PILLET-GAN: Pixel-Level Lesion Traversal Generative Adversarial Network for Pneumonia Localization

Editors: Under Review for MIDL 2021

Abstract

The study of pneumonia localization focuses on the problem of accurately visualizing lesions in the thoracic X-ray image. It is important to provide precisely localized regions to users in terms of that it can lay out a basis of the model decision by comparing the difference of X-ray image between 'Healthy' and 'Disease' class. In particular, for the medical image analysis, it is essential not only to make a correct prediction for the disease but also to provide evidence to support accurate predictions. To address this issue, many generative adversarial network (GAN) based approaches are employed to show the pixel-level changes via domain translation technique. Although the most previous research tried to improve the performance of localization through understanding the hidden effects for better domain translation, it still remains challenging to capture the pixel-level changes for the specific category. For this reason, we focus on the stage of domain translation for a better understanding of category attributes. In this paper, we propose a Pixel-Level Lesion Traversal Generative Adversarial Network (PILLET-GAN) that mines spatial features for the category via convolutional masking technique and fuse them into original feature map extracted from the generator for better domain translation. Then, these aggregated features are combined with the input image to generate clear localization results. Our experimental results show that PILLET-GAN achieves superior performance compared to the state-of-the-art models on both qualitative results and quantitative results on the RSNA-Pneumonia dataset.

Keywords: Medical image analysis, Generative adversarial network, Pneumonia localization, Computer-aided system

1. Introduction

With the development of deep learning techniques, the demand for explainable artificial intelligence (XAI) study has risen above the surface of the water in Computer Vision (CV) and Natural Language Processing (NLP) communities. In particular, the XAI system is essential in the medical domain. Since the doctor, who is a user of the system, does not rely entirely on the predictions of the deep learning model but uses it as an auxiliary tool, it should be able to present not only the prediction results but also the reasonable basis for the model’s decision. Meanwhile, the initial works of XAI propose a method that visualizes the model’s decision using class activation map or posterior gradient information in the classification task problem (Zhou et al., 2016; Selvaraju et al., 2017; Chattopadhay et al., 2018). However, these methods have the main limitation in that it visualizes extensively the model’s decision evidence. Classification-based visualization approaches are vulnerable to providing high-quality visual evidence or multi-region, it is because the classifier filters out relatively fewer discriminative features within the image by learning only the most discriminative area for the accurate prediction. To address this issue, the GAN-based approach has attracted a lot of attention in CV communities. GAN is designed
to learn the class distribution and generate the synthesized images that have attributes from other class domains via transferring ‘styles’ between the class domain. (Baumgartner et al., 2018) propose a method that attributes the pixels to a certain category to find the category-specific features using Wasserstein Generative Adversarial Networks (WGAN) in real 3D neuroimaging data. (Siddiquee et al., 2019) propose a method where GAN can remove the specific regions so that identify a minimal sub-category for the target pixels in the domain translation. However, previous works have mainly investigated the problem of domain translation for accurate localization. To show more accurate pixel-level differences, when turning the normal class into an abnormal class, both inter-domain translation and domain category attribute which represent the originality of the category characteristic should be considered. To address this issue, we propose a Pixel-Level Lesion Traversal Generative Adversarial Network (PILLET-GAN) that mines pixel-level pattern information of the features extracted from the generator for the randomly selected target domain (‘Healthy’ or ‘Disease’) via spatial attention block and then, fuse the mined features into transferring ‘style’. We refer to it as ‘hint’ for the ‘style’. Note that the PILLET-GAN can accurately localize the pixel-level lesions in the input X-ray image. In addition, we adopt a StarGAN (Choi et al., 2018) as our baseline system. Our experimental results show a significant improvement in localization results compared to the state-of-the-art models on the RSNA-Pneumonia dataset.

