LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-PROC-779784

Generative Counterfactual Introspection forExplainable Deep Learning

S. Liu, B. Kailkhura, D. Loveland, H. Yong

June 27, 2019

GlobalSIP
ottawa, Canada
November 11, 2019 through November 14, 2019

Disclaimer

Generative Counterfactual Introspection for Explainable Deep Learning

Shusen Liu, Bhavya Kailkhura
CASC, Computation
Lawrence Livermore National Laboratory
Livermore, USA
{liu42, kailkhura1}@llnl.gov

Donald Loveland, Yong Han
MSD, Physical and Life Science
Lawrence Livermore National Laboratory
Livermore, USA
{loveland4, han5}@llnl.gov

***Abstract*—In this work, we propose an introspection technique for deep neural networks that relies on a generative model to instigate salient editing of the input image for model interpretation. Such modification provides the fundamental interventional operation that allows us to obtain answers to counterfactual inquiries, i.e., what meaningful change can be made to the input image in order to alter the prediction. We demonstrate how to reveal interesting properties of the given classifiers by utilizing the proposed introspection approach on both the MNIST and the CelebA dataset.**

***Index Terms*—Explainable Deep Learning, Model Introspection, Counterfactual Reasoning, Generative Adversarial Network**

I. Introduction

The recent success of deep neural networks has lead to many breakthroughs in various application domains [1]–[3]. However, these advances have also introduced increasingly complex and opaque models with decision boundaries that are extremely hard to understand. Despite many recent developments in explainable AI, there are still enormous challenges for explaining deep neural networks. Most existing model introspection approaches [4]–[6] focus on studying the correlation between inputs and outputs (or predictions), e.g., by identifying regions of the input image that most contributed to the final model decision. However, these methods do not consider alternative decisions or identify changes to the input which could result in different outcomes – i.e., they are neither discriminative nor counterfactual [7]. To reliably address some of the most important introspection questions, the ability to reason about causal relationships beyond correlation is necessary.

Knowing causal reasoning behind a prediction is vital in fields such as drug or material discovery [8] where the aim is to map a known value from the output (i.e., property) space back to a set of input experimental parameters. More importantly, from a given input and output data pair, it is useful to understand how the input data could be changed to produce an output closer to their target. These necessary edits to the input data in the form of actionable knobs (implicit or explicit attribute changes) to achieve the desired results can provide a better understanding of complex decision boundaries.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

A promising technique for investigating decision boundaries of a model is based on the prototype and criticism based explanations approach [9]. In this approach, given a query sample, a prototype is defined as a quintessential data sample that best represents the class that the query sample belongs to, while a criticism is the data sample from a different target class which lay closest to the decision boundary. Explainable AI can take advantage of these relationships, as both prototype and criticism examples help build an intuitive understanding of a model and elucidate the necessary changes in the input space to achieve different responses. However, the current prototype and criticism based explanation approaches are not counterfactual in nature and cannot provide actionable feedback. Existing counterfactual explanation techniques [10] are limited to generating criticisms by intervening the original data space. Specifically, they generate criticisms by replacing part of the query image I with specific regions of a 'distractor' image I' that the classifier C predicts as class c'. However, making changes in the original data space (e.g., square tiles of the image) likely will not provide actionable feedback, which is essential for many use cases, e.g., experimental knobs in a scientific application. Furthermore, such changes may not be semantically meaningful and the solution space of potential explanations is restricted by the number of semantically meaningful changes in the original data space.

To overcome these limitations, in this work, we develop a *generative counterfactual introspection framework* to produce inherently interpretable and actionable counterfactual visual explanations in the form of prototypes and criticisms. The counterfactual explanation generation problem is given as follows: *Given a 'query' image I for which a classifier* $\mathbf{C}$ *predicts class c, a counterfactual visual explanation identifies what aspects (or attributes) of I should be changed such that the classifier would output a different target class c' (i.e., the criticism) or provide a more confident classification to c for modified image I' (i.e., the prototype).*

To solve this problem, we propose to employ powerful generative models along with an attribute (or actionable latent feature) editing mechanism [11] to develop *Generative Counterfactual Explanation*: generative and actionable counterfactual explanations generation framework. To the best of our knowledge, this is the first approach exploring the

decision boundaries between classes and their relationship to the input data by providing actionable feedback and generating counterfactual prototypes and criticism based explanations.

