Formal Explainable AI and Enhancement of Image Attributes

R & D Project I

by

Palle Bhavana

(Roll No. 210050111)

Under the Supervision of

Krishna S.

Ashutosh Gupta

Department of Computer Science and Engineering

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

Mumbai - 400076, India

December, 2023

Contents

Contents

1 Introduction 1
1.1 Introduction . . . 1
1.1.1 Background . . . 1
1.1.1.1 Formal Explanations . . . 1
1.1.1.2 Minimal and Minimum Explanations . . . 2
1.1.1.3 Generative Adversarial Networks (GANs) . . . 2
1.1.1.4 Attribute Vectors . . . 3
1.2 Problem Statement . . . 3

2 Literature Review 5
2.1 Contrastive Examples . . . 5
2.1.1 Essential Lemmas . . . 5
2.2 Main Algorithms . . . 6
2.2.1 Upper Bounding Thread (TUB) . . . 6
2.2.2 Lower Bounding Thread (TLB) . . . 6
2.2.3 Regressor Model and Image Editing Algorithm . . . 6

3 Ideation and Experimentation 9
3.1 Approach 1 . . . 9
3.1.1 Idea Formulation . . . 9
3.1.2 Detailed Steps . . . 9
3.2 Approach 2 . . . 12
3.2.0.1 Detailed Steps . . . 13
3.2.0.2 Loss Functions . . . 14

4 Conclusions and Practical Implications 15
4.1 Conclusion . . . 15
4.1.1 Approach 1 . . . 15
4.1.2 Approach 2 . . . 16

Chapter 1

Introduction

1.1 Introduction

Explaining the decisions made by Deep Neural Networks (DNNs) is a critical aspect of ensuring transparency and trust in any application. In this project, our primary focus revolves around exploring explainable artificial intelligence (XAI), with a specific emphasis on formal explanations within the context of classification problems. The fundamental definitions guiding our exploration are rooted in the concept of Formal Explanations, particularly addressing why a DNN classifies a given input as a specific label.

In the subsequent stage, our attention turns to developing techniques for modifying images by manipulating specific attributes associated with them. We will introduce methodologies employing Generative Adversarial Networks (GANs) to achieve this objective.

1.1.1 Background

1.1.1.1 Formal Explanations

In the context of classification problems, where a model is trained to predict labels for given inputs, a classification problem is represented as a tuple $<F, D, K, N>$. Here, $F = \{1, ..., m\}$ denotes the features, $D = \{D1, D2, ..., Dm\}$ denotes the domains of features,

$K = \{c1, c2, ..., cn\}$ is a set of classes, and $N : F \to K$ is the classification function (often a neural network). A classification instance is denoted as (v, c), where $v \in F$, $c \in K$, and $c = N(v)$. Given an input v with classification $N(v) = c$, an explanation (also known as an abductive explanation or an AXP) is a subset of features $E \subseteq F$, such that:

$$\forall(x \in F).[\bigwedge_{\forall i \in E} (x_i = v_i) \to (N(x) = c)]$$

1.1.1.2 Minimal and Minimum Explanations

A subset $E \subseteq F$ is a minimal explanation of instance (v, c) if it ceases to be an explanation if even a single feature is removed from it:

$$\begin{aligned}&(\forall(x \in F).[\wedge_{i \in E}(x_i = v_i) \to (N(x) = c)])\\ &\wedge (\forall(j \in E).[\exists(y \in F).[\wedge_{i \in E \setminus j}(y_i = v_i) \wedge (N(y) \neq c)]])\end{aligned}$$

A minimum explanation is defined as a minimal explanation of minimum size; i.e., if E is a minimum explanation, then there does not exist a minimal explanation $E'E$ such that $|E'| < |E|$.

1.1.1.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) belong to a class of artificial intelligence algorithms employed in unsupervised machine learning. GANs consist of two essential components: a generator (G) and a discriminator (D). These networks operate simultaneously through adversarial training to produce data, such as images, that closely resembles real data.

The primary goal of GANs is to strike a balance between the generator producing data indistinguishable from real data and the discriminator accurately identifying whether the provided data is real or generated.

In the context of this project, GANs are leveraged to alter images by manipulating specific attributes. The generator is controlled to modify attribute vectors, enabling the generation of images with desired alterations.