2. Related Work

2.1. Localization approaches in Medical domain

Localization for medical image analysis is the task of pointing out the lesions that affect the diagnosis of the diseases. It is required for the localization results to be specific and clear in that they present conclusive evidence not extensive. Initially, classification-based methods (Zhou et al., 2016; Selvaraju et al., 2017) were used to be compatible with the previous classification models. (Zhou et al., 2016) propose class activation map (CAM) to provide a classifier’s decision evidence that visualizes the feature maps where the activated maps are calculated by multiplying the last feature map from the CNN model with the weights of the fully-connected layer. (Selvaraju et al., 2017) propose Grad-CAM that uses gradient information to reflect the model’s decision in detail. Unlike CAM, Grad-CAM can have more flexible structure of the classification model in that they only need a gradient vectors for prediction. However, these approaches basically do not consider spatial information of an image in detail, and it is limited to be used in medical image analysis. To address this problem, various GAN-based localization methods (Baumgartner et al., 2018; Seah et al., 2019; Siddiquee et al., 2019; Taghanaki et al., 2019; Zhang et al., 2019) have been proposed for the medical image analysis. In particular, (Baumgartner et al., 2018) propose a visual attribution (VA) GAN that can obtain a disease affected map which shows the pixel-level difference between the ‘Healthy’ domain ‘Disease’ domain via domain translation techniques using wasserstein distance metric. (Siddiquee et al., 2019) propose a Fixed-Point GAN that captures undesirable changes while it is maintaining the input target domain.
3. Model

In this section, we present Pixel-Level Lesion Traversal Generative Adversarial Network (PILET-GAN) for providing pixel-level pneumonia localization. We train GAN to translate the input X-ray image into the target domain and then localize the pneumonia area through the difference between the input X-ray image and the 'healthy' domain X-ray image synthesized from the generator. Our network consists of two sub-networks to enhance the network’s attention power, as shown in Figure 1. First, spatial attention block is designed to reduce unnecessary pixels from the synthesized image. This sub-network selects both highly weighted features to the disease area and slightly weighted features to other domain area. Second, mixing block use output attention map that is extracted from spatial attention block to maintain the original information of the class with the residual connection. Finally, through two sub-blocks, we obtain a fine-grained localization map that indicates pneumonia regions.

3.1. Spatial Attention Block

The spatial attention block adjusts the transformation area to provide guidance on generator features. Given a feature map $L_{Dec}$ generated by the decoder, the spatial attention block is shown in Eq. (1) and (2).

$$SAB (L_{Dec}) = U \left( \sigma \left( \text{ConvNet} \left( F^{4 \times 4 / 4}_{avg} \left( L_{Dec} \right) \right) \right) \right) \tag{1}$$

$$\text{ConvNet}(\cdot) = f^{3 \times 3 / 1}_{conv} \left( \delta \left( IN \left( f^{3 \times 3 / 1}_{conv}(\cdot) \right) \right) \right), \tag{2}$$
where $L_{avg}^{4 \times 4}$ represents average pooling with kernel size 4 and stride 4, and $f_{conv}^{3 \times 3 / 1}$ represents a convolution network with kernel size 3 and stride 1. We use instance normalization $IN(\cdot)$ to eliminate style variations (Ulyanov et al., 2016). $\delta$ and $\sigma$ denote ReLU and Sigmoid, respectively. $U$ is the up-sampling operation. To sum up, we generate a spatial attention map by considering inter-spatial relationship of the feature maps. In particular, we use average pooling to integrate the neighboring spatial information, followed by average pooling, each down-sampled value becomes the representative value of the grid. Then, we use a convolution network to train the inter-dependency within representative values. We generate a probability map by applying the sigmoid function. The final feature map of the spatial attention block is up-sampled to match the size of $L_{Dec}$.

### 3.2. Mixing Block

After applying the spatial attention block to $L_{Dec}$, a residual connection is used to maintain the well-learned properties of the original features. Then, we reduce the dimension of the feature maps using a $1 \times 1$ convolution network to consider channel-to-channel correlation information. Applying a non-linear activation function, we obtain a lesion map $L_{Mixed}$, and a fake image $x_{fake}$ is generated by combining the lesion map and the original image. This process is formally expressed as follows:

$$L_{Mixed} = \tanh \left( f_{conv}^{1 \times 1 / 1} \left( (SAB \ (L_{Dec}) \oplus 1) \otimes L_{Dec} \right) \right)$$  \hspace{1cm} (3)

$$x_{fake} = x_{origin} \oplus L_{Mixed}$$  \hspace{1cm} (4)

We have found that the attention module especially enhances localization power in the lesion map while maintaining the quality of the synthesized image by reducing unnecessary changes.

