Generating User-Friendly Explanations for Loan Denials Using Generative Adversarial Networks

Ramya Malur Srinivasan Ajay Chander

There are increasing competitive and regulatory incentives to deploy AI mindfully within financial services. An important aspect towards that end is to explain AI decisions to various stakeholders. State-of-the-art explainable AI systems mostly serve AI engineers and offer little value to other stakeholders. Towards addressing this gap, we built the first dataset that is representative of loan-applicant friendly explanations. We also designed a novel generative adversarial network (GAN) that can accommodate smaller datasets to generate user-friendly purpose-driven explanations. In this article, we demonstrate how our system can generate explanations serving multiple purposes, including those that help educate loan applicants and those that help them take appropriate action towards a future approval.

1. Introduction

From customer behavior prediction to identity verification, AI is being deployed within a wide variety of Fintech applications [1]. Amidst this widespread business adoption of AI, customers, policymakers, and technologists are getting increasingly concerned about the blackbox aspects of AI. For example, with an AI-based credit scoring system, in markets where credit risk scoring models are regulated, there is a strong requirement for the models to be explainable [2].

Recently, there have been several initiatives from both the government [3] and a diverse set of industries [4, 5] to make AI explainable and hence trustworthy. Yet, most state-of-the-art methods provide explanations that mostly target the needs of AI engineers [6, 7]. Thus, there is an increasing need for creating AI systems that can explain their decisions to a large community of users.

Effective explanations serve a variety of purposes. They help build user trust and help make AI systems more robust. They also help educate decision makers and broaden their awareness in choosing appropriate actions.

With the above as context, we considered the use case of AI-based loan decisions, in particular loan denials. We collected and built a dataset representative of user-friendly explanations. We then designed a machine-learning system that can generate such explanations. This system can generate explanations serving different purposes such as those that help in educating loan applicants and those that help the applicants in taking appropriate action towards a future approval. These features enhance the trustworthiness of the AI systems.

2. Related work

In this section, we review related efforts with emphasis on the finance industry. We then review related works in explainability from an AI engineer's and end-user's perspectives.

2.1 Explainable AI efforts in financial industry

The new European General Data Protection Regulation (GDPR), ISO/IEC 27001, and the U.S. Defense Advanced Research Projects Agency's explainable AI (XAI) program [3] are a number of note-worthy governmental initiatives towards XAI. In parallel, several industry groups are looking to address issues concerning AI explainability. FICO, a credit analytics company, recently released an XAI toolkit [8] to outline part of the explainability support for their machine learning. The momentum for XAI is only expected to grow in the near future.

2.2 Explanations for AI engineers

A nice summary concerning explainability from an AI engineer's perspective is provided in [9] and [10]. However, most AI explanations are for AI engineers [4] and are not useful to non-AI experts in either understanding the AI's decision or in debugging the model [11]. In [12], on the other hand, the authors discuss the main factors used by the AI system in arriving at a certain decision and also discuss how changing a factor changes the decision. This kind of explanation helps in debugging for the AI engineers. While impressive in helping an AI engineer, these works are not accessible to a wider set of users.

2.3 Explanations for End-users

More recently, there have been several efforts in understanding the human interpretability of AI systems. The authors in [13] provided a taxonomy for human interpretability of AI systems. An effective perspective of user-centered explanations is provided in [14] and [15], wherein the authors emphasize the need for persuasive explanations. The authors in [16] explore the notion of interactivity from the viewpoint of the user. In [17], the authors discuss how humans understand explanations from machine learning systems through surveys. Interpretability performance is measured in terms of time to response and the accuracy of the response. While these efforts are significant in quantifying human interpretability, they fall short of generating user-friendly explanations, which is the focus of this work.

3. Dataset building

Due to the non-availability of any dataset that is representative of user-friendly explanations, we built the first ever dataset in this regard. We refer to the dataset as "*X-Net*."

To build X-Net, we ran a survey on Amazon Mechanical Turk (MTurk) [18] wherein we provided MTurk workers with a loan application scenario and asked them to imagine that they were the loan applicants.

These workers were provided with textual descriptions highlighting reasons for loan denials. These descriptions were then edited for syntactic and semantic correctness. Furthermore, linguists also provided annotations for each description with corresponding broad and specific reasons for each loan denial. For example, a broad reason could be "job" and specific reasons could be "no job," "unstable job," "limited job history," "no job history," and "unstable job history." This resulted in a curated dataset of 2,432 sentences with their corresponding broad and specific reasons.

