Using StyleGAN for Visual Interpretability of Deep Learning Models on Medical Images

Kathryn Schutte[*], **Olivier Moindrot**, **Paul Hérent**, **Jean-Baptiste Schiratti**, **Simon Jégou**

Owkin, Inc.
New York, NY, USA

Abstract

As AI-based medical devices are becoming more common in imaging fields like radiology and histology, interpretability of the underlying predictive models is crucial to expand their use in clinical practice. Existing heatmap-based interpretability methods such as GradCAM only highlight the location of predictive features but do not explain how they contribute to the prediction. In this paper, we propose a new interpretability method that can be used to understand the predictions of any black-box model on images, by showing how the input image would be modified in order to produce different predictions. A StyleGAN is trained on medical images to provide a mapping between latent vectors and images. Our method identifies the optimal direction in the latent space to create a change in the model prediction. By shifting the latent representation of an input image along this direction, we can produce a series of new synthetic images with changed predictions. We validate our approach on histology and radiology images, and demonstrate its ability to provide meaningful explanations that are more informative than GradCAM heatmaps. Our method reveals the patterns learned by the model, which allows clinicians to build trust in the model's predictions, discover new biomarkers and eventually reveal potential biases.

1 Introduction

As of September 2020, the FDA had approved 64 AI-based medical devices (Benjamens et al., 2020), and for the first time the Centers for Medicare & Medicaid Services (CMS) approved the reimbursement of deep-learning powered stroke detector for brain CT scans (Viz.ai, 2020). The advances of deep learning in computer vision (Krizhevsky et al., 2012) are especially promising in medical imaging fields such as radiology (Ardila et al., 2019), histology (Coudray et al., 2018), dermatology (Esteva et al., 2017) or ophthalmology (Gulshan et al., 2016).

While many deep learning techniques may provide state-of-the-art predictive performance, *interpretable* deep learning models are necessary for regulatory approval, as their ability to explain their predictions can reveal potential biases and failure modes, as seen in the case of (Oakden-Rayner, 2017). Additionally, interpretable models also provide new opportunities for biomedical investigation, as evidenced in (Courtiol et al., 2019). Finally, such models are able to make inroads with medical experts, as their explainability helps build confidence in their utility (Holzinger et al., 2019). As illustrated by the COVID-19 crisis (Bai et al., 2020; Li et al., 2020; Wang et al., 2020), the go-to method for model interpretation in the medical imaging field is GradCAM (Selvaraju et al., 2017), which produces a coarse heatmap based on gradient intensity to identify which areas of the input image are responsible for the prediction. However, these heatmaps only highlight the location of predictive features but do not explain how they contribute to the prediction. In an image where