### 3.3. Training Objectives

#### 3.3.1. Adversarial Loss

GAN proposed by (Goodfellow et al., 2014) shows excellent performance in image-to-image translation. However, there exist some limitations, such as mode collapse that produces only one image to deceive discriminators and saturated gradients that backwords meaningless gradients. Thus, we substitute GAN’s loss functions into WGAN (Arjovsky et al., 2017) and WGAN-GP (Gulrajani et al., 2017) as expressed in Eq. (5) and (6).

$$L_{adv}^G = -E_{x,t} \left[ D_{r/f} (G (x,t)) \right]$$  \hspace{1cm} (5)

$$L_{adv}^D = E_x \left[ D_{r/f} (x) \right] - E_{x,t} \left[ D_{r/f} (G (x,t)) \right] - \lambda_{gp} E_{\tilde{x}} \left[ \| \nabla_{\tilde{x}} D_{r/f} (\tilde{x}) \|_2 - 1 \|^{2} \right],$$  \hspace{1cm} (6)

where $x$ is input X-ray image and $t$ is target domain vector. $\lambda_{gp}$ represents a gradient penalty as a hyper-parameter. $\tilde{x}$ means a sampled point from the linearly interpolated distribution between the sampled distribution of the real data and the generated distribution of the synthesized data, and a gradient penalty term is added for the discriminator to approximate 1-lipschitz constraint (Gulrajani et al., 2017). $D$ and $G$ denote discriminator and generator, respectively.
3.3.2. Reconstruction Loss

The reconstruction loss is used to make the generator transform only domain-specific conditions while maintaining the shape of the input image as follows:

$$L_{\text{recon}} = E_{x,t} [\| G(G(x,t),o) - x \|_1],$$

(7)

where $o$ represents the original target domain vector.

3.3.3. Identity Loss

We use the identity loss to change only the most necessary parts representing a domain during image-to-image translation process. Eq. (8) shows our identity loss function that consists of two terms:

$$L_{id} = E_{x,t} [\| G(x,t) - x \|_1] + E_{x,o} [\| G(x,o) - x \|_1]$$

(8)

The first term is the target identity loss, represented as the L1 difference between the input image $x$ and the generated image $G(x,t)$. The second term is the original identity loss, which is used by the generator to preserve the identity of the original domain.

3.3.4. Classification Loss

To control the domain of the generated fake image, we additionally use a classification loss as follows:

$$L_{cls}^{\text{real}} = E_{x,o} [- \log D_{cls}(o|x)]$$

(9)

$$L_{cls}^{\text{fake}} = E_{x,t} [- \log D_{cls}(t|G(x,t))]$$

(10)

Using Eq. (9), the discriminator learns the distribution of real data. Eq. (10) encourages that the generator produces a fake image with the target domain vector to approximate the real data distribution using the discriminator.

3.3.5. Overall Loss

The total loss function of the discriminator and the generator is defined as follows:

$$L_D = -L_{\text{adv}}^D + \lambda_{cls}L_{cls}^{\text{real}}$$

(11)

$$L_G = L_{\text{adv}}^G + \lambda_{cls}L_{cls}^{\text{fake}} + \lambda_{recon}L_{recon} + \lambda_{id}L_{id},$$

(12)

where $\lambda_{cls}$, $\lambda_{recon}$, $\lambda_{id}$ are hyperparameters that control each loss term.

4. Experiments

4.1. Datasets

Our model is trained and evaluated on the RSNA-Pneumonia dataset. The RSNA-Pneumonia dataset is composed of 14,862 X-ray images with a frontal side of the thoracic. We divide the training set and validation set into 9:1 ratios. Both training set and validation set contain each 13,386 images (Pneumonia: 5,443, Healthy: 7,943) and 1,487 images (Pneumonia: 579, Healthy: 908).
4.2. Implementation and Training Details