As a result of analyzing these reasons, the number of unique ones turned out to be less than a hundred. Thus, we observed that the set of user-friendly explanations is a rather small set. Furthermore, we observed that the reasons provided by MTurk workers seldom appeared as features in machine-learning datasets. The most frequent set of broad reasons included credit, job, income, and debt. There were a few others that were mentioned in small numbers such as failed background check, incomplete applications, etc.

Next, to generate explanations serving different purposes such as those that educate loan applicants and those that help in taking appropriate actions in the financial domain, we curated a dataset consisting of 2,432 sentence pairs corresponding to two different purposes of education and action. This was done in collaboration with a linguist (Ph.D. in Rhetoric). We refer to this dataset as the "Extended X-Net." A sample education-action explanation pair from Extended X-Net is provided below:

"The record of finances associated with this application suggests that there is a record of outstanding loan payments" (educates).

"Please complete all remaining loan payments before applying for a new loan" (suggests action).

More details about the dataset can be found in [19].

4. Method of generating user-friendly examples

Our first goal is to automatically generate user-friendly explanations such as those in X-Net. Furthermore, to ensure that explanations can be generated in a controlled manner, we want conditional generation, i.e., the generated explanation should match with a specified reason for a loan denial. Our second goal is to generate explanations serving different purposes, such as those that educate and those that help in taking an action.

Motivated by the recent success of GANs, we designed a conditional GAN to address the

aforementioned goals. In general, a GAN can generate different kinds of data such as images and text. A GAN consists of two neural networks; a generator and discriminator (**Figure 1**). The generator takes in noise as the input and generates the data that is then predicted by the discriminator as being real or fake.

In a conditional GAN, certain conditions can be specified to the generator to guide the data generation as per the application requirements. We also incorporate the GAN model with a style transfer mechanism to generate explanations serving different purposes.

However, the biggest challenge we face is the problem of limited training data (2,432 sentences with 100 unique reasons). Below we describe the main aspects of our solution strategy. Reference [20] includes more details.

To address the first goal of conditional explanation generation using limited training data, we use the adversarially regularized autoencoder (ARAE) [21] architecture. In particular, we propose three modifications to the architecture as listed below.

1) We consider a mixture of normal distributions to model the underlying conditions of the GAN model [22]. Since mixture models can approximate arbitrarily complex distributions given a sufficiently large number of normal components, by using a mixture of normal instead of the conventional distribution, we can improve the expressiveness of the model.
2) We incorporate conditions in a hierarchical manner. The motivation comes from the fact that children learn from very limited data in a hierarchical manner [23]. Thus, we incorporate conditions for loan denial on the basis of a broad reason (e.g. income) and specific reason (e.g. unstable income) as a two-level conditioning scheme.
3) To generate more relevant sentences, we introduce two new loss functions called labeler and anti-labeler loss [24] to classify the real and generated sentences on the basis of their reasons, respectively.

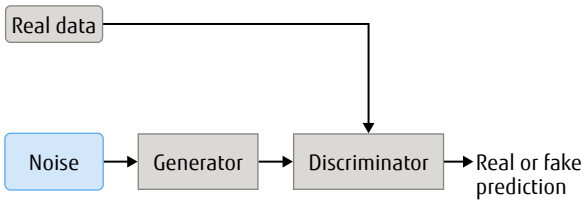

Figure 1
Block diagram of general GAN system architecture.

Figure 2 shows the block diagram of our proposed architecture that is built on top of the ARAEGAN architecture. For details please refer to [19].

The goal here is to generate explanations serving different purposes, namely those that educate the loan applicant about the denial and those that help them to take future actions. Limited training data poses a big challenge in this task as well. We considered pairs of explanations corresponding to "education" and "action" in the training data and use the ARAEGAN model along with a mixture of Gaussian noise. In particular, we consider two setups. One in which there are matching explanation pairs corresponding to explanations that educate and those that help in guiding action; we call this the aligned case. The other is the unaligned case, where there are no corresponding explanation pairs in the training data.

5. Results

In this section, we provide a number of illustrations of generated explanations. We evaluate the ability of our model to utilize the reason information in generating meaningful and relevant sentences using perplexity and accuracy of pre-trained classifiers as evaluation metrics. For details pertaining to model hyper-parameters and experimental setup, please refer to [20].

The generated explanations are illustrated in **Table 1**. The bold text in the explanation 1 denotes the reason for a loan denial. The bold and italic text in the explanation 2 and 3 denote the incorrect reasons in the generated sentences. The bold and underlined text in the explanation 4 denote the correct reasons in the generated sentences. As is shown, the unaligned ARAEGAN model was incapable of generating explanations serving different purposes. On the other hand, the aligned GAN model not only generated meaningful sentences but could also preserve the reasons from the reference sentences (bold text).

