ISSN 2186-7437

NII Shonan Meeting Report

No. 166

Visualization for XAI (Explainable AI)

Seok-Hee Hong
Daniel Keim
Issei Fujishiro

February 6–9, 2023

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Visualization for XAI (Explainable AI)

Organizers:
Seok-Hee Hong (University of Sydney)
Daniel Keim (University of Konstanz)
Issei Fujishiro (Keio University)

February 6–9, 2023

Background and introduction

Summary

Explainable AI (XAI) refers to artificial intelligence (AI) techniques which can be easily understood by humans which increases their trustworthiness. In contrast, many AI techniques generate "black box" models where even their designers cannot explain why they arrive at specific decisions. Visualization of data or the generated models plays a significant role in making AI methods understandable and accessible to the user.

The technical challenge of XAI is to make AI decisions understandable and interpretable, leading to a higher degree of trust by the users of AI methods. In general, AI methods learn useful rules from the test-set; however, they may also learn incorrect rules, which fail to generalize outside the test set, or even inappropriate rules, which lead to an undesired social, political, or racial profiling. XAI methods help humans understand the learned models as well as the relationship to the training and application data. The goal is to get an idea how likely the AI decisions will generalize to future real-world data outside the test-set and understand their economic, social, and political implications.

XAI is of interest to a wide variety of user groups, including

- AI method developers who are interested in understanding and improving their AI algorithms but they are in general application-agnostic
- Application end users who are interested in understanding the AI decisions since they are responsible for the consequences but they are in general model-agnostic
- Data scientists who are interested in understanding which AI methods they should apply to the concrete data set at hand.

The role of visualization in XAI gained significant attention in recent years. Visualizations play an important role since they allow users to quickly get an overview of the learned models as well as their relationship to various data sets involved (training data, test, data, and application data). With the growing complexity of AI models, the critical need for understanding their inner-workings

has increased and the complex relationships between training, test, and application data and their effect on the usefulness and trustworthiness of the AI decisions has to be carefully taken into account. Visualizations are powerful techniques to fill such critical needs.

The main goal of this workshop is to discuss the role of visualization for XAI methods. In particular, we aim at identifying research opportunities for using visualization in XAI, focusing on the Asia- Pacific context. We believe the visualization community can leverage their expertise in creating visual narratives to bring new insight into the often-obfuscated complexity of AI systems.

Overview of the meeting

Aims

This workshop aims at bringing world-renowned researchers in visualization and AI together, and collaboratively develop innovative visualization approaches for XAI with specific applications of large and complex networks to solve the scalability and complexity issues for analyzing big data arising from various application domains including social networks, business intelligence, and network security.

Objectives

Our specific objectives are:

- We will identify research opportunities in XAI, focusing on the visualization perspectives in the Asia-Pacific context.
- We will form a broader research community with cross-disciplinary collaboration, including computer science and machine learning, with a particular focus on visualization and visual analytics for XAI.
- We will foster exchange between visualization researchers and practitioners, and draw more researchers in the Asia-Pacific region to enter this rapidly growing area of research.
- We will assist emerging researchers to link to international researchers, find industrial contacts, and apply for competitive research grants.

Significance and Innovation

XAI is one of the biggest fundamental challenges in IT research due to the wide spread use of AI methods in research and industry. In many application contexts, the applicability of the results, however, suffers from the black-box nature of many AI methods which prevents understandability and interpretability, and ultimately limits the trust into the results. Visualization and visual analytics provide a great potential to overcome the current limits of AI methods and significantly increase understandability, interpretability, and trust in AI methods. Innovative visualization- and visual analytics-based XAI methods may therefore be the key enabler for researchers and end users in many application domains and other disciplines. We believe that this workshop has the potential to set a new research agenda for visualization and visual analytics based XAI research.

Expected Outcomes

- Innovative visualization and visual analytics techniques and solutions for XAI, which will be used by domain experts and end users in various applications.
- Joint publications at top conferences and journals in visualization and visual analytics, jointly authored by visualization and AI researchers.
- Joint funding applications for long-term research collaborations continuing the research.

Impact

- Academic impact: Research publications at top conferences and journals, with high number of citations.
- Societal impact: Visualization and visual analytics techniques for XAI will be in high- demand by researchers and practitioners in various applications and disciplines to solve complex problems in their domains.