II. Related Work

Recently, quite a few model introspection methods have been proposed to allow for interpretability of a given prediction. Many CNN interpretation methods [4]–[6], [12], [13], such as GradCAM [12], utilize backpropagation to conduct sensitivity analysis by attributing the prediction to the input domain (e.g., image pixels). Alternatively, we can build a simpler localized model to approximate the complex nonlinear model [14], [15]. To understand how components of the network work, a variety of the methods have been introduced to visualization the feature (or pattern) the given neuron or layer aim to capture [16]–[18] or examining the representation of the high-level concept in the latent representations [19].

With the pressing need to obtain causal understanding of model behavior, interpretation approaches [20]–[23] focusing on counterfactual reasoning have been proposed. In [20], the counterfactual query is utilized as the fundamental tool for evaluating the fairness of the high impact social application. In the counterfactual visual explanation [22] work, a patch based editing of input image is optimize in order to satisfy the intended changes in the prediction. In the ground visual explanation [23] work, text based explanation are generated to provide counterfactual explanation for image classification task. Beside the causal interpretation methods, as demonstrated in [9], examining the relationship between the trained model and training dataset can also help interpret model behavior.

The safety of deep neural nets have been challenged by the existence of adversarial samples [24]–[26], in which the appearance of small but intentionally worst-case perturbations will lead to change in the prediction. Conceptually, the adversarial examples can also be considered as an answer to a counterfactual query, as it reveals a modification to the input that lead to change of the prediction. However, as the adversarial changes are imperceptible, they cannot reveal the potential bias to humans. We address this problem by utilizing generative adversarial networks (GANs) [11], [27] to generate modification of the input, which ensures a meaningfully edited image rather than an adversarial example.

III. Method

In order to explain a query image with respect to decision boundaries of some trained classifier on image set $\mathcal{I}$, we aim to produce counterfactual prototypes and criticisms. Next we formalize this problem and then present our solution.

A. Minimal Change Counterfactual Example Generation

Given a query image I for which the classifier C predicts class c, we seek to identify the key attribute changes in I such that making these changes in I would lead the network to either change its decision about the query to the target class (i.e., criticism) or make it more confident about the query class. We consider both of these following cases: 1) attributes are known and given for $\mathcal{I}$, or 2) attributes are unknown in which case will be learned from $\mathcal{I}$. Furthermore, these attributes are expected to be actionable, i.e., we should be able to change these attributes and generate corresponding changes in the query image. To enable this, we employ a powerful generative machine learning model called "generative adversarial network (GAN)" [27]. GANs transform vectors of generated noise (or latent factors) into synthetic samples resembling data gathered in the training set. GANs (and corresponding latent space) are learned in an adversarial manner, i.e., a concept taken from the game theory which assumes two competing networks, a discriminator D (differentiating real vs. synthetic samples) and a generator G (learning to produce realistic synthetic samples by transforming latent factors). This adversarial learning is shown to learn salient attributes of the data in an unsupervised manner which can later be manipulated using the generator G. GANs can also be used for simultaneously generating and manipulating the images with known and desired attributes [11]. We use both of these formulations in our framework depending on whether actionable attributes are known or unknown, where the latter uses the latent representations as our attributes.

Generative editing models are denoted as $G(I; A)$ or $G(I; L_o)$ depending on whether actionable attributes are known or unknown respectively. The goal is to manipulate single or multiple attributes $A = \{a_1, \cdots, a_N\}$ of an image I, i.e., to generate a new image I^* with desired attributes $\{a_1^*, \cdots, a_M^*\}$ while preserving other details $\{a_{M+1}, \cdots, a_N\}$, or to manipulate a latent vector L_o in a similar fashion. Given these generative editing mechanism, we formulate minimal change counterfactual explanation generation problem given image I, image attribute A, and a target attribute vector A', where L_o and L'_o can be used in place of A and A', as follows:

$$\begin{aligned} \min_{A'} \quad & \|I - I(A')\|_p \\ \text{s.t.} \quad & c' = C(I(A')) \\ & I(A') = G(I; A') \end{aligned} \tag{1}$$

where $p = 1$ and c' is the target criticism class. When the goal is to generate prototypes, we set $c' = c$ as the original class label of the query image and formulate an alternating loss function to promote solution which maximize class confidence instead of having a trivial solution, i.e., $A' = A$.