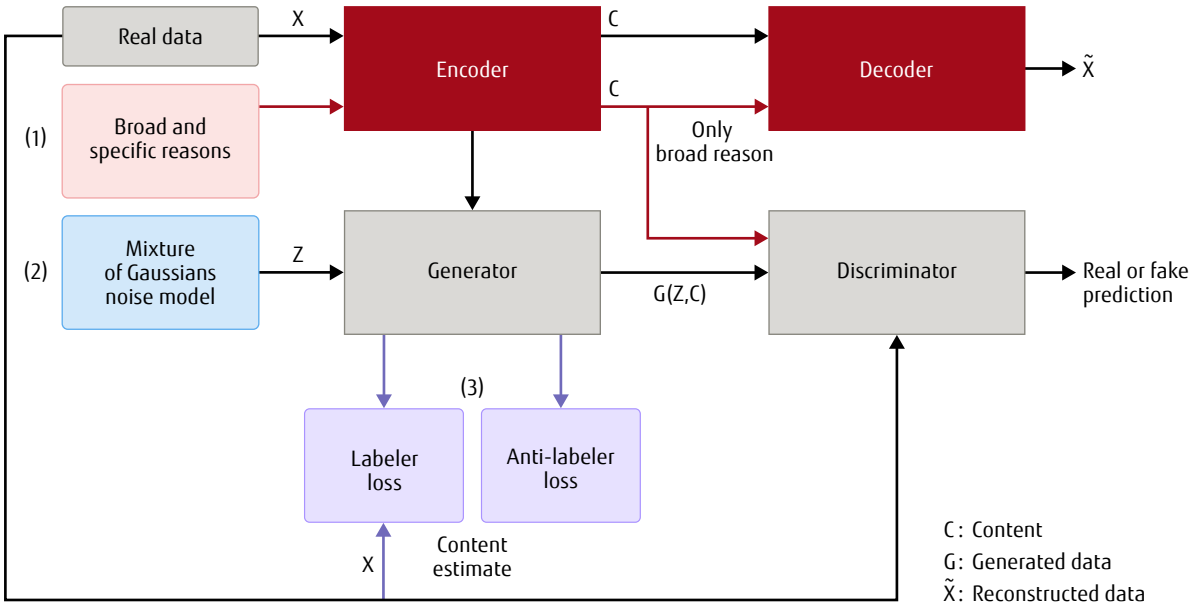

Figure 2
Block diagram of proposed system architecture on top of ARAEGAN.

Table 1
Generating explanations serving different purposes.

Explanations	Models	Transfer from education-oriented explanation to action-oriented explanation	Transfer from action-oriented explanation to education-oriented explanation
1	Reference sentence corresponding to training data	there is a record of inconsistent loan **payments**.	please re-consider applying for a loan of a different **amount** that may better align with your **income**.
2	Generated sentence: Unaligned ARAEGAN model	please re-consider applying for a loan of a different amount that may better align with your ***income***.	the applicant has only been ***employed*** at their current employer for a limited period of time.
3	Generated sentence: Unaligned ARAEGAN model+ GM	talk to your bank about finding ways to improve your ***credit***.	the ***credit*** associated with this application is, unfortunately, not high enough to be considered eligible for this loan.
4	Generated sentence: Aligned GAN model	maintain a consistent record of timely loan **payments** moving forward.	the **income** listed on this application is not high enough to match the **amount** requested for a loan.

GM: Gaussian mixture model for noise

6. Conclusion

In this work, we explored the problem of explainability from a user perspective. In particular, we considered the use case of explaining loan denials and built the first ever dataset that is representative of user-friendly explanations. We observed that the reasons acceptable to users seldom appear as features in machine-learning datasets, which makes model-centric explanations less useful for users. To address this, we designed a novel conditional GAN to generate user-friendly explanations on the basis of the reasons specified in the dataset. This GAN architecture was then extended with style transfer to generate explanations serving different purposes: to educate and to help take actions.

We hope our work will help bridge the gap between research and practice in the financial industry by catering to the needs of the wider community of users seeking explanations, and by generating multiple explanations serving different purposes.