Overview of Talks

Visual, Interactive, and Explainable AI: Overview and Open Research Challenges

Mennatallah El-Assady, ETH, Switzeralnd

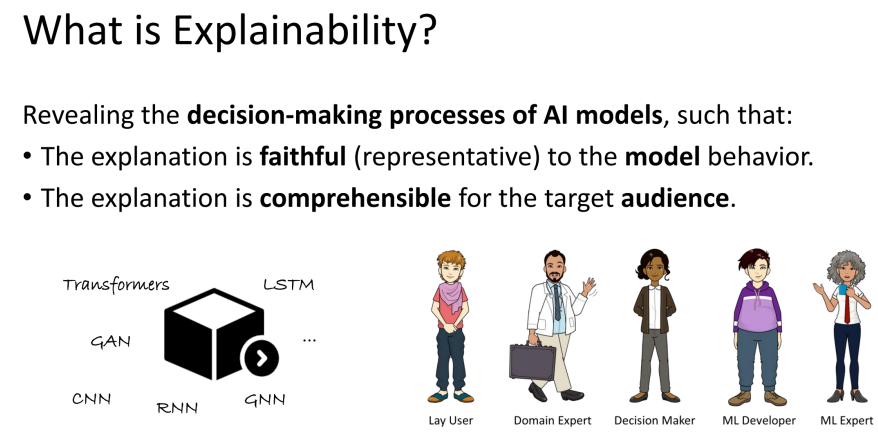

Figure 1: XAI definition based on faithfulness and comprehensibility.

Interactive and explainable machine learning can be regarded as a process encompassing three high-level stages: (1) Understanding machine learning models and data; (2) Diagnosing model limitations using explainable AI methods; and (3) Refining and optimizing models interactively.

In this talk, I review the current state-of-the-art of visualization and visual analytics techniques by grouping them into these three stages. In addition, I argue for expanding our approach to explainability by adapting concepts, such as metaphorical narratives, verbalization, and gamification. I further introduce the explAIner.ai framework for structuring the process of explainable artificial intelligence (XAI) and interactive machine learning (IML), operationalizing it through a TensoBoard plugin.

Lastly, to derive a robust XAI methodology, I present some first steps to extract XAI strategies and mediums by transferring knowledge and best practices from other disciplines.

Visual Analytics for Explainability to understand High Dimensional Data

Kwan-Liu Ma, University of California at Davis

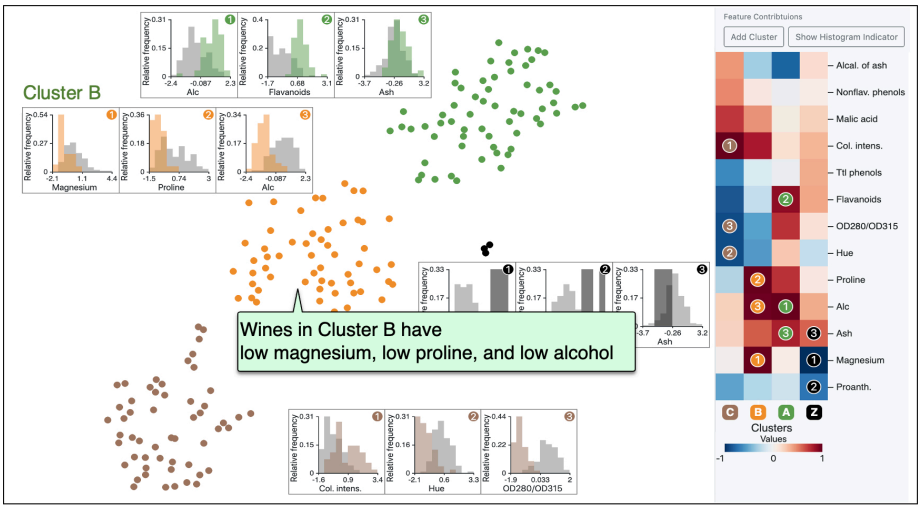

Figure 2: The results of applying ccPCA to the UCI wine dataset.