1.1.1.4 Attribute Vectors

Attribute vectors are numerical representations associated with specific features or characteristics of data. In the context of this project, attribute vectors are employed to describe and control attributes of images. An attribute vector for an image (x) is denoted as $\text{AttributeVector}(x) = [a_1, a_2, ..., a_n]$, where each a_i represents a specific attribute.

The attribute vectors play a crucial role in altering images using Generative Adversarial Networks (GANs). A regressor model is utilized to obtain the attribute vector for an image, and this vector, along with random noise, is fed into the GAN's generator. The generator then alters the attribute vector, resulting in the modification of the image's attributes.

1.2 Problem Statement

Our primary focus in this project is to validate the methods and implementations designed for computing explanation sets. The central objective is to analyze the behavior of a given image when subjected to alterations without changing its explanation set. In essence, we aim to investigate whether an image, when modified, not only retains the same explanation set as the original but also maintains its classification. Specifically, we inquire whether two images with identical explanation sets will be classified similarly or diverge under the influence of alterations. To achieve this, we propose and explore various methods that can be employed in the project.

Chapter 2

Literature Review

We reviewed two papers in the beginning to learn about how to obtain explanations for an image. These papers were (1; 2). Here are the important aspects discussed in relation to obtaining explanations:

2.1 Contrastive Examples

A subset of features $C \subseteq F$ is called a contrastive example or a contrastive explanation (CXP) if altering the features in C is sufficient to cause the misclassification of a given classification instance (v, c):

$$\exists(x \in F).[\wedge_{i \in F \setminus C}(x_i = v_i) \wedge (N(x) \neq c)]$$

2.1.1 Essential Lemmas

We present two essential lemmas related to contrastive examples:

Lemma 1. Every contrastive singleton is contained in all explanations.

Lemma 2. All explanations contain at least one element of every contrastive pair.

2.2 Main Algorithms

2.2.1 Upper Bounding Thread (TUB)

The TUB Algorithm, aims to find a minimal explanation by iteratively removing features. It uses a verification query for each feature, converging to a minimal explanation. To improve efficiency, a heuristic model prioritizes the removal of less important features, potentially leading to smaller explanations. it needs not to converge to a minimum explanation.

2.2.2 Lower Bounding Thread (TLB)

TLB identifies contrastive singletons and pairs to establish a lower bound on the size of the minimum explanation. It efficiently discovers singletons by leveraging the sensitivity of neural networks to adversarial attacks. The algorithm extends its search to contrastive pairs, excluding those with contrastive singletons. The lower bound is computed as the sum of singleton count and the size of the minimum vertex cover for pairs. This approach, adaptable to larger contrastive examples, offers effective approximations without incurring substantial computational costs.

We then explored an additional paper that leverages Generative Adversarial Networks (GANs) and latent spaces for image attribute manipulation. This work, referenced as (3), delves into the application of GANs to alter specific attributes within an image using latent vectors. The methodology involves the use of a regressor model in conjunction with attribute vectors to achieve image alterations.

2.2.3 Regressor Model and Image Editing Algorithm

In the investigated paper (3), a regressor model is employed to manipulate images effectively. The regressor utilizes attribute vectors to guide the alteration of specific features within an image. The latent vector representation is transformed by adding a perturbation, defined as $T\epsilon = \sum_{i=1}^{N} \epsilon_i d_i$, to achieve the desired attribute changes.

The main algorithm for image editing involves transforming the latent vector using the attribute vectors and regressor. This process aims to generate images with modified attributes while maintaining overall visual coherence.

This approach showcases the capability of GANs and latent spaces in achieving controlled and interpretable image editing, providing a foundation for further exploration and advancements in image attribute manipulation.

Chapter 3

Ideation and Experimentation

3.1 Approach 1

3.1.1 Idea Formulation

The primary objective of Approach 1 is to generate altered images from an original image by synergizing Generative Adversarial Networks (GANs) with an existing implementation designed for extracting the explanation set of an image. The foundational concept involves acquiring the explanation set of an image and utilizing it, alongside random noise, as inputs for training a GAN. The ultimate goal is to train the GAN to produce images that not only share similar explanations but also exhibit visual similarity to the original image. Validation of the generated images involves comparing their explanations and visual appearance using a discriminator.