B. Approximate Solution

Most deep neural network based models make formulation (1) non-linear and non-convex, making it hard to find a closed-form solution. Thus, we formulate a relaxed version of this optimization problem which can be solved efficiently using gradient descent algorithms. The proposed approach relaxes the optimization problem 1 as follows:

$$\min_{A'} \quad \lambda \cdot loss_{C,c'}(I(A')) + \|I - I(A')\|_p \tag{2}$$

where loss $loss_{C,c'}$ is cross-entropy loss for predicting image $I(A')$ to label c' using classifier C. Note that both classifier C and generator G are differentiable. The gradient of the objective function is computed by back-propagation, and the

minimal change counterfactual example generation problem is solved using gradient descent. Furthermore, to generate an explanation with minimum change $\delta = \|I - I(A')\|_p$, one can repeatedly solve this optimization problem using gradient descent, continually updating λ using bisection search or any other method for one-dimensional optimization.

IV. Experiments

Here we demonstrate the effectiveness of the proposed counterfactual explanation generation approach on two datasets (one with known attributes and another one with unknown). The proposed method outputs modified images to satisfy counterfactual queries along with actionable attribute values to achieve these results, in turn, providing a comprehensive understanding of decision boundaries of the classifier C.

A. MNIST dataset [28]

In this experiment, we consider the problem of classifying a given image of a handwritten digit into one of 10 classes (0 to 9). We use the MNIST dataset which contains 60,000 training and 10,000 test images of handwritten digits. The classifier [29] is trained on MNIST training set and achieves 99.10% accuracy on the test set. We utilize a pretrained DCGAN architecture [30] (with a 10D latent space) as our image generator. Given 10D latent vector (L_o), the generator produces a digit image. The proposed optimization method will update the L_o to generate meaningful modification of the image that answers the counterfactual query.

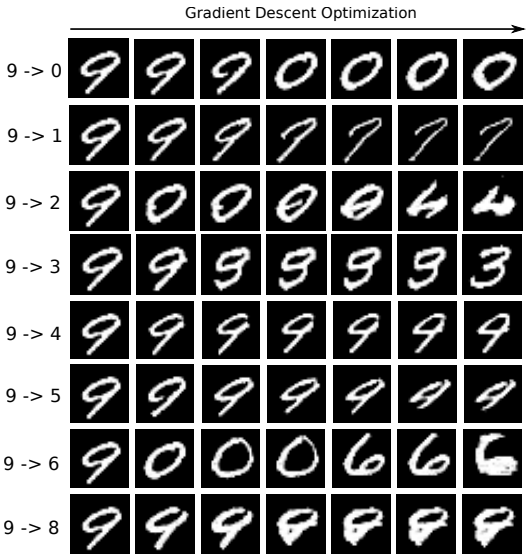

Fig. 1: Finding criticism of the digit 9 class.

As shown in Fig. 1, we illustrate meaningful changes to the image of digit 9 to alter its prediction. We start from the same image in each row and illustrate the optimization path from the original image to the images that altered the classified label to a predefined target label. Compared to a direction optimization in the image space [24] that leads to an adversarial example, the utilization of a GAN guarantees that we end up exploring the "manifold" of all possible meaningful images. As a result, these edits provide us with valuable insights regarding classifier decision boundary, i.e., what are the boundary image patterns between different classes of digits, and what kind of changes are most likely to alter the prediction. Interestingly, we see that for certain target labels (9->2, 9->6), the image first change to a different digit (in this case 0) before morphing into the target digits. Alternatively, as shown in Fig. 2, we also utilize a similar optimization to find the *prototype* for each digit by"walking" toward the center of the class on the digit image manifold. We can see the starting digits morphed into a more "regular" handwriting style, which are easier for human to recognize. These observations not only help in revealing the inherent structure of the digit image manifold but also indicate the preference of the classifier regarding similarity between digits.

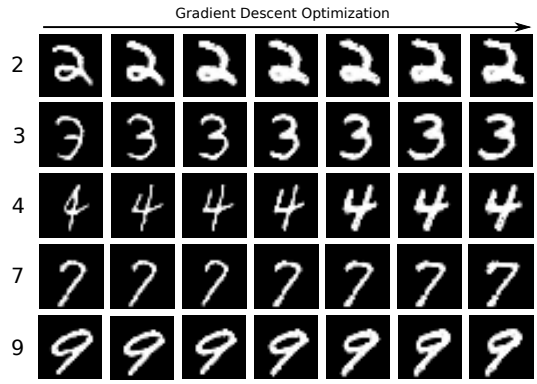

Fig. 2: Finding prototypes of different digits.

