Failure-Resistant Intelligent Interaction for Reliable Human-AI Collaboration

Hiromu Yakura

Graduate School of Science and Technology, University of Tsukuba / Tsukuba, Japan
hiromu.yakura@aist.go.jp

Introduction

My primary research interests are at the intersection of human–computer interaction and machine learning. In particular, I am focusing on how we can overcome the gap people have against machine learning techniques that require a well-defined application scheme and can produce wrong results. I acknowledge that a number of researchers have contributed to improving the accuracy of machine learning, but I believe that it is also important to investigate how humans can benefit from imperfect machine learning systems. In addition, to promote sustainable collaboration between humans and machine learning systems, we need to carefully design how the systems obtain feedback from and intervene in humans. Considering that an increasing number of machine learning systems are employed in many people's daily lives, we need not only to improve machine learning itself but also to cultivate effective ways of leveraging it.

My previous projects are in line with this context; for example, the one on malware analysis support (Yakura et al. 2018b, 2019) focused on the gap that, while computer security experts wanted to automate costly analysis, the introduction of machine learning had not been so favored because it might put errors in the analysis results that are used for forensic purpose. Then, instead of automated end-to-end analysis, I proposed a new approach that offers hints for experts who are beginning to analyze by highlighting characteristic behaviors using the attention mechanism. Another project proposing a music recommender dedicated to being used while working (Yakura et al. 2018a) stemmed from the gap behind the scientific result that people should listen to songs they feel mediocre for keeping concentrated. That is, it is difficult for humans to find their mediocre songs by themselves, while it is much easier for a recommender to find such songs rather than to find their best songs.

In my thesis, I am planning to discuss the principle of the interaction design that fills such a gap, namely "failure-resistant intelligent interaction," based on a series of research I have conducted and am working on. My approach is on both recognizing the limitation of current technology and providing users with control to overcome the limitations, as presented in the following sections.

Computational Support for Coaching

In executive coaching, coaches are required to analyze the nonverbal behavior of a coachee during a conversation, which is mentally demanding. Therefore, I started this independent project with my undergraduate friend, assuming that computationally analyzing such nonverbal behavior can help coaches. However, in the early attempts, we found that conventional methods based on supervised machine learning cannot consider the contextual semantics of the behavior and often contradicted the intuition of skilled coaches.

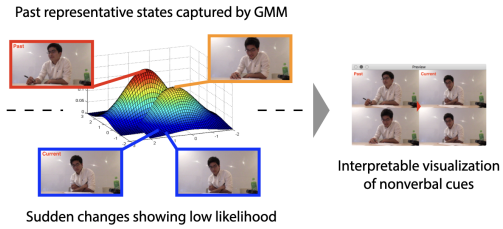

Figure 1: GMM allows us to capture sudden changes in the nonverbal behavior of a coachee, which help coaches infer the coachee's internal status (Arakawa and Yakura 2019).

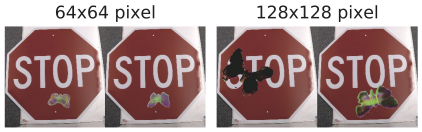

Figure 2: By placing a moth-like patch on a 'STOP' sign, we can make autonomous driving cars misrecognize the sign as 'Speed Limit 80' (Yakura et al. 2020).

Hence, we developed a dedicated system that captures sudden changes in coachees' nonverbal behavior in an unsupervised manner using the Gaussian mixture model (GMM), as in Fig. 1. It presents a coach with cues but entrusts their interpretation to the coach to be free from the contradiction. We confirmed the effectiveness of this approach with expert coaches, and this work was accepted to ACM CHI '19 and '20 (Arakawa and Yakura 2019, 2020).

