CheXplaining in Style

1. General Info

Contact Person: Matthias Keicher, Kristina Mach, Ashkan Khakzar

Contact Email: matthias.keicher@tum.de, kristina.mach@tum.de, ashkan.khakzar@tum.de

Outcome: The result of the project will potentially be published in MICCAI 2022

2. Project Abstract

In this project, we interpret classifiers trained on Chest X-ray datasets[1][2][3] via counterfactual explanation. I.e. we will change specific features within the Chest X-rays and observe the corresponding change in the output. For instance, one can change the size of a nodule in the X-ray and observe whether the classifier's output is sensitive to this change. In order to identify what features to change, we will use a StyleGAN based methodology[4], which captures disentangled directions in the latent space. The objective is to find out whether changing X-ray features along these directions are human-interpretable.

3. Background and Motivation

Deep learning models have demonstrated promising potential in diagnosing pathologies on Chest X-rays. However, the black-box nature of deep learning methods raises concerns regarding their reliability. For their adoption in the clinical routine, it is required to know what patterns the models rely on for the diagnosis. There exist a plethora of approaches that identify what features the model use[5][6]. One of the main approaches is counterfactual explanations. In this paradigm, the input is changed and the corresponding change in output is observed. For instance, one can change specific features that clinicians use in their diagnosis, and observe whether the model is sensitive to such changes. In our project, we use a GAN-based methodology to capture disentangled latent representations and change the features along these latent directions. Ideally, we will observe that these latent directions correspond to clinically relevant features.

4. Technical Prerequisites

- Background in deep learning and GANs
- "Mad" Python/Pytorch skills

5. Benefits:

- Working on a state of the art deep learning explanation approach
- Working on the largest public medical datasets available (CheXpert, NIH Chest Xray and MIMIC)
- Possibility of finding novel findings and writing a scientific paper.

---

[1] https://stanfordmlgroup.github.io/competitions/chexpert/
[2] https://physionet.org/content/mimic-cxr/2.0.0/
[3] https://paperswithcode.com/dataset/chestx-ray14
[4] Lang, Oran, et al. "Explaining in Style: Training a GAN to explain a classifier in StyleSpace." arXiv preprint arXiv:2104.13369 (2021).
https://explaining-in-style.github.io/
[5] Khakzar, Ashkan, et al. "Explaining COVID-19 and Thoracic Pathology Model Predictions by Identifying Informative Input Features." arXiv preprint arXiv:2104.00411 (2021).
[6] Khakzar, Ashkan, et al. "Towards Semantic Interpretation of Thoracic Disease and COVID-19 Diagnosis Models." arXiv preprint arXiv:2104.02481 (2021).

6. Work-packages and Time-plan:

	Description	#Students	From	To
WP1	Group 1: Understanding chest X-ray models' literature	Group 1	01.11	01.12
WP2	Group 2: Understanding the background works on GAN based explanations and feature disentanglement	Group 2	01.11	01.12
WP3	Group 1: Modeling and implementing the classification problem on multiple chest x-ray datasets	Group 1	01.12	intermediate presentation
WP4	Group 2: Implementing the StyleEx method on the architecture developed by Group 1	Group 2	01.12	intermediate presentation
M1	Intermediate Presentation II	all		
WP5	Group 1&2: Explore and analyze the results	all		
WP7	(Optional) Explore other GAN based approaches for the purpose of counterfactual explanations	all		
WP8	Documentation	all		
M2	Final Presentation	all		