CheXplaining in Style: Counterfactual Explanations for Chest X-rays using StyleGAN

Matan Atad [*1] **Vitalii Dmytrenko** [*1] **Yitong Li** [*1] **Xinyue Zhang** [*1] **Matthias Keicher** [1] **Jan S. Kirschke** [1 2] **Bene Wiestler** [1 2] **Ashkan Khakzar** [1] **Nassir Navab** [1]

Abstract

Deep learning models used in medical image analysis are prone to raising reliability concerns due to their black-box nature. To shed light on these black-box models, previous works predominantly focus on identifying the contribution of input features to the diagnosis, i.e., feature attribution. In this work, we explore *counterfactual explanations* to identify what patterns the models rely on for diagnosis. Specifically, we investigate the effect of changing features within chest X-rays on the classifier's output to understand its decision mechanism. We leverage a StyleGAN-based approach (StyleEx) to create counterfactual explanations for chest X-rays by manipulating specific latent directions in their latent space. In addition, we propose EigenFind to significantly reduce the computation time of generated explanations. We clinically evaluate the relevancy of our counterfactual explanations with the help of radiologists. Our code is publicly available.[1]

1. Introduction

Chest X-ray, benefiting from its simple accessibility and fast availability, is currently one of the most common ways for the screening and diagnosis of a variety of thoracic diseases. Deep learning models have demonstrated promising potential for automated interpretation of chest X-rays at the level of practicing radiologists (Rajpurkar et al., 2017; Irvin et al., 2019; Wang et al., 2017). However, the black-box nature of deep learning models raises concerns about their reliability in clinical applications (Khakzar et al., 2021b;a).

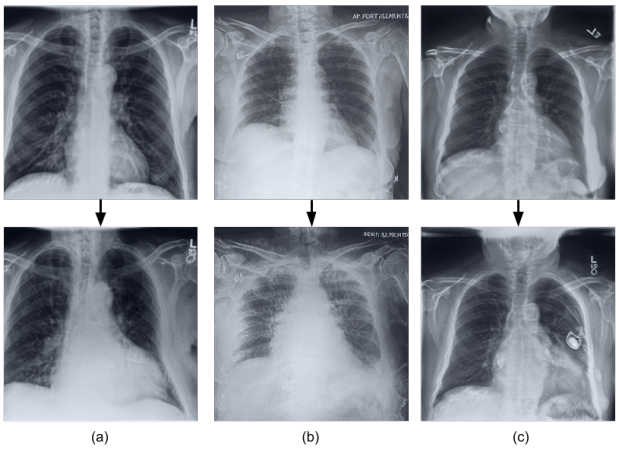

Figure 1. Some examples of the comparison of original chest X-ray images (first row) and their counterfactual explanations (second row) generated by our method. For each set of images, the emerging features in the counterfactuals are pathologically relevant: in (a) the width of the heart silhouette has been increased, corresponding to one of the main features of cardiomegaly; in (b) the appearance of the pleural recessus obstruction is a clear evidence of pleural effusion; in (c) a pacemaker was added, a demonstration of how the model learned to associate heart diseases (in this case cardiomegaly) with relevant indicators.

It is essential to know which patterns the models rely on for diagnosis in the clinical routine.

To interpret deep learning models in chest X-ray analysis, so far, most works leverage feature attribution (saliency) methods (Wang et al., 2017; Khakzar et al., 2021b; Rajpurkar et al., 2017). These methods identify the contribution of input features to the diagnosis. Despite providing valuable information to the users, they only show which regions are important for the prediction on medical images, but there is ambiguity surrounding *what* these features are. Some works (Wu et al., 2018; Khakzar et al., 2021a) provide further information regarding these features by analyzing neuron activation patterns on images with different concepts.

However, there is an emerging avenue for neural network interpretation that is not explored in medical applications

[*]Equal contribution [1]Technical University of Munich, Germany [2]Klinikum rechts der Isar, Munich, Germany. Correspondence to: Matan Atad <matan.atad@tum.de>, Yitong Li <yi_tong.li@tum.de>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML). Copyright 2022 by the author(s).

[1]https://github.com/CAMP-eXplain-AI/Style-CheXplain

and is known as counterfactual explanations. Counterfactual explanations in medical image diagnosis models translate to identifying the feature changes required in the medical images in order to lead the model to a different diagnosis.

To create counterfactual explanations, we need a method to extract visual features in images and change them in a semantic way. GAN-based models turn out to be an appropriate choice. Specific GAN structures can capture latent representations from the input data and control their features along these latent directions. Lang et al. (2021) presented a novel framework StyleEx to create a classifier-specific latent space and counterfactual explanations.