High dimensional data is commonly found in real-world applications. Dimensionality reduction methods are often used to study the intrinsic structure of such data. But the results from most of these methods, as 2D projections, are not directly interpretable. That is, while these methods, such as t-SNE, visually show us clusters, we don't at a glance know what determine each cluster. In addition, even though the methods that we use, such as PCA which preserves data variance and t-SNE which preserves local neighbors, can extract information from one dataset, they are not suitable for comparative analysis of multiple datasets (or multiple groups in a dataset.)

ccPCA (Contrasting Clsuters in PCA), instead, not only can identify the salient factors/attributes, but also can derive the level of contribution of each attribute, which allows us to explain what attributes/dimensions dominate the

forming of a group/class. ccPCA has been applied to a real-world dataset collected from both colon cancer patients and healthy people for the study of gut microbiota composition. Bacteria in a human gut highly influence the development of colorectal cancer. Understanding the impact of bacteria composition is challenging as the colon environments are influenced by a chain of interactions between multiple bacteria.

ccPCA can greatly help clinical researchers find such combinational factors. First, ccPCA can visually show a clear separation between cancerous and normal samples. More importantly, with ccPCA, we can also explain what bacteria dominate the separation of the two groups of samples with the contribution value computed for each type of bacteria.

Measuring and Explaining the Inter-Cluster Reliability of Multidimensional Projections

Jinwook Seo, HCI Lab, Department of Computer Science and Engineering, Seoul National University

Explainable AI (XAI) is concerned with the development of Artificial Intelligence systems that can provide a clear and interpretable explanation of their decision-making process. Visualizations can play a crucial role in making the process more transparent and understandable to humans by providing insight into how AI systems arrive at their decisions. As visualizations are the key frontend components of XAI, the graphical integrity of visualizations is an important factor in designing XAI systems.

In this talk, we are going to focus on the graphical integrity of low-dimensional projections (i.e., 2D scatterplots) of high-dimensional data since most AI systems are built upon large high-dimensional data. We are presenting two novel metrics, Steadiness and Cohesiveness, for determining the inter-cluster consistency of multidimensional projections (MDPs).

Our aim is to assess how well the relationships between clusters in the high-dimensional space are maintained in the low-dimensional projection space. This consistency is crucial for the accuracy of tasks that involve understanding the relationships between clusters, such as identifying cluster connections in the original space from a projected view. However, existing metrics, like Trustworthiness and Continuity, do not provide an adequate measure of inter-cluster consistency.

Our metrics evaluate two aspects of consistency: Steadiness measures the correlation between clusters in the projected space and those in the original space, while Cohesiveness measures the opposite. We quantify the distortion of randomly selected clusters of different shapes and positions in one space and determine how much they are stretched or dispersed in the other space. Our metrics also produce a reliability map, which visualizes the consistency of the projection and provides useful information for choosing the proper projection technique or hyperparameters.

We conclude by proposing a future research direction to design and development of measures of misrepresentations (i.e., visualization integrity measures) for common XAI visualization techniques.

XAI for the Sciences – Recent Developments in XAI

Klaus-Robert Müller, Machine Learning Group, TU Berlin, Berlin, Germany; Department of Artificial Intelligence, Korea University, Seoul, Korea; MPIII, Saarbrücken, Germany; BIFOLD, TU Berlin, Berlin, Germany

In recent years, machine learning (ML) and artificial intelligence (AI) methods have begun to play a more and more enabling role in the sciences and in industry. In particular, the advent of large and/or complex data corpora has given rise to new technological challenges and possibilities.

In this talk, I will touch upon the topic of ML applications in the sciences, in particular in medicine and physics. I will also discuss possibilities for extracting information from machine learning models to further our understanding by explaining AI (XAI) models.

In particular also recent XAI developments are discussed. E.g. Machine Learning Models for Quantum Chemistry can, by applying XAI, contribute to furthering chemical understanding. Finally, I will briefly outline perspectives and limitations.

Steering Deep Neural Networks towards Interactive Data Labeling, Classifier Debiasing, and Generative Tasks

Jaegul Choo, KAIST, Korea

Human intervention can play a crucial role in utilizing deep neural networks in one's own manner. In this talk, I will present my research on (1) leveraging neural networks for efficient labeling of sentence classification via word-level attention guidance, (2) debiasing classifiers using synthetic data augmentation via image-to-image translation, and (3) interactive image generation and editing via conditional generative adversarial networks.

Finally, I will conclude the talk by discussing challenges and opportunities towards interpretable and interactive deep neural networks.

Visual and Linguistic Explanations in Semantic Machine Intelligence

Komei Sugiura, Keio University, Japan

Models that generate both visual and linguistic explanations have the potential to provide insight into previously unexplained phenomena and to offer user-friendly interfaces for novice users. In this talk, our recent work on XAI models that generate both visual and linguistic explanations is presented.