B. CelebA dataset [31]

In several case, attributes are known explicitly, thus, optimization can be carried out in the attribute space A that a generator is conditioned on ($I' = G(I; A)$), where explicitly defined physical attributes can provide actionable feedback. We use CelebFaces Attributes dataset (CelebA) which is a large-scale face attributes dataset with more than 200K celebrity images, each with explicit attribute annotations. We consider a classification problem of classifying a celebrity face image in CelebA dataset into young or old. The classifier [11] is trained on CelebA dataset and achieves average accuracy of 90.89% on CelebA testing set. Next, we use the AttGAN [11] as our generative editing method to generate modification to the query face image. The AttGAN can make edits to the original query image I based on additional attributes (e.g., hair color, glass, bang, bald).

Since the AttGAN generator has a young/old input attribute, a direct optimization in the entire attribute space will likely lead to the degenerate case, in which the young/old attribute is used to edit the image (to make it appears older for the classifier). Therefore, in our experiment, we fixed the young/old attribute to the original label and only make changes to rest of the attributes (12 in total). In other words, we ask what kind of attributes changes (beside the young/old attribute) will make a given image appear older or younger for the given classifier.

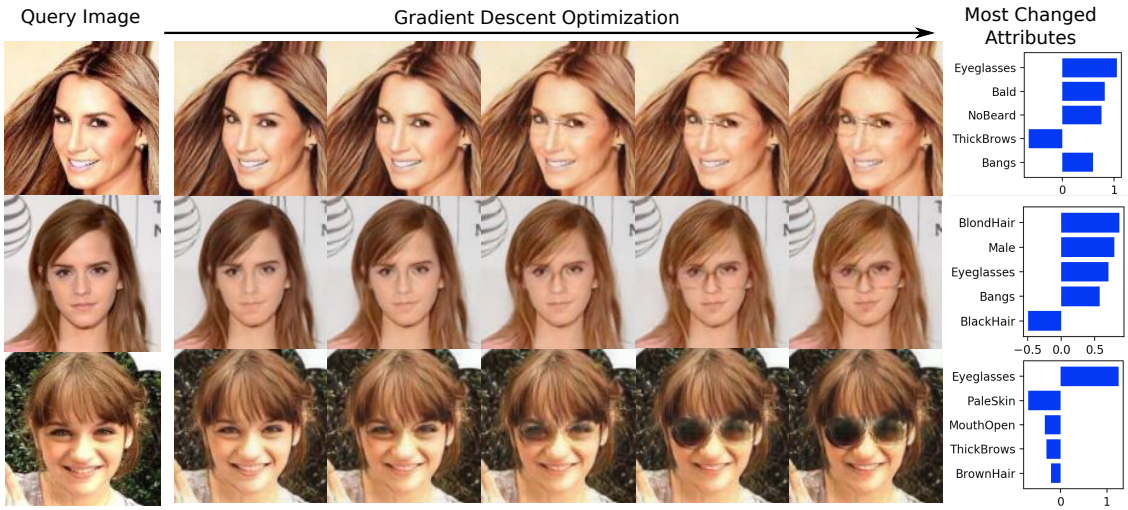

Fig. 3: Illustrate attributes changes (beside the young/old attribute) that will make the image appears older for the classifier.