3.1.2 Detailed Steps

Explanation Set Extraction:

- Employ the existing implementation to extract the explanation set of a given original image.

- The explanation set represents a subset of features crucial for the classification of the image.

GAN Training Setup:

- Prepare a GAN architecture comprising a generator(G) and a discriminator(D).
- The generator takes random noise and the extracted explanation set as input to generate altered images.

GAN Training Objective:

- Train the GAN to produce images that align with the explanation set and visually resemble the original image.
- Utilize a combination of loss functions, including Explanation Set Loss, Adversarial Loss, and Diversity Regularization Loss.

Validation Process:

- Generate images using the trained GAN and extract their explanation sets.
- Employ the discriminator to compare the generated images with the original image.
- Ensure that the GAN does not produce identical images and that the generated images maintain visual similarity to the original.

Fine-Tuning and Optimization:

- Iteratively fine-tune GAN parameters to enhance quality of generated images.
- Optimize the training process to balance diversity and fidelity in the altered images.

Results Analysis:

- Evaluate results by comparing explanation sets and visual similarity between generated and original images.
- Validate the effectiveness of the GAN in producing diverse and visually similar altered images.

This approach leverages the power of GANs to generate altered images while aligning with the original image's explanation set, providing a controlled and interpretable means of image transformation.

Loss Functions:

The training objective incorporates the following loss functions:

- **Explanation Set Loss**: The Explanation Set Loss is a custom loss function designed to ensure that the generated images align with the critical features identified in the explanation set of the original image. It penalizes deviations from the explanation set, guiding the GAN to produce images that reflect the essential characteristics for classification. Mathematically:

$$L_{\text{explanation}} = \sum_{i=1}^{n} (E_{\text{gen}}[i] - E_{\text{original}}[i])^2$$

- **Adversarial Loss**: Adversarial Loss is fundamental for training the GAN to generate realistic images. Implemented with binary cross-entropy, this loss function encourages the generator to produce images that are indistinguishable from real images according to the discriminator. It enhances the overall realism of the generated images. Mathematically, Adversarial Loss ($L_{\text{adversarial}}$) is commonly implemented using binary cross-entropy. For a given discriminator output (D), target label (y), and the generated image ($G(z)$), the loss is calculated as:

$$L_{\text{adversarial}} = -\frac{1}{N}\sum_{i=1}^{N}(y \cdot \log(D(G(z))) + (1-y) \cdot \log(1 - D(G(z))))$$

where $G(z)$ is the generated image, and z is the random noise input.

- **Diversity Regularization Loss**: The Diversity Regularization Loss ($L_{\text{diversity}}$) penalizes similarities between generated samples. Let I_1 and I_2 be two generated images. The loss is defined as the negative cosine similarity between their feature representations:

$$L_{\text{diversity}} = -\frac{\text{cosine\_similarity}(f(I_1), f(I_2)) + 1}{2}$$

Here, $f(\cdot)$ represents the feature extraction function, and `cosine_similarity` calculates the cosine similarity between the feature vectors.

The combined loss function is designed to strike a balance between fidelity to the original image, alignment with the explanation set, and diversity in the generated images.

Unfortunately, due to a lack of communication from the university stakeholders and the unavailability of access to the complete implementation, we encountered a significant constraint.. The absence of complete access to the necessary resources hindered our ability to proceed with the detailed execution of this approach. So, we propose another approach.

3.2 Approach 2

In this approach, we explored the idea of utilizing Generative Adversarial Networks (GANs) exclusively for image alteration. While the alteration precision may not meet the previous expectations, we decided to use the concept of attribute vectors. The overarching idea involves training a regressor to output attribute vectors for an image. Subsequently, these vectors, along with random noise, are fed into the GAN. The GAN's task is to modify the attribute vector to fine-tune and alter the attributes of the image.