The focus of the first topic is on visual explanation through the use of light-weight transformers for solar flare prediction. A large-scale solar flare can result in estimated damage of 160 billion, however its theoretical background is not fully understood. In order to provide scientific insights to experts on solar flares, our approach generates attention maps about salient regions based on the incorporation of a branching network into light-weight transformers.

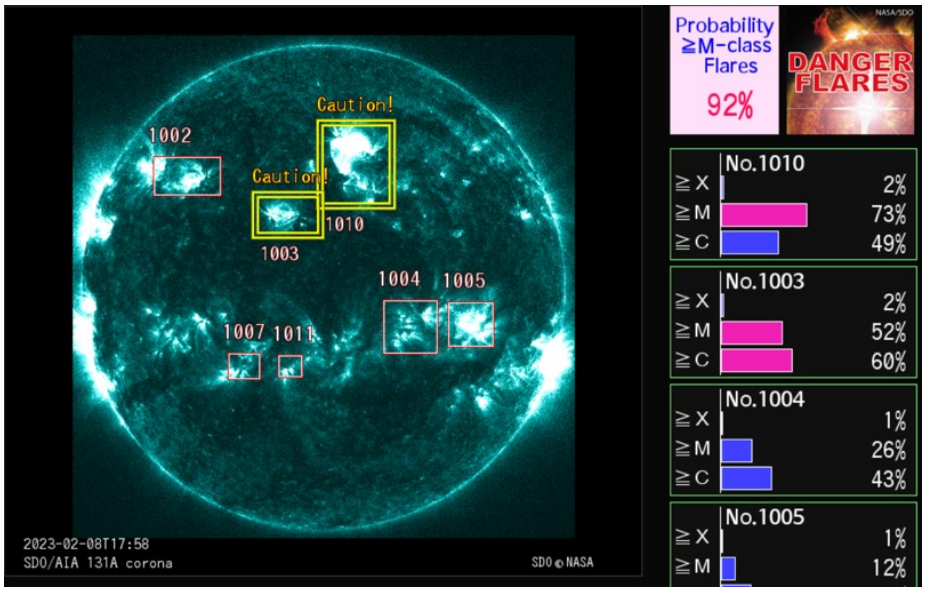

Figure 3: Deep flare net.

The second topic of this talk will address future captioning methods for generating linguistic explanations. These future captioning models generate descriptions of future scenes based on video clips or images. Their applications to cooking videos and domestic service robots are discussed.

XAI - A User- and Domain-Centric Perspective

Jörn Kohlhammer, Fraunhofer IGD, Germany

Successful xAI approaches have to be designed for its target user group and in most cases for a specific domain. Depending on the domain and user role, the model, the learning process, and the results need to be explained in a certain way. In some domains, xAI also fulfils the need for trust building and overcoming skepticism.

My talk introduces the medical domain as one such domain, where there are high-stakes decisions and where black-box algorithms are only acceptable for small sections of diagnostic processes, e.g. for supporting the detection of cancer cells in medical imagery. The context of use for medical experts or patients, and the prior knowledge and stances of these user groups, influences the level of detail that is required of explanatory visualizations.

My talk also briefly summarizes some results of two earlier Dagstuhl seminars on Visualization for Trustworthy AI. These included, most prominently, a manifesto of the participants of the first workshop, and a first model that distinguishes subjective and objective trust during human AI-supported decision making.

Analysis of CNNs and Its Applications

Yasushi Yamaguchi, University of Tokyo, Japan

Figure 4: Examples of unexpected region removal.

This talk introduces three studies aiming at analyzing the internal functionalities of convolutional neural networks for images and some applications based on the analysis.

The first study is a technique for visualizing class-targeted features of classification networks. Even if we generate a feature visualization from the class-targeted neuron attribution, the result can be contaminated with the other features located in the different regions in the image. We proposed a region conscious visualization method which generates a spatial mask as well as a feature visualization related to the target class simultaneously to filter out the unrelated features.

The second one is a pruning technique to obtain a light-weight single-class detection network from the multi-classes detection network. We developed the way to calculate output target related to objects to be detected for the detection network and examined the pruning effect in each layer.

