This explains *That*: Congruent Image–Report Generation for Explainable Medical Image Analysis with Cyclic Generative Adversarial Networks

Abhineet Pandey⋆[1], Bhawna Paliwal⋆[1], Abhinav Dhall[1,2], Ramanathan Subramanian[1,3], and Dwarikanath Mahapatra[4]

[1] Indian Institute of Technology (IIT) Ropar, India
[2] Monash University, Melbourne, Australia
[3] University of Canberra, Canberra, Australia
[4] Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Abstract. We present a novel framework for *explainable labeling and interpretation* of medical images. Medical images require specialized professionals for interpretation, and are explained (typically) via elaborate textual reports. Different from prior methods that focus on medical report generation from images or vice-versa, we novelly generate congruent image–report pairs employing a cyclic-Generative Adversarial Network (cycleGAN); thereby, the generated report will adequately explain a medical image, while a report-generated image that effectively characterizes the text visually should (sufficiently) resemble the original. The aim of the work is to generate *trustworthy and faithful explanations* for the outputs of a model diagnosing chest x-ray images by pointing a human user to similar cases in support of a diagnostic decision. Apart from enabling transparent medical image labeling and interpretation, we achieve report and image-based labeling comparable to prior methods, including state-of-the-art performance in some cases as evidenced by experiments on the *Indiana Chest X-ray* dataset.

Keywords: Explainability · Medical image analysis · Multimodal

1 Introduction

Medical images present critical information for clinicians and epidemiologists to diagnose and treat a variety of diseases. However, unlike natural images (scenes) which can be easily analyzed and explained by laypersons, medical images are hard to understand and interpret without specialized expertise. [1]

Artificial intelligence (AI) has made rapid advances in the last decade thanks to deep learning. However, the need for accountability and transparency to explain decisions along with high performance, especially in healthcare, has spurred the need for *explainable machine learning.* While natural images can be analyzed

⋆ contributed equally

and explained by *decomposing* them into semantically-consistent and prototypical visual segments [18], multimodal approaches for prototypical explanations are essential for interpreting and explaining medical imagery given the tight connection between image and text in this domain.

1.1 Prior Work

Prior works on medical image interpretation and explainability have either attempted to characterize (chest) x-rays in terms of multiple pathological labels [19] or via automated generation of imaging reports [1,11,15]. The Chexnet framework [19] employs a 121-layer convolution network to label chest x-rays. A multi-task learning framework is employed to generate both tags and elaborate reports via a hierarchical long short-term memory (LSTM) model in [1]. Improvements over [1] are achieved by [11] and [15] by employing a topic model and a memory driven transformer respectively. While the above report-generation works achieve excellent performance, and effectively learn mappings between the image and textual features, they nevertheless do not *verify* if the generated report characterizes the input x-ray. It is this constrained characterization in our suggested work that helps us generate prototypical chest x-ray images serving as explanations. In more recent work saliency maps have been used to select informative xray images [16]

1.2 Our Approach

This work differently focuses on the generation of *coherent* image–report pairs, and posits that if the image and report are conjoined counterparts, one should inherently describe the characteristics of the other. It is the second part of the radiology report generation model i.e. generation of prototypical images from generated reports that serve as explanations for the generated reports. The explainable model proposed can be characterised as a model having post hoc explanations where an explainer outputs the explanations corresponding to the output of the model being explained. The approach to explanations in such an explanation technique as ours is different from methods which propose simpler models such as decision trees that are inherently explainable. Having prototypical images as explanations has been used in case of natural images in [18] (discussed earlier) and [26]. None of the approaches explores the paradigm of prototypical image generation as explanations in case of medical images which has been proposed in this work novelly with a multimodal approach.

1.3 Contributions

Overall, we make the following research contributions:

1. We present the first multimodal formulation that enforces the generation of *coherent* and *explanatory* image–report pairs via the cycle-consistency loss employed in cycleGANs [14].

2. Different from prior works, we regenerate an x-ray image from the report, and use this image to quantitatively and qualitatively evaluate the report quality. Extensive labeling experiments on textual reports and images generated via the Indiana Chest X-ray dataset [20] reveal the effectiveness of our multi modal explanations approach.
3. We evaluate the proposed model on two grounds namely: the quality of generated reports and the quality of generated explanations. Our method achieves results comparable to prior methods in report generation task, while achieving state-of-the-art performance in certain conditions. The evaluations done for post-hoc explanations show the employability of cycle consistency constraints and multimodal analysis as an explanation technique.
4. As qualitative evaluation, we present Grad CAM [4]-based *attention maps* conveying where a classification model *focuses* to make a prediction.