References and Notes

[1] GLOBAL FINTECH INVESTMENT ROBUST ON BACK OF STRONG VC FUNDING: KPMG. Digital News Asia, 2017. *https://www.digitalnewsasia.com/digital-economy/global-fintech-investment-robust-back-strong-vc-funding-kpmg*

[2] FICO: Machine Learning and FICO Scores: An Evolution in ML innovations that helps both lenders and consumers. White Paper: Business Technology Overview 2018. *https://www.fico.com/en/latest-thinking/white-paper/machine-learning-and-fico-scores*

[3] D. Gunning: Explainable Artificial Intelligence (XAI). DARPA/I2O, 2017 *https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf*

[4] KYNDI: How 'Explainability' is Driving the Future of Artificial Intelligence. A Kyndi White Paper, 2018. *https://kyndi.com/wp-content/uploads/2018/01/Kyndi-final-Explainable-AI-White-Paper.pdf*

[5] PWC: Explainable AI: Driving business value through greater understanding. White paper, Intelligent Digital 2018. *https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf*

[6] R. Selvaraju et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. The IEEE International Conference on Computer Vision, pp. 618–626, 2017. *https://ieeexplore.ieee.org/document/8237336*

[7] D. Park et al.: Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. Arxiv, 2018. *https://arxiv.org/abs/1802.08129*

[8] A. Flint et al.: xAI Toolkit: Practical, Explainable Machine Learning. White Paper, 2018. *https://www.fico.com/sites/default/files/2018-06/FICO_xAI_Toolkit-Practical_Explainable_Machine_Learning_4547WP_EN.pdf*

[9] Z. Lipton: The Mythos of Model Interpretability. ICML Workshop, 2016. *https://arxiv.org/abs/1606.03490*

[10] D. Doran: What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. ArXiv 2017. *https://arxiv.org/abs/1710.00794*

[11] A. Chandrasekaran et al.: Do Explanations make VQA Models more Predictable to a Human? EMNLP, 2018. *https://arxiv.org/pdf/1810.12366.pdf*

[12] F. Doshi-Velez et al.: Accountability of AI Under the Law: The Role of Explanation. ArXiv, 2017. *https://arxiv.org/pdf/1711.01134.pdf*

[13] F. Doshi-Velez et al.: Towards A Rigorous Science of Interpretable Machine Learning. ArXiv, 2017. *https://arxiv.org/pdf/1702.08608.pdf*

[14] T. Millers et al.: Explainable AI: Beware of Inmates Running the Asylum. ArXiv, 2017. *https://arxiv.org/pdf/1712.00547.pdf*

[15] B. Herman: The Promise and Peril of Human Evaluation for Model Interpretability. NIPS Workshop 2017. *https://arxiv.org/abs/1711.07414*

[16] S. Amershi et al.: Power to the People: The Role of Humans in Interactive Machine Learning. AI Magazine, pp. 105–120, 2014. *https://www.aaai.org/ojs/index.php/aimagazine/article/view/2513*

[17] M. Narayanan et al.: How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. 2018. *https://arxiv.org/pdf/1802.00682.pdf*

[18] One of the Amazon web services. Combining computer pro-grams with human intelligence can handle tasks that are impossible with computers alone.

[19] A. Chander et al.: Creation of User Friendly Datasets: Insights from a case study concerning explanation of loan denials. ICML HILL Workshop 2019. *https://arxiv.org/abs/1906.04643*

[20] R. Srinivasan et al.: Generating User-friendly Explanations for Loan Denials using GANs. Neurips workshop on Challenges and Opportunities for AI in Financial Services: The Impact of Fairness, Explainability, Accuracy, and Privacy. *https://arxiv.org/pdf/1906.10244.pdf*

[21] J. Zhao et al.: Adversarially Regularized Autoencoders. ArXiv, 2018. *https://arxiv.org/pdf/1706.04223.pdf*

[22] S. Gurumurthy et al.: DeLiGAN: Generative Adversarial Networks for Diverse and Limited Data, 2017. *https://arxiv.org/pdf/1706.02071.pdf*

[23] C. Kemp et al.: Learning overhypotheses with hierarchical Bayesian models. Developmental Science, 10(3), pp. 307–321, 2007. *https://web.mit.edu/cocosci/Papers/devsci07_kempetal.pdf*

[24] M. Kocaoglu et al.: CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. arXiv preprint arXiv:1709.02023, 2017. *https://arxiv.org/pdf/1709.02023.pdf*

Ramya Malur Srinivasan
Fujitsu Laboratories of America Inc., Solutions for Augmented Intelligence Laboratory
Dr. Srinivasan is currently engaged in research and development of explainable AI technologies.

Ajay Chander
Fujitsu Laboratories of America Inc., Solutions for Augmented Intelligence Laboratory
Dr. Chander is currently engaged in research and development of new human-centric technologies and products.