4.2.1. Implementation

We optimize PILLET-GAN using the Adam (Kingma and Ba, 2014) optimizer with a batch size of 8. We set the learning rate of the generator and discriminator as 1e-4. We normalize the input X-ray image from [0, 255] to [-1,1]. We set $\lambda_{cls}=1$, $\lambda_{recon}=10$, $\lambda_{id}=10$ and $\lambda_{gp}=10$ for the hyperparameter of each classification loss, reconstruction loss, identification loss and gradient penalty loss. Our model is built upon the baseline StarGAN model (Choi et al., 2018). For stable learning, we balance the learning of generators and discriminators by the ratio of 1:5 to have the generator quickly learn from the meaningful gradient of the discriminator. We train PILLET-GAN for about 100K iterations and took about 10 hours on a single Titan Xp GPU.

4.2.2. Training Details

We train our generator with adversarial loss, classification loss, identity loss, and discriminator with adversarial loss and classification loss on the RSNA-Pneumonia dataset. During the training phase, we train the generator by randomly setting the target vector so that the generator can learn the distribution of disease and healthy domains. In the test phase, we perform localization using the lesion map, and the target vector is fixed to the ‘health’ class.

4.3. Comparison with State-Of-The-Arts Models

4.3.1. Qualitative Results

We compare PILLET-GAN to a classifier-based localization; CheXNet-CAM (Rajpurkar et al., 2017) and generative model-based localization methods; VA-GAN (Baumgartner et al., 2018), StarGAN (Choi et al., 2018) and Fixed-Point GAN (Siddiquee et al., 2019). We show the localization performance using pneumonia images containing ground-truth boxes in Figure 2. For the evaluation of localization, we calculated the difference map between a synthesized image from PILLET-GAN and an input X-ray image, and the map was normalized. We observe that CheXNet-CAM(2nd column) does not visualize multi-lesions and visualize a wide range of areas. Other models show poor performance in visualizing lesions due to too many unnecessary changes. On the other hand, PILLET-GAN shows a more precise localization of the disease area by focusing on the disease area using the spatial attention block. More qualitative results can be found in the appendix section.

4.3.2. Quantitative Results

For quantitative evaluation, we use an accuracy metric to evaluate the quality of synthesized images using a pretrained CheXNet-CAM (Rajpurkar et al., 2017) network and the pixel change rate to evaluate how few areas are used for the image translation. The pixel change rate is defined as

\[
\text{Pixel change rate} = \sum_{i,j} \frac{d_{ij}(x)}{WH}, \quad \text{where} \quad d_{ij}(x) = \begin{cases} 
0, & \text{if } x_{ij} = G_{ij}(x, t) \\
1, & \text{otherwise.}
\end{cases}
\]  

(13)
Figure 2: Qualitative evaluation of localization performance in various methods. (a) Chest X-ray images with ground truth boxes. (b) Classifier-based method (CheXNet). (c)-(f) Generative model-based methods including our PILLET-GAN. The left images of each figure show the lesion maps generated by each GAN model. The right images of each figure indicate overlapped images using the input image and the lesion maps.

Table 1: Accuracy using the pretrained CheXNet network after synthesizing an input image to the ‘Healthy’ domain. All methods show similar performance for the accuracy evaluation. P and H represent ‘Pneumonia’ and ‘Healthy’, respectively.

<table>
<thead>
<tr>
<th></th>
<th>VA-GAN</th>
<th>StarGAN</th>
<th>Fixed-Point GAN</th>
<th>PILLET-GAN (w/o SAB)</th>
<th>PILLET-GAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>P → H</td>
<td>75.34%</td>
<td>85.31%</td>
<td>88.42%</td>
<td>88.42%</td>
<td>87.56%</td>
</tr>
<tr>
<td>H → H</td>
<td>98.56%</td>
<td>99.55%</td>
<td>99.44%</td>
<td>99.66%</td>
<td>98.12%</td>
</tr>
<tr>
<td>P + H → H</td>
<td>86.82%</td>
<td>94.01%</td>
<td>95.15%</td>
<td>95.29%</td>
<td>94.01%</td>
</tr>
</tbody>
</table>

Tables 1, 2 show the accuracy and pixel change rates respectively. In the tables, ‘X → Y’ denotes that images belong to domain ‘X’ are synthesized to images corresponding to domain ‘Y’. PILLET-GAN not only translates much fewer pixels than other generative models for the synthesis to normal but also maintains similar accuracy compared to others.