2 Method

2.1 Coherent Image-Report Pairs With CycleGANs

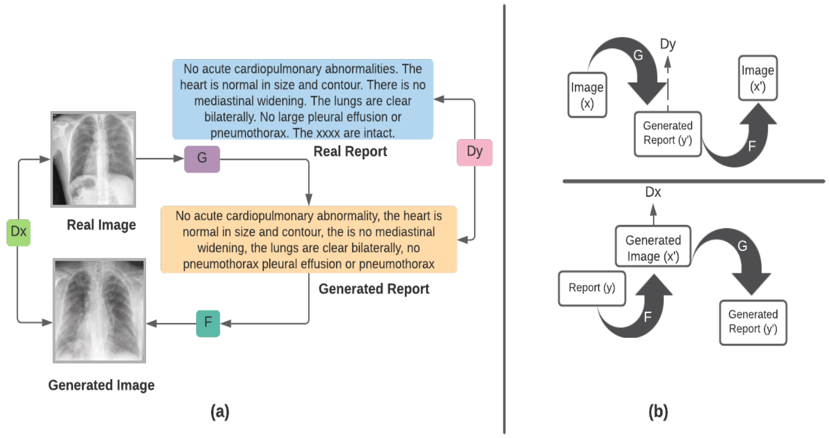
We aim to model the tight coherence between image and textual features in the chest x-ray images and reports through our multi-modal report generation model. Reports generated should be such that an x-ray image generated with just these generated reports as the input should be similar to ground truth x-ray images; and prototypical x-ray images generated as explanations should be such that a report generated from these images as inputs resembles original report. We hence devise a multimodal, paired GAN architecture explicitly modeling the cycle consistent constraints based on CycleGAN [14] with data of type {*image*, *text/labels*}.

2.2 CycleGAN

Given two sets of images corresponding to domains X and Y (for example, two sets of paintings corresponding to different styles), cycleGAN enables learning a mapping $G : X \rightarrow Y$ such that the generated image $G(x) = y'$, where $x \in X$ and $y \in Y$, looks similar to y.

The generated images y' are also mapped back to images x' in domain X. Hence, cycleGAN also learns another mapping $F : Y \rightarrow X$ where $F(y') = x'$ such that x' is similar to x. The structural cycle-consistency assumption is modeled via the cycle consistency loss, which enforces $F(G(x))$ to be similar to x, and conversely, $G(F(y))$ to be similar to y. Hence the objective loss to be minimized enforces the following four constraints:

$$G(x) \approx y, F(y) \approx x \quad and \quad F(G(x)) \approx x, G(F(x)) \approx y \tag{1}$$



No acute cardiopulmonary abnormalities. The heart is normal in size and contour. There is no mediastinal widening. The lungs are clear bilaterally. No large pleural effusion or pneumothorax. The xxxx are intact.

No acute cardiopulmonary abnormality, the heart is normal in size and contour, the is no mediastinal widening, the lungs are clear bilaterally, no pneumothorax pleural effusion or pneumothorax

Fig. 1. a) Representation of our multimodal cycleGAN framework with exemplar inputs and generated outputs for the image and text modalities. b) Application of the cyclic GAN [14] framework to generate coherent image–report pairs.

We exploit the setting of Cycle-GAN in a multimodal paradigm i.e. the domains in which we work are text (reports) and image (chest x-ray). As shown in Figure 1, our multimodal cyclic GAN architecture comprises (i) two GANs F and G to respectively generate images from textual descriptions and vice-versa, and (ii) two deep neural networks, termed *Discriminators* D_X and D_Y, to respectively compare the generated images and reports against the originals. Figure 1(a) depicts the mappings G and F, while Figure 1(b) depicts how cycle-consistency is enforced to generate coherent image-report pairs.