[*]Corresponding author: kathryn.schutte@owkin.com

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

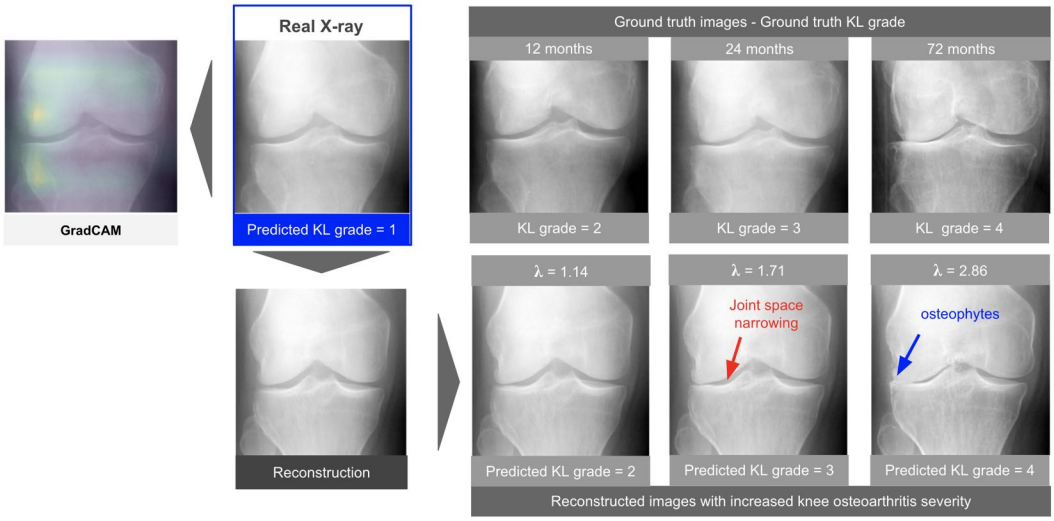

Figure 1: Our method applied to knee osteoarthritis severity prediction on an X-ray image. The input image is gradually modified to increase the osteoarthritis severity. The GradCAM heatmap is computed on the input image to compare both interpretability methods. X-rays of the patient's later visits are displayed to visually assess the clinical relevance of our method.

information is diffuse, the heatmap cannot highlight any specific region so GradCAM is not sufficient to interpret the model predictions.

In this paper, we propose a new interpretability method that generates small synthetic transformations of the original image that would lead to different model predictions. We train a generative model called StyleGAN (Karras et al., 2019, 2020) and find the minimal modification in the latent space that changes the model prediction, which ensures that generated images remain as close as possible to the original image. Seah et al. (2019) explore a similar idea by using an older GAN algorithm to create heatmaps highlighting features of congestive heart failure, but their method cannot be applied to any black-box model. Fetty et al. (2020) manipulate three attributes of the StyleGAN latent space in order to enlarge datasets with synthetic images. We validate our interpretability method on two different imaging modalities and demonstrate its ability to provide meaningful explanations of the predictions, and its potential to discover new biomarkers.

2 Method

We propose to create StyleGAN-generated visualizations that explain the predictions of a deep neural network in an interpretable manner. Let f be a classifier (e.g. a fully convolutional neural network) trained on a dataset $\mathcal{D} = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes a set of 2D images and $\mathcal{Y}$ a finite set of labels. Our method consists of three steps. First, the images in $\mathcal{X}$ are used to train a StyleGAN2 (Karras et al., 2019), which is an improved GAN whose *generator* $G : \mathcal{W} \to \mathcal{X}$ has a linearly disentangled intermediate *latent space* $\mathcal{W} \subset \mathbb{R}^{512}$. The generator G is used to generate a set of synthetic images $(G(\mathbf{w}_i))$, where the $\mathbf{w}_i$ are sampled in the latent space $\mathcal{W}$. Then, we train (using a Mean Squared Error loss) a ResNet50 (He et al., 2016) *encoder* $E : \mathcal{X} \to \mathcal{W}$ on the synthetic dataset $(\mathbf{w}_i, G(\mathbf{w}_i))$ to retrieve the latent representation $\mathbf{w}_i$ from a generated image $G(\mathbf{w}_i)$. Finally, a logistic regression classifier $\tilde{f}(\mathbf{w_i}) = \sigma\left(\boldsymbol{\alpha}^\top \mathbf{w_i} + \boldsymbol{\beta}\right)$ is trained on the latent space $\mathcal{W}$ to predict the estimated labels $\tilde{y}_i = f(G(\mathbf{w_i}))$ associated to each latent vector $\mathbf{w}_i \in \mathcal{W}$. Given a new input image $\mathbf{x} \in \mathcal{X}$, our method translates the latent vector $\mathbf{w} = E(x)$ along the direction $\boldsymbol{\alpha}$. We can then create new images from the latent representation via $G(\mathbf{w} + \lambda\boldsymbol{\alpha})$ associated with a lower or a higher prediction depending on the value of $\lambda \in \mathbb{R}$.