The last study is an interactive tuning method to remove unexpected regions from generated images with GANs. The method allows users to interactively modify the pretrained generator network by ablating some channels related to the artifacts. It is capable of modifying the network to avoid watermarks or captions even if the network has learned with watermarks and captions from the training dataset.

List of Participants

- Jaegul Choo (KASIT, Korea)
- Mennatallah El-Assady (ETH, Switzeralnd)
- Issei Fujishiro (Keio University, Japan)
- Seok-Hee Hong (University of Sydney, Australia)
- Yun Jang (Sejong University, Korea)
- Daniel Keim (University of Konstanz, Germany)
- Jörn Kohlhammer (Fraunhofer IGD, Germany)
- Bum Chul Kwon (IBM Research, USA)
- Kwan-Liu Ma (UC Davis, USA)
- Klaus-Robert Müller (TU Berlin, Germany)
- Jinwook Seo (Seoul National University, Korea)
- Komei Sugiura (Keio University, Japan)
- Yasushi Yamaguchi (University of Tokyo, Japan)

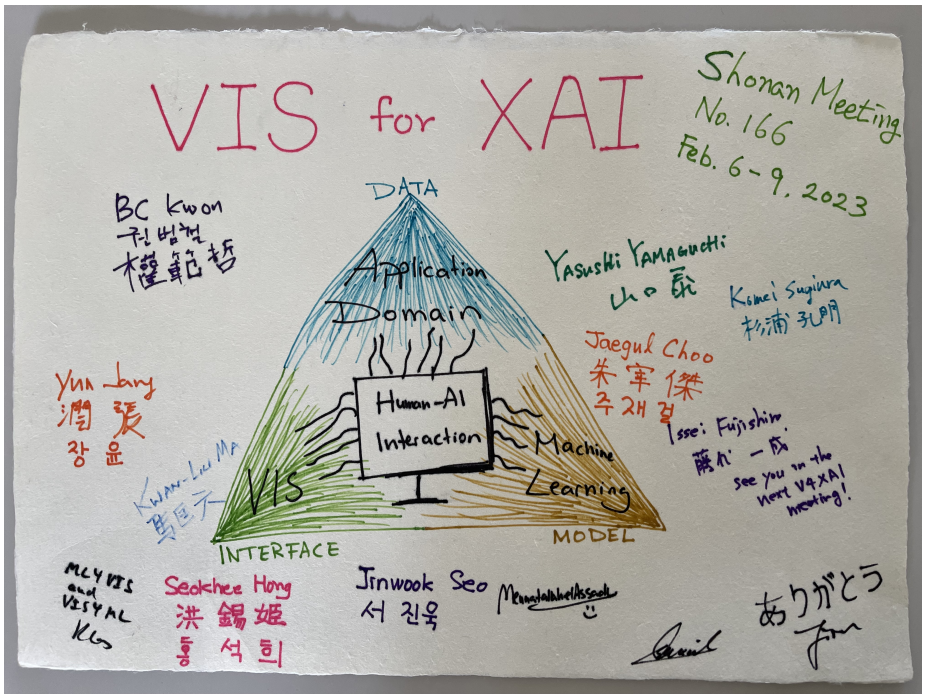

Figure 5: Shonan meeting 166 on Vis for XAI

Meeting Schedule

Day 0: February 5 (Sun)

- 19-21 Welcome Banquet

Day 1: February 6 (Mon)

- 9-9:30 Introduction
- 9:30-10:30 Talk by Menna
- 11-11:30 Talk by Kwan-Liu
- 11:30-12 Talk by Jinwook
- Group Photo Shooting
- 13:30-14:00 Talk by Klaus
- 14-14:30 Talk by Jaegul
- 14:30-15:00 Talk by Komei
- 15:30-16 Talk by Joern
- 16-17 Open problem session and Group formation 1

Day 2: February 7 (Tue)

- 9-9:30 Talk by Yasushi
- 9:30-10:30 Open problem session and Group formation 2
- 11-12 Group Discussion 1
- 13:30-15 Group Discussion 2
- 15:30-16:30 Group Discussion 3
- 16:30-17 Group Report 1

Day 3: February 8 (Wed)

- 9-10 Group Discussion 4
- 10:30-11:30 Group Discussion 5
- 11:30-12 Group Report 2
- 13:30-21 Excursion and Main Banquet

Day 4: February 9 (Thu)

- 9-10 Group Discussion 6
- 10:30-11:30 Group Discussion 7
- 11:30-12 Group Report 3 and Wrap up