2.3 Explanatory image–report pairs

Our model learns mappings between prototypical image–text decompositions (termed visual or textual words in information retrieval) akin to the *this looks like that* formulation [18] and synthetic image based explanations in [26]. Since our setting is multimodal instead of image to image setting in cycle-gans, GAN G (report-to-image generator) in our setting is based on a CNN-plus-LSTM based generative model similar to the architecture proposed in [1]. GAN F (image-to-report generation) uses a hierarchical structure composed of two GANs similar to [10]. First, GAN F_1 takes the text embedding as input and generates a low-resolution (64×64) image. The second GAN F_2 utilizes this image and the text embedding to generate a high-resolution (256×256) image.

2.4 Dataset

We used the Indiana University Chest X-Ray Collection (IU X-Ray) [20] for our experiments, as it contains free text reports essential for the report generation task. IU X-Ray is a set of chest x-ray images paired with their corresponding

Table 1. Natural Language Metrics for Generated Reports

Methods:	Ours-Cycle*	Ours-no-cycle	R2Gen [15]	Multiview [25]
BLEU-1	0.486	0.520	0.470	**0.529**
BLEU-2	0.360	**0.388**	0.304	0.372
BLEU-3	0.285	0.302	0.219	**0.315**
BLEU-4	0.222	0.251	0.165	**0.255**
ROUGE	0.440	**0.463**	0.371	0.453

* Reduction in training data (only frontal image-report pairs used)

diagnostic reports. The dataset contains 7,470 images, some of which map to the same free text report. 51% of the images are frontal, while the other 49% are lateral.

The frontal and lateral images map to individual text reports, at times corresponding to the same report. Consequently, mapping reports to images may confound the generator F regarding which type of image to generate. To avoid this confusion, we work only with frontal images, thus reducing the dataset to 3793 image-text pairs. Each report consists of the following sections: impression, findings, tags, comparison, and indication. In this work, we treat the contents of impression and findings as the target captions to be generated. We adopt a 80:20 train-test split for all experiments.

2.5 Implementation

All images were resized to 244 × 224 size. We used 512 × 512 images for initial experiments involving the 'Ours-no-cycle' method (see Table 1) and observed a better performance with respect to natural language metrics. However, low-resolution x-rays were used for subsequent experiments due to computational constraints. The input and hidden state dimensions for Sentence-LSTM are 1024 and 512 respectively, while both are of 512 length in the case of Word-LSTM. Learning rate used for the visual encoder is 1e-5, while 5e-4 is used for LSTM parts. Embedding dimension used for input to the text-to-image framework is 256, with learning rate set to 2e-4 for both the discriminator and the generator. We used PyTorch [5]-based implementations for all experiments.

Firstly, we individually trained the image-to-text and text-to-image generator modules. In the text-to-image part, we first trained the Stage 1 generator, followed by Stage 2 training on freezing the Stage 1 generator. Note that this individual training of the text-to-image module was done on original reports from the training set. However, when we trained the cycleGAN architecture, the text-to-image part took in the generated text as input. While directly training both the modules together, oscillations in loss values were observed.

3 Evaluation

3.1 Evaluation of Generated Reports

We first evaluate the quality of the generated reports via the BLEU and ROUGE metrics [23,24]; we compare our performance against other methods [15,25] in Table 1. Our methods with and without cycle-consistency loss are referred to as *Ours-cycle* and *Ours-no-cycle*. Since only frontal images were used for training *Ours-cycle* (see Section 2.5), the training set is reduced to 3793 image–report pairs. We get comparable performance with the multi-view network [25] based on NLG metrics. There is a small drop in these metrics with the addition of the cycle component, mainly due to the reduction in training data (as the number of image-report repairs is approximately halved).

3.2 Evaluation of Explanations

To evaluate the explanations, we first assess if the generated images truly resemble real input images because the quality of the generated images is also a representative of the quality of the model generated reports as discussed in earlier sections. Secondly, we consider the aspects of trust and faithfulness of our explanation technique based on ideas in [27] for post-hoc explanations.

