05	0.10	0.15	0.20
IID Test	0.76	0.81	0.80	0.80	0.81
OOD Test	0.81	0.78	0.82	0.79	0.80
Difference	+0.05	-0.03	+0.02	-0.01	-0.01
Average	0.79	0.80	**0.81**	0.79	0.80

Table 1. Ablation analysis of hyper-parameter k. Note that when $k = 0$, the model is trained by AT only.

λ	0	0.5	1	1.5	2
IID Test	0.82	0.78	0.80	0.78	0.78
OOD Test	0.75	0.81	0.82	0.81	0.82
Difference	-0.07	+0.03	+0.02	+0.03	+0.04
Average	0.79	0.79	**0.81**	0.79	0.80

Table 2. Ablation analysis of hyper-parameter λ. Note that when $\lambda = 0$, the model is trained by NT.

The selection of k depends on how much non-relevant information is in a training set. We chose small k for our COVID-19 detection experiment because CXR images have little and clear backgrounds. Table 1 show the result of ablation study on the hyper-parameter k, which is tuned from 0 to 0.2. Compared to AT, the model's performance increased by masking small amount of non-relevant features and achieve the best average performance on IID and OOD test sets when $k = 0.1$.

Hyper-parameter λ λ is used to balance the contributions of NT and SGA regularization in the training objective. As shown in Table 2, compared to NT, the models trained with SGA have significantly better performance on OOD test set.

5. Conclusions

Existing DNN training methods can exploit non-relevant and shortcut features for prediction, which may account for the performance degradation on test set, particularly on OOD test sets. To overcome this limitation, we propose a novel saliency-guided adversarial training scheme for learning good features and empirically demonstrate its strong performance on CXR based OOD test sets, opening a new avenue for tackling the failure of medical imaging system in new hospitals or on new test sets.

References

Anonymous. Informative outlier matters: Robustifying out-of-distribution detection using outlier mining. In *Submitted to International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dhQHk8ShEmF. under review.

Boopathy, A., Liu, S., Zhang, G., Liu, C., Chen, P.-Y., Chang, S., and Daniel, L. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*, pp. 1014–1023. PMLR, 2020.

Chen, J., Wu, X., Rastogi, V., Liang, Y., and Jha, S. Robust attribution regularization. *Advances in Neural Information Processing Systems*, 32, 2019.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. Informative outlier matters: Robustifying out-of-distribution detection using outlier mining, 2020a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020b.

Cohen, J. P., Hashir, M., Brooks, R., and Bertrand, H. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, pp. 136–155. PMLR, 2020a.

Cohen, J. P., Morrison, P., and Dao, L. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020b.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Ismail, A. A., Corrada Bravo, H., and Feizi, S. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34, 2021.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pp. 10215–10224, 2018.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Lakhani, P., Mongan, J., Singhal, C., Zhou, Q., Andriole, K. P., Auffermann, W. F., Prasanna, P., Pham, T., Peterson, M., Bergquist, P. J., et al. The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs. 2021.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Li, X. and Zhu, D. Robust detection of adversarial attacks on medical images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1154–1158. IEEE, 2020.

Li, X., Li, C., and Zhu, D. Covid-mobilexpert: On-device covid-19 patient triage and follow-up using chest x-rays. In *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 1063–1067. IEEE, 2020a.

Li, X., Li, X., Pan, D., and Zhu, D. Improving adversarial robustness via probabilistically compact loss with logit constraints. *arXiv preprint arXiv:2012.07688 accepted by AAAI-21*, 2020b.

Li, X., Pan, D., and Zhu, D. Defending against adversarial attacks on medical imaging ai system, classification or detection? *arXiv preprint arXiv:2006.13555*, 2020c.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Maguolo, G. and Nanni, L. A critic evaluation of methods for covid-19 automatic detection from x-ray images. *Information Fusion*, 2021.

Meng, D. and Chen, H. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 135–147, 2017.

Mustafa, A., Khan, S., Hayat, M., Goecke, R., Shen, J., and Shao, L. Adversarial defense by restricting the hidden space of deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3385–3394, 2019.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Hybrid models with deep and invertible features. *arXiv preprint arXiv:1902.02767*, 2019.

Pan, D., Li, X., and Zhu, D. Explaining deep neural network models with adversarial gradient integration. In *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. Improving adversarial robustness via promoting ensemble diversity. *arXiv preprint arXiv:1901.08846*, 2019.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Sehwag, V., Bhagoji, A. N., Song, L., Sitawarin, C., Cullina, D., Chiang, M., and Mittal, P. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 105–116, 2019.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pp. 3353–3364, 2019.

Shao, R., Perera, P., Yuen, P. C., and Patel, V. M. Open-set adversarial defense. *arXiv preprint arXiv:2009.00814*, 2020.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2, 2017.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Srinivas, S. and Fleuret, F. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Tack, J., Yu, S., Jeong, J., Kim, M., Hwang, S. J., and Shin, J. Consistency regularization for adversarial robustness. *arXiv preprint arXiv:2103.04623*, 2021.

Taghanaki, S. A., Abhishek, K., Azizi, S., and Hamarneh, G. A kernelized manifold mapping to diminish the effect of adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11340–11349, 2019.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Uddin, A., Monira, M., Shin, W., Chung, T., Bae, S.-H., et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*, 2020.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Wei, X., Wang, H., Scotney, B., and Wan, H. Minimum margin loss for deep face recognition. *Pattern Recognition*, 97:107012, 2020.

Winkens, J., Bunel, R., Roy, A. G., Stanforth, R., Natarajan, V., Ledsam, J. R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Yang, C.-H. H., Liu, Y.-C., Chen, P.-Y., Ma, X., and Tsai, Y.-C. J. When causal intervention meets adversarial examples and image masking for deep neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3811–3815. IEEE, 2019.

Yi, M., Hou, L., Sun, J., Shang, L., Jiang, X., Liu, Q., and Ma, Z. Improved ood generalization via adversarial training and pretraing. In *International Conference on Machine Learning*, pp. 11987–11997. PMLR, 2021.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, pp. 227–238, 2019.

Zhang, H., Li, A., Guo, J., and Guo, Y. Hybrid models for open set recognition. *arXiv preprint arXiv:2003.12506*, 2020.