Group 1 Report

Visualization and Machine Learning: Bridging Gaps towards Insight

Summary of discussions

We discuss the need for integrating visualization and machine learning beyond the current research in XAI, such that visualizations also help to understand the data and the problem, and allow the domain experts to steer the modeling process and incorporate their knowledge. To solve complex application problems, domain experts, modeling experts, and visualization experts need to work together by incorporating their complementary strengths to generate new insights. For example, the domain experts can contribute their data and problem understanding, the modeling experts can contribute their knowledge of the modeling process, and the visualization experts can contribute their expertise in creating visual narratives to bring new insight to enable more effective usage of the AI systems.

We discuss more specific details of the three phases, namely the *data and problem*, *modeling*, and *insight generation*, where the three expert groups work together to successfully address the complex application problems. Finally, we conclude with challenges and future research directions that require a close collaboration of ML and VIS researchers to set the joint research agenda.

Summary of new findings

In general, the ML researchers can benefit from a better understanding of the targeted perceptual and interactive tasks facilitated by the visualization approach, which needs to be tightly integrated with the modeling approach. The VIS researchers can benefit from a more targeted development and use of an ideal model and parameterization for the automated part of the solution. Therefore, enabling the explanation and integration of automated results with an interactive visualization approach can enable domain users a truly resonate solution with the combined expertise of both ML and VIS communities.

- *Data and Problem Specification*: Collaboration between the domain experts, ML experts, and VIS experts is essential for this stage. Specifically, data specification is crucial for assessing the data quality, such as data correctness, data completeness, as well as potential biases. For problem specification, domain experts need to work with the ML experts to define the problem to be solved and select an appropriate model and the necessary inputs.

  Moreover, the data and problem need to be specified using more general terms than the domain-specific terms so that the ML experts and VIS experts can focus on the gist of the problem. Therefore, close communication between the three expert groups is crucial for the data and problem specification to be effective.

- *Modeling, Explanation, and Representation*: For this process, ML experts need to work with domain experts to select the best ML models and provide guidance on how to set important hyperparameters.

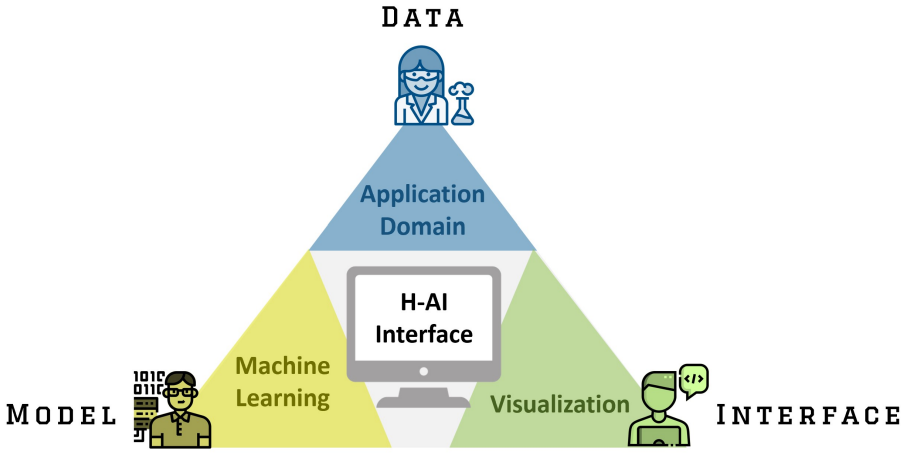

Figure 6: Collaboration between domain experts, ML experts and VIS experts.

XAI techniques can be deployed to enable better understanding of the results. Namely, VIS experts can help to shortcut the model selection, training, and optimization process. With the help of VIS experts, the models and results can be reviewed by the ML experts, with the feedback from the domain experts. Note that an iterative loop between the modeling and insight generation is necessary to develop the appropriate solutions and to improve the results.

- *Insight Generation*: Domain experts use the ML and XAI systems to solve their application problems or to generate new insights. Therefore, effective interaction with the ML and XAI systems plays an important role in this process.

  An appropriate visual analytic interface need be implemented to support effective interaction between the ML and XAI models, minimizing perceptual biases. However, designing effective Human-AI interfaces poses significant challenges due to their complexity.

Identified issues and future directions

We now discuss challenges and future research directions, which require a close collaboration between the domain experts, ML researchers and VIS researchers.