3.2.0.1 Detailed Steps

1. **Problem Statement**:
    - We consider controllable semantic image editing via latent space navigation in GANs.
    - The goal is to discover semantically meaningful latent-space GAN directions to manipulate the attributes of synthetic images.
2. **Proposed Approach**:
    - Utilize a GAN model comprising a generator (G), discriminator (D), and a pre-trained regressor (R).
    - Discover latent-space GAN directions (T) to manipulate attributes with an assigned step size (ϵ).
    - Transform the latent vector z by adding

    $$T\epsilon = \sum_{i=1}^{N} \epsilon_i d_i$$

        – T represents the set of latent-space GAN directions, where each d_i is a specific direction

        $$T = \{d_1, d_2, ..., d_N\}$$

        – ϵ is a vector containing step sizes for each direction in T

        $$\epsilon = \{\epsilon_1, \epsilon_2, ..., \epsilon_N\}$$

        – ϵ_i represents the step size assigned to the i-th direction

        – The summation $\sum_{i=1}^{N} \epsilon_i d_i$ implies that the latent vector z is transformed by adding a weighted combination of the directions in T, where the weights are determined by the step sizes ϵ_i
    - Evaluate the recovered and edited images using the regressor R to predict attribute values.

3. **Objective Function**:

$$\min_{T} L, \lambda_1 L_{\text{reg}} + \lambda_2 L_{\text{disc}} + \lambda_3 L_{\text{content}}$$

where L_{reg}, L_{disc}, and L_{content} represent regression, discriminator, and content loss, respectively.

3.2.0.2 Loss Functions

1. **Regression Loss (L_{reg})**:

$$L_{\text{reg}} = -\hat{\alpha} \log \alpha' - (1 - \hat{\alpha}) \log(1 - \alpha')$$

where α' is from the distribution generated by z and ϵ.

2. **Discriminator Loss (L_{disc})**:

$$L_{\text{disc}} = \log(1 - D(G(z')))$$

where Z' is conditioned on z.

3. **Content Loss (L_{content})**:

$$L_{\text{content}} = \left\| F(G(z')) - F(G(z)) \right\|^2$$

where $F(\cdot)$ denotes a feature function, and D_{content} indicates the layers used as features.

Chapter 4

Conclusions and Practical Implications

4.1 Conclusion

In this project, we explored two intriguing strategies for image alteration, grounded in theoretical frameworks leveraging Generative Adversarial Networks (GANs) and attribute vectors. While practical experiments were not conducted due to unforeseen constraints, the theoretical foundations of these approaches suggest promising avenues for controllable and interpretable image transformations.

4.1.1 Approach 1

Integrating GANs with explanation sets to generate altered images exhibited substantial potential. The envisioned workflow, involving the extraction of explanation sets and subsequent GAN training, presents a novel approach to ensure altered images not only retain visual fidelity but also align with the original image's explanation set. The incorporation of diverse loss functions, including perceptual and adversarial terms, promises a comprehensive strategy for achieving both diversity and fidelity in the generated images.

4.1.2 Approach 2

The exclusive use of GANs, guided by the manipulation of attribute vectors through a regressor, provides an alternative avenue for image alteration. Despite lacking empirical validation, the method's reliance on attribute vectors and latent space navigation offers a controllable means for semantic image editing. The proposed objective function, integrating regression, adversarial, and content losses, highlights the importance of maintaining image identity during transformations.

While practical experimentation was hindered, the conceptualization of these strategies lays the groundwork for future exploration. These theoretical frameworks open doors to further investigations into interpretable image transformations, controllable attribute edits, and the interplay between GANs and explanation sets.

In conclusion, the strategies presented here contribute valuable insights into the potential realms of controllable image alteration. Future work should focus on practical implementation, empirical validation, and the refinement of proposed frameworks for a deeper understanding of their efficacy in real-world scenarios.

Bibliography

[1] Shahaf Bassan and Guy Katz, *Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks*, Journal Name, 2023,
The Hebrew University of Jerusalem, Jerusalem, Israel, https://arxiv.org/abs/2210.13915,

[2] Shahaf Bassan1, Guy Amir1 , Davide Corsi2, Idan Refaeli1, and Guy Katz1 *Formally Explaining Neural Networks within Reactive Systems*, 2023, The Hebrew University of Jerusalem, https://arxiv.org/abs/2308.00143

[3] Peiye Zhuang, Oluwasanmi Koyejo, Alexander G. Schwing, *ENJOY YOUR EDITING: CONTROLLABLE GANS FOR IMAGE EDITING VIA LATENT SPACE NAVIGATION*, ICLR 2021, University of Illinois https://arxiv.org/abs/2102.01187

[4] https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/