3 Experiments

3.1 Knee osteoarthritis severity prediction on X-ray images

We first demonstrate our method by explaining the predictions of an osteoarthritis severity predictor on X-ray images. The dataset on which the predictor has been trained consists of 20,123 X-rays of patients suffering from knee osteoarthritis collected by the Osteoarthritis Initiative (OAI) (Nevitt et al., 2006). Each patient has one to eight 12-month follow-up X-rays, as well as associated clinical data, including the Kellgren and Lawrence (KL) grade (Kohn et al., 2016). The KL grade describes a degree of osteoarthritis severity and ranges from 0 to 4: grades 0 and 1 mean no or doubtful osteoarthritis, while grades 2 to 4 mean mild to severe osteoarthritis.

The image classifier f is a ResNet50 trained on the multi-class prediction task. To fit this multi-class setting to our method, we transform it to a binary classification task by pooling grades 0 and 1 versus grades 2 to 4. The predictor f obtains 89% test AUC on this binary task, while $\tilde{f}$ obtains 80% test AUC on the latent space. Three radiologists evaluated the quality of the StyleGAN generator with a Turing test. They reach 58% accuracy on average, showing that synthetic and real X-rays are almost indistinguishable.

In Figure 1, our interpretability method is applied to a real X-ray image. The GradCAM heatmap provides topographical information by showing that the osteoarthritis features are located in lateral femorotibial space. Our method provides more than topographical information by showing the gradual emergence of the different osteoarthritis features as the KL grade increases, such as the joint space narrowing (red arrow) and osteophytes (blue arrow). By comparing the synthetic evolution of the image to the real evolution of the patient at 12, 24 and 72 months after baseline, we observe that the direction found in the latent space corresponds to a biologically plausible osteoarthritis progression.

3.2 Tumor detection on histology images of metastatic lymph nodes

We apply the same method to histology images, to explain the predictions of a metastasis detector on Camelyon16 (Bejnordi et al., 2017). The dataset contains 224,166 patch images from breast cancer lymph node whole-slide images, each with a binary label indicating the presence of tumor cells.

The image classifier f is a ResNet50 trained on this dataset, obtaining 92% test AUC, while the latent predictor $\tilde{f}$ reaches 95% test AUC. Figure 2 shows our interpretability method on two images: patch B contains tumoral cells while patch A does not. The GradCAM heatmaps are not relevant here because the informative features are spread over the entire image. On the contrary, our approach reveals clinically relevant features. On patch A, it shows the appearance of tumor cells (blue arrow) and the disappearance of lymphocytes (red arrow) as the tumor probability increases, and inversely on patch B.

We can see that the encoder-decoder model is not able to perfectly reconstruct histology images, as opposed to knee X-rays. A possible explanation is that the StyleGAN model does not generate images that are under-represented in the training set. This issue is highlighted in this particular use-case as there is more variability in the histology images than in the knee X-ray images. Recently, Yu et al. (2020) propose to overcome this data coverage challenge by harmonizing adversarial training with reconstructive generation.

4 Conclusion

In this study we explored the potential of StyleGANs to explain the predictions of black-box models on medical images. Although heatmap-based methods dominate the interpretability field, they only highlight the localization of predictive features in the image. Our method provides an intuitive way for medical researchers to understand *where* are located the predictive features in the image and *how* they impact the prediction by showing modified views of the input image that would produce different predictions. This method shows how the model learned to solve the prediction task which allows clinicians to build trust in the model's predictions, discover new biomarkers and eventually reveal potential biases. In both experiments, our method proved that the models learned clinically relevant features.