- *Interdisciplinary Design Process*: It is highly challenging to collaborate effectively due to the difference in the languages used in different domains. Therefore, the team of diverse experts needs to actively learn the language of others to collaboratively identify problems and brainstorm solutions together. More specifically, the team must discuss initial development, critical feedback, and refinement together with the domain experts.

  Moreover, it is equally important to build a trust between the experts in different domains, to facilitate the communication. Trust building can start by better understanding each other's expertise.

Figure 7: Research challenges.

- *Bias and Uncertainty Amplification*: When building a visual, interactive and explainable ML systems, it is highly challenging to analyze the confound effects of the multi-stage pipelines developed in the system. In general, the underlying pipelines consist of one or more data processing and machine learning step, as well as various visualization and interaction steps. Specifically, each of these steps involves a transformation of some input (i.e., data) to some output (i.e, representation), which may introduce some artifacts.

  Therefore, a significant challenge for both ML and VIS researchers is to effectively measure the amplification effects in such multi-stage pipelines. For example, a possible future research direction is to develop systematic and automatic tools to explore the limits of combining machine learning and visualization.

- *Unified Evaluation*: The existing evaluation methods for explainable ML systems use either algorithm-centric evaluation methods (such as quality metrics and benchmarks) or human-centric evaluation methods (such as user studies).

  A significant challenge is to design unified evaluation techniques that integrates both the fidelity of the ML models as well as their comprehensibility to the targeted user group. Therefore, a possible future research direction for ML and VIS researchers is to collaboratively design unified evaluation schemes that consists of both the algorithm-centric and the human-centric evaluation methodologies.

Group 2 Report

Visual Analytics for Evaluating Large-scale Pretrained Models

Summary of discussions

The use of large-scale pretrained models, such as GPT-3 and ChatGPT, has been beneficial in various AI applications, however there may be a domain gap between the pretraining data and new tasks, making transfer learning necessary. Finetuning can lead to overfitting, but techniques such as adding a

fully-connected layer have been proposed to mitigate this issue. In a zero-shot setting, prompt tuning is used to optimize natural language textual templates for the model to perform new tasks. However, this approach can be ad-hoc and heuristic, requiring a visual analytic system to analyze the resulting outputs and form various textual query templates. This system must rely on a black-box approach for interpreting the model behavior, as internal states may be inaccessible. Pretrained models may also have limitations, requiring additional information on model uncertainty.

In response, visual analytic approaches can help to address these challenges in leveraging pretrained models for new tasks, and various potential approaches and research directions have been discussed.

Summary of new findings

Large-scale pre-trained models are making a significant impact on various applications of artificial intelligence. For example, in natural language processing domains, GPT-3 [1] and ChatGPT [2,3] are making a huge impact by working as a general AI agent that can answer various questions describing one's own tasks. Consequently, information retrieval tasks such as search engines are now being replaced by a more intelligent agent, such as ChatGPT, that can understand various natural language queries whether it is asking the document summarization, machine translation, document classification, beyond just a simple keyword-based information retrieval of web documents. Another example of an impactful large-scale pretrained model is Stable Diffusion [4], which can synthesize high-quality images reflecting the users' own textual description given as input.

Given these powerful, general-purpose AI models, our task is often domain-specific, and it may involve narrow tasks. For example, in the medical domain, we may want to automatically detect some phrases and sentences from free-text medical notes. We may also want to identify the area of nodules from lung CT images. In these cases, there may exist a significant domain gap between the training data used during the pretraining stage and the data for new tasks, limiting the effectiveness of utilizing these preptrained models.

To properly leverage the power of the above-described pretrained models for new target tasks, we often adopt the *transfer learning* approach that fine-tunes the given pre-trained models by using our own, possibly small-scale training data. Alternatively, we can even ask directly the model to solve our task by describing our tasks in a natural language form, without finetuning the pretrained model at all. For example, we can directly ask the ChatGPT model, by saying "find out the treatment sentences from the following sentences." Here, the former approach that uses finetuning can be referred to as a few-shot approach while the latter as a zero-shot approach.

In each of these two scenarios of leveraging the pretrained models for our own tasks, new challenges emerge. In the former case, the main issue is potential *overfitting*. The training data for finetuning is relatively small compared to the large-scale data used during pretraining. Thus, it can cause finetuning, or so-called catastrophic forgetting of the general knowledge learned from the pretraining stage. This does not only harm the generalization capability of the downstream tasks but also prevents the finetuned model from being recycled for various future use, e.g., adapting to domain-shifted data.