In this paper, we explore counterfactual explanations for chest X-ray diagnosis models. We evaluate whether the counterfactual explanations are clinically relevant with the help of radiologists from our university hospital. We employed a method based on StyleEx (Lang et al., 2021) and applied it to chest X-ray models to generate the counterfactual explanations. We improved the original method by factorizing the latent space instead of working on it directly, thus reducing the search time considerably.

2. Method

In this section, we introduce the details of deploying the StylEx (Lang et al., 2021) inspired methodology on chest X-ray models to generate counterfactual explanations. We further proceed with improving the style space search method to increase computational efficiency.

Given an input X-ray image $x \in X$ and its matching classifier label $C(x) = y$, to create its counterfactual explanation, the aim is to change x in a meaningful way, such that the changed image $\tilde{x}$ is as close as possible to x but $C(\tilde{x}) = \tilde{y}$ where $\tilde{y} \neq y$.

The method we use for counterfactual generation is based on StylEx proposed by Lang et al. (2021). Our implementation of StylEx is trained on the CheXpert dataset (Irvin et al., 2019). The architecture is comprised of a conditional StyleGAN2 (Karras et al., 2020), a frozen pretrained Classifier and an Encoder (Figure 2).

We first pretrained a DenseNet (Iandola et al., 2014) classifier on a Positive vs. Healthy binary setting per pathology in the dataset. Both the Generator and the Discriminator will then be conditioned on the class labels y predicted from the classifier by concatenating an embedding of the image labels to their inputs. The encoder is based on the Discriminator architecture while removing the batch-normalization layer. The main purpose of the encoder is to allow for mapping of any kind of image into the latent space, by which we will be able to create counterfactuals for real images.

Algorithm 1 EigenFind

Input: Classifier C, Encoder E, Generator G, number of Eigenvectors to consider k, Degree of change d, Images X classified as y
Return: Counterfactuals $X_{explained}$

$X_{explained} \leftarrow \varnothing$
$V_{max} \leftarrow \varnothing$
$V \leftarrow PCA(G)[1:k]$
for v in V **do**
  for x in X **do**
    $\tilde{x} \leftarrow G(E(x) + d * v)$
    $\delta[x, v] \leftarrow C(\tilde{x}) - C(x)$
  end for
  $\bar{\Delta}[v] = \frac{1}{|X|} \sum_{x \in X} \delta[x, v]$
end for
repeat
  $v_{max} \leftarrow \text{argmax}_v \bar{\Delta}$
  for x in X **do**
    $\tilde{x} \leftarrow G(E(x) + d * v_{max})$
    if $C(\tilde{x}) = \bar{y}$ **then**
      $X_{explained} = X_{explained} \cup \tilde{x}$
      $X = X \setminus x$
    end if
  end for
  $V_{max} = V_{max} \cup v_{max}$
  delete $\bar{\Delta}[v_{max}]$
until $|X| = 0$ **or** $|\bar{\Delta}| = 0$

We trained the entire architecture at once, passing in each iteration both latents originating in noise and in encoded images. Training along with the raw noise input will help us to maintain the optimal training direction of the conditional StyleGAN. The trained StyleSpace will capture the features that are decisive for the classifier's prediction, the most significant features of which will be later detected and extracted using specific searching algorithm for counterfactual generation.

We propose our algorithm EigenFind (Algorithm 1) for a more efficient counterfactual search. In the previous paper, Lang et al. (2021) presented the AttFind algorithm which iterates over all coordinates in the StyleSpace while changing them one by one, searching for coordinates with the largest affect on the classifier decision. We factorize the StyleSpace with PCA (Shen & Zhou, 2021) and modify the algorithm to iterate over Eigenvectors instead.

In EigenFind, for images X classified as y, we calculate how moving their latents in the direction of each of the top k StyleSpace Eigenvectors affects the classifier decision[2].

[2]Both positive and negative directions are evaluated, but the negative directions are omitted here for brevity.