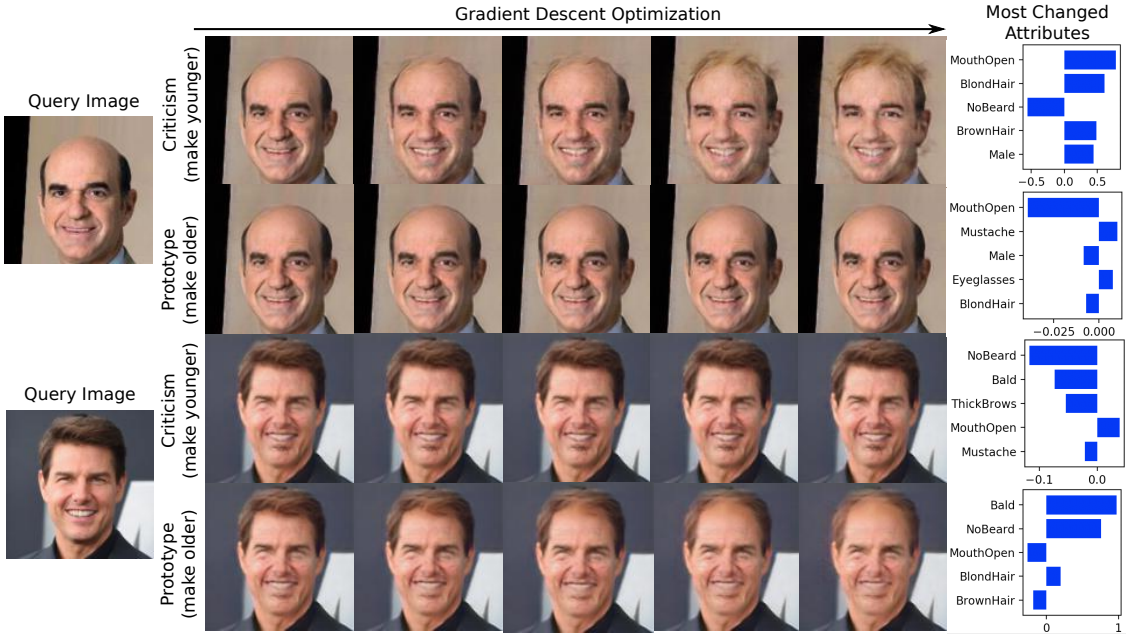

Fig. 4: Prototype and criticism for the images with ground truth label “old”.

In Fig. 3, we have three female celebrity faces (query images) which are classified as “young”. Here, we show the optimization path that eventually leads to an “old” classification. The right most column shows the top five most changed attributes and their relatively changes. This result is particularly interesting as all three examples show eyeglasses in the modified images that result in an “old” classification. One possible explanation for such an observation is that the classifier learns these patterns from the training data (the model may try to obtain better accuracy by picking up features that are not aligned with human perception [32]). To investigate this hypothesis, we explore the distributions of attributes across the training data, where we notice a clear difference regarding eye glass frequency between the young and old population. This result demonstrate that counterfactual query can be an extremely powerful tool to reveal unexpected behaviors of classifiers and highlight the potential bias in the training data and model.

To further illustrate how counterfactual examples help explain the behavior of the classifier, in Fig. 4, we investigate the prototype and criticism examples for two male celebrity faces that both have a ground truth label “old”. When searching for the prototypes (i.e., making them older), we see a minimal changes for the first face (second row) while observe significant change for the second face (fourth row). This distinction indicates that the first person seems to have a prototypical look for the “old” class, whereas the second person does not. For the criticisms (row one and three), the opposite holds true, which indicates the image of the second person is an outlier

for "old" samples, and is closer to a typical "young" image. Finally, the right most column provides "actionable insights" to achieve these changes. The top five most changed attributes are reasonable with hair features being most important factors in discriminating the age group.

V. Discussion and Future Work

In this work, we present preliminary results on utilizing generative models to obtain counterfactual explanations for a given classifier. Despite the simplicity of the optimization, we demonstrate that the effectiveness of the proposed approach for revealing insights regarding the behavior of deep neural network models. For future directions, we plan to explore the potential application of such interpretation method for scientific application, where explainability are essential for model validation and domain discovery.

Acknowledgement

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Review and release under LLNL-PROC-779784.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, vol. 5, p. 4308, 2014.

[3] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular systems biology*, vol. 12, no. 7, p. 878, 2016.

[4] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[5] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[6] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[7] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.

[8] Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," *Journal of Materiomics*, vol. 3, no. 3, pp. 159–177, 2017.

[9] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems*, 2016, pp. 2280–2288.

[10] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," *CoRR*, vol. abs/1904.07451, 2019. [Online]. Available: http://arxiv.org/abs/1904.07451

[11] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, 2019.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[15] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy, "Treeview: Peeking into deep neural networks via feature-space partitioning," *arXiv preprint arXiv:1611.07429*, 2016.

[16] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks," *arXiv preprint arXiv:1602.03616*, 2016.

[17] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[18] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, p. e7, 2017.

[19] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," *arXiv preprint arXiv:1711.11279*, 2017.

[20] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[21] T. Narendra, A. Sankaran, D. Vijaykeerthy, and S. Mani, "Explaining deep learning models using causal inference," *arXiv preprint arXiv:1811.04376*, 2018.

[22] J. E. D. B. D. P. S. L. Yash Goyal, Ziyan Wu, "Counterfactual visual explanations," in *ICML*, 2019, pp. 264–279.

[23] L. Anne Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 264–279.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[25] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, 2019.

[26] T. A. Hogan and B. Kailkhura, "Universal hard-label black-box perturbations: Breaking security-through-obscurity defenses," *arXiv preprint arXiv:1811.03733*, 2018.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[28] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[29] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[30] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[32] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," *arXiv preprint arXiv:1805.12152*, 2018.