3.2.1 Evaluating Similarity of Generated Images and Real X-ray Images We quantitatively assess the images using CheXNet [19] (state-of-the-art performance on multi-label classification for chest x-ray images). We use CheXNet on ⟨input image–generated image⟩ pairs for checking the amount of disparity present between the *true* and *generated* images. We achieve a KL-Divergence of 0.101. We also introduce a 'top-k' metric to identify if the same set of diseases are identified from the *input* and *generated* images. The metric averages the number of top predicted diseases which are *common* to both input and the generated images.

$$top - k = \frac{\sum_{All\,pairs}|(top{-}k\,labels(input\,image))\cap(top{-}k\,labels(generated\,image))|}{Number\,of\,pairs}$$

We compare the output labels of CheXNet on both real and generated image using the top-k, Precision@k and Recall@k metrics. From Table 2, on average 1.84 predicted disease labels are common between the input and generated images, considering only the top-two ranked disease labels. In Table 2, we have also shown a comparison against images generated from our text-to-image (report-to-x-ray-image) model on the reports generated by the recently proposed transformer-based R2gen algorithm [15]. Our representative generated images perform better on the top-x, precision and recall metrics, quantitatively showing that the reports generated by our cycleGAN model better describe the input chest x-ray image.

Table 2. Metrics for generated images by using CheXNet for multi-label classification.

k	Top-k (Ours)	Top-k (R2Gen) [15]	Precision@k (Ours)	Precision@k (R2Gen) [15]	Recall@k (Ours)	Recall@k (R2Gen) [15]
2	**1.84**	0.64	**0.92**	0.32	**0.13**	0.05
5	**3.01**	2.55	**0.60**	0.51	**0.21**	0.18
8	**6.45**	5.82	**0.81**	0.73	**0.46**	0.42

Table 3. Accuracy Metric for the Reports Generated from Prototypical (Generated) Images

Label	No Finding	Cardiomediastinum	Cardiomegaly	Lung Lesion	Lung Opacity	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural(E)	Pleural(O)	Fracture	Support Devices
Accuracy	0.78	0.92	0.84	0.96	0.82	0.97	0.96	0.97	0.94	0.98	0.95	0.99	0.96	0.94

3.2.2 Evaluating Trustability of the Explanations We build upon the idea of trust in an explanation technique suggested in [27] for post-hoc explanations. An explanation method can be considered trustworthy if the generated explanations are able to characterize the kind of inputs on which the model performs well or closer to the ground truth. We evaluate our explanations on this aspect of trustability by testing if the explanations or prototypical x-ray images generated are the images on which reports generated are very close to ground truth reports. We evaluate the similarity of the two reports (ground truth reports and reports generated from prototypical images) by comparing the labels output by a naive Bayes classifier on the input reports. The results for accuracy metric for each of the 14 labels is summarised in the Table 3. We can clearly infer that the x-ray images generated as explanations have been able to understand the model's behaviour and hence the good accuracy (around 0.9 for most of the labels).

3.2.3 Evaluating Faithfulness of Explanations Another aspect which has been explored in some of the explanation works is faithfulness of the technique i.e. whether the explanation technique is reliable. Reliability is understood in the sense that it is reflecting the underlying associations of the model rather than any other correlation such as just testing the presence of edges in object detection tasks [12]. We test the faithfulness of the explanations generated by randomising the weights of the report generation model and then evaluating the quality of prototypical images to check if the explanation technique can be called faithful to the model parameters. The metric values for Top-2, Precision@2 and Recall@2 for generated images in this case are 0.90, 0.45 and 0.06 respectively significantly less than corresponding metrics in Table 2. As evident, the prototypical images generated as explanations from randomised weights model are unable to characterize the original input images because the model they are explaining doesn’t contain the underlying information it had previously learnt for characterizing given chest x-ray images.

3.2.4 Qualitative Assessment of Generated Images using Grad-CAM We used GradCAM [4] for highlighting the salient image regions focused upon by the CheXNet [19] model for label prediction from the real and generated image pairs. Two examples are shown in Fig. 2. In the left sample pair, real image shows fibrosis as the most probable disease label, as does the generated image. As observable, the highlighted region showing the presence of a nodule is the same in both x-ray images except for the flip from the left and right lung. This shows that the report generation model was able to capture these abnormalities with great detail, as the report-generated image also captures these details visually. Similarly, two of the top-three labels are the same in both real and generated images as predicted by CheXNet in sample pair 2.