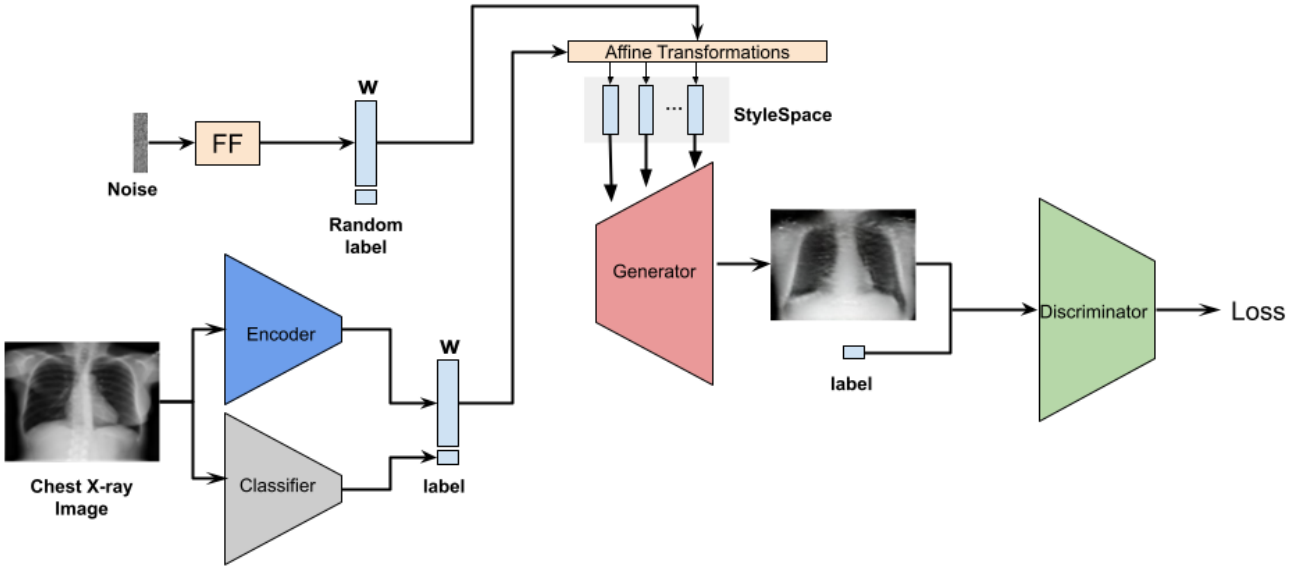

Figure 2. The pipeline captures classifier features in the StyleSpace, which are later used to generate counterfactual explanations. The whole architecture consists of a conditional StyleGAN and a pretrained frozen classifier, along with an encoder which allows mapping of images to the latent space. The Generator and Discriminator are conditioned on a label predicted by the classifier for the original input image. We use our EigenFind (Algorithm 1) to find the classifier specific directions in the learnt StyleSpace.

Next, we follow Lang et al. (2021) and estimate the most significant Eigenvectors by calculating the average difference between the classifier logit before and after the change on all input images X. Finally, for each image, we find which of the most significant Eigenvectors is able to flip the image label to $\tilde{y}$. The resulting image $\tilde{x} \in X_{explained}$ is the counterfactual.

The time complexity of EigenFind is $\mathcal{O}(kn)$, where n is the number of images and k is the number of top Eigenvectors we consider. The time complexity of AttFind (Lang et al., 2021) is also linear $\mathcal{O}(mn)$, where m is the number of channels in the StyleSpace feature map. In practice, m is determined by the StyleGAN architecture based on the size of the input image (for 256×256 $m = 3040$[3]) and we chose $k = 8$. Since $k \ll m$ the running time is considerably reduced.

3. Experiments & Evaluation

We trained the pipeline separately for three common thoracic pathologies in the CheXpert dataset (Irvin et al., 2019): Cardiomegaly, Pleural Effusion and Atelectasis. Based on our EigenFind, we then generated counterfactual explanations for these pathologies. Some examples are shown in Figure 1 - the first row demonstrates the original healthy chest X-ray images, while the second row concludes the corresponding generated counterfactuals. From each pair of images, the features emerging in the counterfactuals are representative of the main features of each disease, which visually affirms the pathological-relevance of the features found by EigenFind.

Furthermore, in order to evaluate whether the counterfactual explanations found by our method are indeed clinically relevant, we cooperated with radiologists from our university hospital. They helped to diagnose which features changed in counterfactuals vs. the originals for these three pathologies (Table 2).

In our evaluation setting, the radiologists first listed the main features and possible secondary findings during diagnosis of each disease. After randomly selecting 10 images that have been classified as *Healthy* samples (i.e., originals), we separately moved their latent representations in the direction of each of the three most significant Eigenvectors, which were obtained with EigenFind (Algorithm 1). After each movement, the radiologists evaluated whether the previously listed disease features existed in the newly generated images (i.e., counterfactuals) or not.

The evaluation results are demonstrated in the last three columns of Table 2. The results indicate that most of the main features could be spotted in the counterfactuals generated by our EigenFind, thus indicating that our most signifi-

[3]We do not consider the StyleGAN ToRGB layers as part of the StyleSpace given to the AttFind search, since these are shown to affect only the output image color (Wu et al., 2021).

Table 1. Comparison of the percentage of explained images for each of the three pathologies in CheXpert (Irvin et al., 2019) (i.e., ones for which a counterfactual could be created), along with the search time between the two algorithms: AttFind by Lang et al. (2021) and our EigenFind. Both algorithms achieved comparable results on counterfactual generation, while our method reduced the searching time considerably.