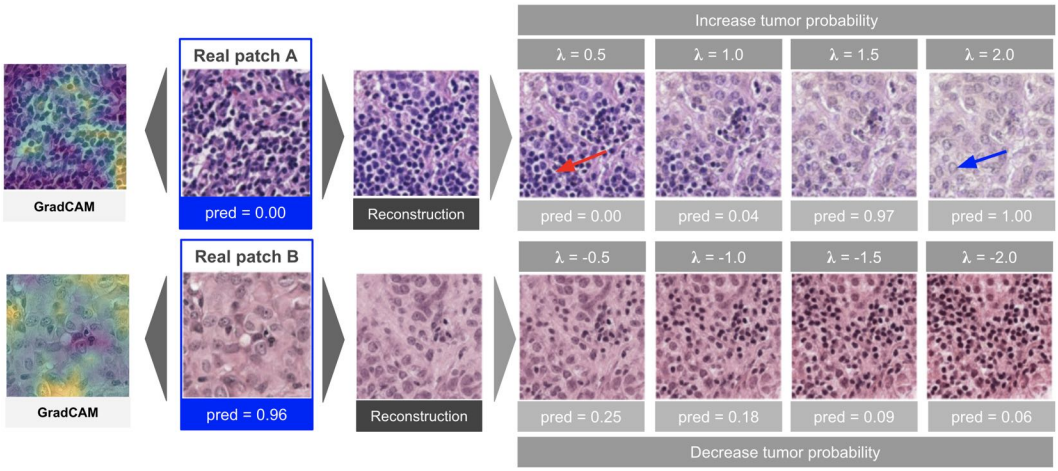

Figure 2: Our method applied to tumor probability prediction on two histology tiles of metastatic lymph nodes. The input image is gradually modified to increase (on patch A) or decrease (on patch B) the tumor probability. The GradCAM heatmap is computed on the input images to compare both interpretability methods.

Acknowledgments

We thank Eric W. Tramel for his valuable feedback on the manuscript. We thank the three radiologists Eric Pessis, François Legoux and Thibaut Emorine for their participation in the Turing Test.

References

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961.

Bai, H. X., Wang, R., Xiong, Z., Hsieh, B., Chang, K., Halsey, K., Tran, T. M. L., Choi, J. W., Wang, D.-C., Shi, L.-B., Mei, J., Jiang, X.-L., Pan, I., Zeng, Q.-H., Hu, P.-F., Li, Y.-H., Fu, F.-X., Huang, R. Y., Sebro, R., Yu, Q.-Z., Atalay, M. K., and Liao, W.-H. (2020). Artificial intelligence augmentation of radiologist performance in distinguishing covid-19 from pneumonia of other origin at chest ct. *Radiology*, 296(3):E156–E165. PMID: 32339081.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210.

Benjamens, S., Dhunnoo, P., and Meskó, B. (2020). The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *npj Digital Medicine*, 3(1):1–8.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567.

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

Fetty, L., Bylund, M., Kuess, P., Heilemann, G., Nyholm, T., Georg, D., and Löfstedt, T. (2020). Latent space manipulation for high-resolution medical image synthesis via the stylegan. *Zeitschrift für Medizinische Physik*.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410.

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

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

Kohn, M. D., Sassoon, A. A., and Fernando, N. D. (2016). Classifications in brief: Kellgren-lawrence classification of osteoarthritis.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., Cao, K., Liu, D., Wang, G., Xu, Q., Fang, X., Zhang, S., Xia, J., and Xia, J. (2020). Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: Evaluation of the diagnostic accuracy. *Radiology*, 296(2):E65–E71. PMID: 32191588.

Nevitt, M., Felson, D., and Lester, G. (2006). The osteoarthritis initiative. *Protocol for the Cohort Study*, 1.

Oakden-Rayner, L. (2017). Exploring the chestxray14 dataset: problems. `https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/`.

Seah, J. C. Y., Tang, J. S. N., Kitchen, A., Gaillard, F., and Dixon, A. F. (2019). Chest radiographs in congestive heart failure: Visualizing neural network learning. *Radiology*, 290(2):514–522. PMID: 30398431.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Viz.ai (2020). Viz.ai granted medicare new technology add-on payment. `https://www.prnewswire.com/news-releases/vizai-granted-medicare-new-technology-add-on-payment-301123603.html`.

Wang, Z., Liu, Q., and Dou, Q. (2020). Contrastive cross-site learning with redesigned net for covid-19 ct classification. *IEEE Journal of Biomedical and Health Informatics*.

Yu, N., Li, K., Zhou, P., Malik, J., Davis, L., and Fritz, M. (2020). Inclusive gan: Improving data and minority coverage in generative models. *arXiv preprint arXiv:2004.03355*.