Lastly, we used the localization metric that evaluates how much the predicted area is included in the ground truth boxes (Taghanaki et al., 2019). As shown in Figure 3, our method outperforms others in terms of the localization performance with 0.83 AUC.
Table 2: Pixel change rate when synthesizing an input image to the ‘Healthy’ domain. PILLET-GAN modifies smaller areas than others in image-to-image translation. P and H represent ‘Pneumonia’ and ‘Healthy’, respectively.

<table>
<thead>
<tr>
<th></th>
<th>VA-GAN</th>
<th>StarGAN</th>
<th>Fixed-Point GAN</th>
<th>PILLET-GAN (w/o SAB)</th>
<th>PILLET-GAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>P → H</td>
<td>0.7639</td>
<td>0.6212</td>
<td>0.5226</td>
<td>0.2391</td>
<td>0.2612</td>
</tr>
<tr>
<td>H → H</td>
<td>0.6995</td>
<td>0.3958</td>
<td>0.2425</td>
<td>0.1474</td>
<td>0.1061</td>
</tr>
<tr>
<td>P+H → H</td>
<td>0.7247</td>
<td>0.4836</td>
<td>0.3515</td>
<td>0.1831</td>
<td>0.1664</td>
</tr>
</tbody>
</table>

Figure 3: ROC curve on the RSNA-pneumonia dataset. Our model has the highest AUC value (0.83) on the image-level localization. Best viewed in color.

5. Conclusion

In this work, we present an approach to pixel-level localization for pneumonia disease. In particular, to preserve the key attributes within the domain, we employ a spatial attention block to remove noises that are irrelevant to domain distribution. Then, we obtain an attention map from the spatial attention block. The mixing block masks generator features with an attention map followed by adding the residual connection to masked features. Finally, we synthesize the original input image with output features from the mixing block. Our quantitative results show that PILLET-GAN effectively improves the performance of localization on the RSNA-Pneumonia dataset. Besides, our qualitative results show that our localized results are superior compared to both classification-based models and GAN-based models.

References


Appendix A. Comparison results of changed pixel area

<table>
<thead>
<tr>
<th>Input</th>
<th>VA-GAN</th>
<th>StarGAN</th>
<th>Fixed-Point GAN</th>
<th>PILLET-GAN (w/o SAB)</th>
<th>PILLET-GAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel change rate</td>
<td>0.7624</td>
<td>0.4128</td>
<td>0.2176</td>
<td>0.1626</td>
<td>0.0756</td>
</tr>
<tr>
<td>Pixel change rate</td>
<td>0.7778</td>
<td>0.6416</td>
<td>0.4470</td>
<td>0.2130</td>
<td>0.1470</td>
</tr>
</tbody>
</table>

Figure 4: Comparison results of changed pixel area when synthesized from ‘Pneumonia’ to ‘Healthy’. Compared to other methods, PILLET-GAN shows that focusing more on the lesion area than other networks because it tends to synthesize the essential parts during the generative process.
Appendix B. Success and failure cases with our PILLET-GAN

![Image of success and failure cases](image)

Figure 5: Successful cases with our PILLET-GAN. (a) Input x-ray images. (b) The activation map using a classifier-based method (CheXNet). (c)-(g) The lesion map using generative model-based methods. 1st and 2nd rows: Synthesized from “Pneumonia” to “Healthy”. 3rd and 4nd rows: Synthesized from “Healthy” to “Healthy”. Our method can localize lesions correctly even when an image looks similar to normal (e.g., 1st row versus 3nd or 4nd rows).
Figure 6: Failure cases of all localization methods. (a) Input x-ray images. (b) The activation map using a classifier-based method (CheXNet). (c)-(g) The lesion map using generative model-based methods. 1st–3rd rows: Synthesized from “Pneumonia” to “Healthy”. 4th row: Synthesized from “Healthy” to “Health”. Some cases such as thoracic images containing gas or fluid, pediatric chest disease, whole body chest images fail to produce correct lesion localization for all methods due to the limited data distribution of the RSNA-pneumonia dataset.