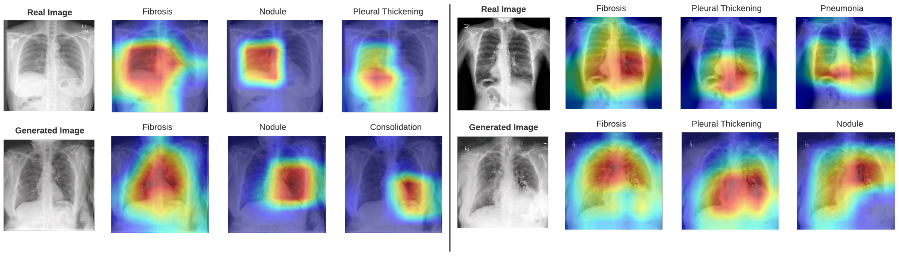


Fig. 2. Grad CAM saliency maps for top 3 predicted labels by CheXNet for real (top row) and generated (bottom row) image pairs; Sample pair 1 (left) and Sample pair 2 (right)

4 Conclusion

A cycleGAN-based framework for explainable medical report generation and synthesis of coherent image-report pairs is proposed in this work. Our generated images visually characterize the text reports, and resemble the input image with respect to pathological characteristics. We have performed extensive experiments and evaluation on the generated images and reports, which show that our report-generation quality is comparable to the state-of-the-art in terms of natural language generation metrics; also the generated images depict the disease attributes both via attention maps and other quantitative measures (precision analysis, trust, and faithfulness) showing the usefulness of a cycle-constrained characterization of chest x-ray images in an explainable medical image analysis task.

References

1. B.Jing, P.Xie, and E.Xing, On the Automatic Generation of Medical Imaging Reports. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018)
2. R. Girshicka, J. Donahuea, T. Darrell, and J. Malik, 2014, rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In CVPR.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, 2012, imagenet classification with deep convolutional neural networks. In NIPS.
4. R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," CoRR, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391
5. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlche-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch:an-imperative-style-high-performance-deep-learning-library.pdf
6. P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," Inf. Sci.,vol. 225, pp. 1–17, 2013. [Online]. Available: https://doi.org/10.1016/j.ins.2012.10.039
7. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient based visual explanations for deep convolutional networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar 2018. [Online]. Available: http://dx.doi.org/10.1109/WACV.2018.00097
8. S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," CoRR, vol. abs/1605.05396, 2016. [Online]. Available: http://arxiv.org/abs/1605.05396
9. T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," CoRR, vol. abs/1711.10485, 2017. [Online]. Available: http://arxiv.org/abs/1711.10485
10. H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," CoRR, vol. abs/1612.03242, 2016. [Online]. Available: http://arxiv.org/abs/1612.03242
11. G. Liu, T. H. Hsu, M. B. A. McDermott, W. Boag, W. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest x-ray report generation," CoRR, vol. abs/1904.02633, 2019. [Online]. Available: http://arxiv.org/abs/1904.02633
12. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," 2020
13. M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.
14. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020.

15. Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," 2020.
@articleMahapatraTMI2021, author = Dwarikanath Mahapatra and Alexander Poellinger and Ling Shao and Mauricio Reyes, title = Interpretability-Driven Sample Selection Using Self Supervised Learning For Disease Classification And Segmentation, journal = IEEE TMI, pages = 1-15, year = 2021
16. D. Mahapatra, A. Poellinger, L. Shao and Mauricio Reyes, "AInterpretability-Driven Sample Selection Using Self Supervised Learning For Disease Classification And Segmentation," in IEEE Trans. Med. Imag., 2021,
17. J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in Computer Vision and Patterm Recognition (CVPR), 2017
18. C. Chen, O. Li, A. Barnett, J. Su, and C. Rudin, "This looks like that: deep learning for interpretable image recognition" CoRR, vol. abs/1806.10574, 2018. [Online]. Available: http://arxiv.org/abs/1806.10574
19. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. Langlotz, K. S. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," CoRR, vol. abs/1711.05225, 2017. Available: http://arxiv.org/abs/1711.05225
20. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc. 2016 Mar;23(2):304-10. doi: 10.1093/jamia/ocv080. Epub 2015 Jul 1. PMID: 26133894; PMCID: PMC5009925.
21. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385
22. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.
23. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," 2002, in Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.
24. C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries." 2004, in Text summarization branches out: Proceedings of the ACL-04 workshop, volume 8. Barcelona, Spain
25. Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 - 21st International Conference, 2018, Proceedings
26. Olah, Chris Mordvintsev, Alexander Schubert, Ludwig, "Feature Visualization", Distill, 2017.
27. Zachary C. Lipton, The Mythos of Model Interpretability, Workshop on Human Interpretability in Machine Learning (WHI 2016)