Researchers recently proposed various techniques to avoid it. For example, we can add a simple *fully-connected layer* between some adjacent layers of the pretrained model and train only these newly added layers, while keeping the original pretrained model parameters intact. Also, one can restrict those trainable regions to be a certain layer of the pretrained model. Moreover, researchers have recently proposed the technique to finetune only the existing fully-connected layer playing a role of transforming an input vector to a query, a key, and a value vector in a self-attention block [5].

However, the current approaches in this process are mostly designed in an ad-hoc manner, without conducting a systematic, in-depth analysis. Therefore, a *visual analytic approach* can help this stage. For example, when finetuning the pretrained model, one can summarize and visualize which layers has changed most significantly. Furthermore, one can apply an interactive what-if scenario approach of analyzing the effect of a certain layer by excluding the rest of the network from the trainable parameters during the finetuning stage.

Additionally, from the perspective of data-item level analysis and interpretability, one can highlight which training or test data items are mainly affected as to whether they are correctly predicted or not, in terms of the original pretraining tasks and/or the finetuning tasks. Here, the reason why we need to care about the accuracy of the original pretraining tasks is because our finetuning tasks can change over time or the characteristics of the data for finetuning can evolve over time. In this manner, visual analytic approaches can help to analyze how the network behavior changes due to different finetuning strategies.

Identified issues and future directions

In this entire process, the main challenge is that the number of parameters of the *pretrained model* is huge, easily going beyond billions of parameters for a single pretrained model. How we can properly summarize and visualize such a large-scale set of parameters in the network would be a significant challenge.

On the other hand, the latter case of using the pretrained model in a zero-shot setting poses another important challenge. For example, ChatGPT has shown surprisingly great potentials across a wide variety of tasks that are described in a natural-language-form input. In this case, to improve the accuracy of the target task, people often perform prompt tuning, which identifies the optimal natural language textual template to ask the model to perform our own tasks. However, this problem is inherently ill-defined since the variables to optimize are those words in a particular order, which are basically categorical variables.

Tackling this problem is often performed in an ad-hoc and heuristic manner, therefore a *visual analytic system* that can greatly help users form various textual query templates and analyze the resulting outputs from the pretrained model would be an interesting research direction. In this process, the visual analytic system can facilitate the comparisons between different textual templates and allows users to find an optimal template for their own tasks. Moreover, such *human-in-the-loop analysis* in a visual analytic environment can enlighten users with a crucial hypothesis towards an optimal prompt template.

However, a caveat in this process is that we may not be able to access the internal state of the pretrained model, which is currently the case of ChatGPT. In this case, our visual analytic approach should mainly rely on black-box

approach for interpreting and understanding the model behavior, which poses another important challenge in leveraging the pretrained model.

In addition, the strategy of using the pretrained model in a zero-shot setting bears the limitations of the pretrained model, where it may not perform particularly poorly in certain situations. For example, when performing a sentiment classification of a product review, the pretrained ChatGPT may give incorrect answers more frequently for longer sentences than short ones. In that case, one may have no other choice but to accept such limitations in a zero-shot setting. Therefore, when deploying the model to end users, it would be desirable to provide additional information about the model uncertainty as to how much the predicted answer by the pretrained model is trustworthy.

In conclusion, the capability of a large-scale pretrained model is significantly improving in terms of its wide applicability, towards getting a general artificial intelligence system. However, as we rely more on this pretrained model, the techniques and strategies on how we properly utilize these models for our own tasks and purposes are also fast evolving, having to adapt to various characteristics of these pretrained models. These techniques and strategies are heuristic, not often based on a standard practice of setting up a clear problem definition and optimization via a properly designed loss function and its associated training data. Therefore, *human-in-the-loop visual analytic approaches* can be an effective alternative.

References

[1] Brown, T. et al. Language models are few-shot learners. Advances in neural information processing systems 33 (2020): 1877-1901.
[2] https://openai.com/blog/chatgpt/
[3] Ouyang, L. et al. Training language models to follow instructions with human feedback. in NeurIPS 2022.
[4] Rombach, R. et al. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[5] Tumanyan, N. et al. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. in arXiv 2022.