Pathology	AttFind Lang et al. (2021)	EigenFind Ours
Atelectasis	94%	94%
Cardiomegaly	96%	95%
Pleural Effusion	94%	91%
Search Time	12 hours	5 minutes

cant Eigenvectors can help with identifying clinical-relevant features. By doing so, we can not only easily spot which regions are crucial for the classifiers to predict different pathologies, but more importantly, we can also understand *what* these determined features are, by comparing the vivid changes between the original images and their counterfactuals.

In addition, we compared the ability to explain images (i.e to create counterfactuals) of our EigenFind algorithm with Lang et al. (2021) AttFind. In Table 1 our method achieves results on par with Lang et al. (2021), while requiring a fraction of the search time on a Tesla P100 GPU.

In our experiments, the StyleGAN2 pipeline was trained with an Adam optimiser for $40K$ iterations with a batch size of 32 on images of size 256×256. The learning rate was set to 0.0016 for the Generator, 0.0018 for the Discriminator and 0.002 for the Encoder. Path length regularization was applied every 4 epochs for the Generator and $R1$ regularization was applied every 16 epochs for the Discriminator. For the evaluation of the counterfactual search algorithms, we took 600 random images for each pathology. For EigenFind we considered the top $k = 8$ Eigenvectors and a degree of change of $d = 10$.

4. Conclusion

To address the concerns regarding the explainability of deep learning models in medical image analysis, we explored the chest X-ray domain and investigated counterfactual explanations to identify the feature changes in chest X-ray images that can lead the classifier to a different diagnosis.

To create such counterfactual explanations, we leveraged a StyleGAN-based approach by manipulating specific latent directions in the pre-trained latent space of the generator. The newly generated images after the latent space manipulation become our counterfactuals. We also propose the EigenFind algorithm to significantly reduce the computation time for counterfactuals generation by factorizing the latent space with PCA, and working on the Eigenvectors instead.

Table 2. Radiologists' evaluation of our generated counterfactuals. Common disease features and secondary findings for each of the three pathologies are listed by the radiologists in the first column. Then the radiologists evaluated if the corresponding features existed, after moving the latent representation of a set of 10 Healthy images in the direction of the 3 most significant Eigenvectors respectively (listed as the last three columns).

Features of Cardiomegaly	v_{17}	v_2	v_{18}
Increased cardiothoracic ratio	✓	✓	✓
Secondary findings:			
Reduced lung tissue opacity	✓		
Pleural Effusion	✓	✓	
Pacemaker			✓
Older patients		✓	
Features of Pleural Effusion	v_{15}	v_7	v_{12}
Obstruction of the pleural recessus	✓	✓	✓
Opaque lower lungs			✓
Secondary findings:			
Increased cardiac diameter	✓		
Fluid overload	✓		
Pneumonia		✓	
Features of Atelectasis	v_{15}	v_7	v_{12}
Mediastinal shift	✓	✓	✓
Wide barrel-like thorax	✓		✓
Secondary findings:			
Pleural Effusion	✓	✓	✓
Infiltration	✓	✓	
Older patients	✓		

Furthermore, we evaluated whether such classifier-decisive features spotted by our EigenFind algorithm are clinically relevant with the help of radiologists. The results demonstrate that the most significant Eigenvectors obtained from EigenFind are able to help with identifying clinical-relevant features in chest X-rays. The feature changes in the generated counterfactuals are in accordance with the main diagnosing features for common thoracic diseases. In addition, the model also learned to associate thoracic diseases with relevant indicators such as pacemaker and age.

References

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., and Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, 2019. URL http://arxiv.org/abs/1901.07031.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Khakzar, A., Musatian, S., Buchberger, J., Valeriano Quiroz, I., Pinger, N., Baselizadeh, S., Kim, S. T., and Navab, N. Towards semantic interpretation of thoracic disease and covid-19 diagnosis models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 499–508. Springer, 2021a.

Khakzar, A., Zhang, Y., Mansour, W., Cai, Y., Li, Y., Zhang, Y., Kim, S. T., and Navab, N. Explaining covid-19 and thoracic pathology model predictions by identifying informative input features. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 391–401. Springer, 2021b.

Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W. T., Isola, P., Globerson, A., Irani, M., and Mosseri, I. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Shen, Y. and Zhou, B. Closed-form factorization of latent semantics in gans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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. *CoRR*, 2017. URL http://arxiv.org/abs/1705.02315.

Wu, J., Zhou, B., Peck, D., Hsieh, S., Dialani, V., Mackey, L., and Patterson, G. Deepminer: Discovering interpretable representations for mammogram classification and explanation. *arXiv preprint arXiv:1805.12323*, 2018.

Wu, Z., Lischinski, D., and Shechtman, E. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872, 2021.