Practical Adversarial Examples

To overcome the gap between humans and machine learning, we also need to understand the limitations of machine learning techniques. To this aim, I have worked on research about adversarial examples to reveal the risk of machine learning systems being intentionally misguided. In IJCAI '19, we showed that, by broadcasting a subtle noise, we can make deep-learning-based speech recognition devices in the phys-

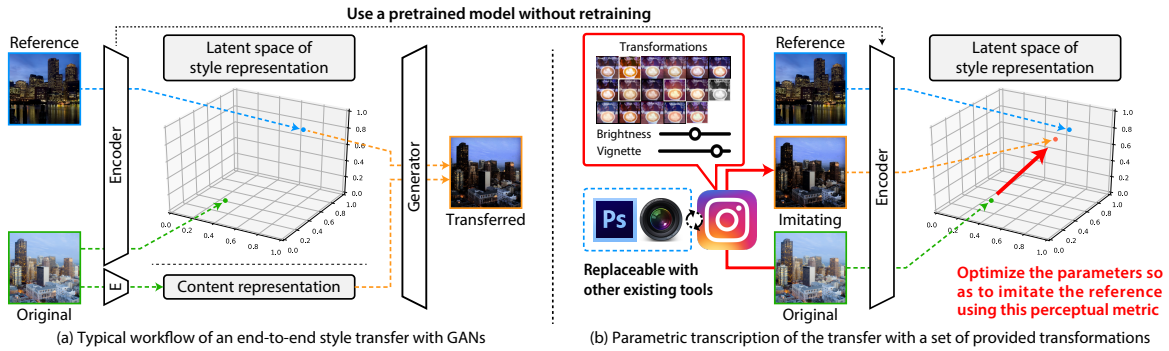

Figure 3: Our approach presents a way to stylize so that a user can explore the variations of a stylized result in a familiar tool.

ical world mistranscribe human utterances, which can result in abusing voice assistants (Yakura and Sakuma 2019).

In AAAI '20 (Yakura et al. 2020), we showed that placing a moth-like patch on a STOP sign can make autonomous driving cars misrecognize it as Speed 80 (Fig. 2). This work revealed the existence of the gap from another perspective; while human drivers would not feel the patch suspicious by assuming it to be a moth stopping on the sign, it can fool the cars as an adversarial example. Given that adversarial examples are common to deep learning models, this work emphasizes the importance of designing interactions in which humans can cope with such malfunction of machine learning by providing users with transparent control.

Explorable Content Editing

The importance of such transparent control is also emphasized in my other project that focused on the gap between deep-learning-based content editing techniques and the creative process of humans. Here, our design process is often exploratory, that is, an open-ended journey starting with an under-specified goal. Therefore, "one-shot" editing provided by end-to-end techniques is not optimal for serving such a serendipitous process. For instance, many people are still enjoying editing photos using Instagram, exploring possible results, whereas state-of-the-art photo enhancement transfer methods can produce high-fidelity results.

Given that, I came up with the idea of letting users know how to obtain a stylized result in a tool they are familiar with, instead of providing only the stylized result (Fig. 3). In photo editing, users can know which transformations should they apply to what extent in order to obtain a stylized result, and then, they can enhance contrast or disable a specific filter. This is enabled by combining black-box optimization with a perceptual metric retrieved from a pretrained style transfer model. Importantly, this approach allows us to leverage various pretrained models accumulated via machine learning research. I confirmed in experiments its applicability to stylizing pictures and making up facial photos, which was accepted to IJCAI '21 (Yakura et al. 2021).

Future work

As described above, I have consistently worked on how to fill the gap between machine learning techniques and humans. Currently, I am working on investigating the effect of supporting intellectual tasks (e.g., writing texts or preparing presentations) using large language models, which includes not only making progress but also maintaining users' task engagement (i.e., avoiding procrastination). For the latter purpose, models do not always have to be accurate but can produce some interesting or funny outputs, which makes their relationship with the users failure resistant.

Toward completing my thesis, I want to deepen the discussion about the principle of the interaction design that unifies all of my past projects. This requires a comprehensive perspective that bridges artificial intelligence and human-centered computing. In these few years, I relatively put more emphasis on the latter, as I conducted empirical research centered on semi-structured interviews (Yakura 2021). Thus, discussing with other doctoral students in the AI community certainly helps me to construct bold design principles.

References

Arakawa, R.; and Yakura, H. 2019. REsCUE: A framework for REal-time feedback on behavioral CUEs using multimodal anomaly detection. In *ACM CHI*, 572.

Arakawa, R.; and Yakura, H. 2020. INWARD: A computer-supported tool for video-reflection improves efficiency and effectiveness in executive coaching. In *ACM CHI*, 1–13.

Yakura, H. 2021. No more handshaking: How have COVID-19 pushed the expansion of computer-mediated communication in Japanese idol culture? In *ACM CHI*, 645:1–645:10.

Yakura, H.; and Sakuma, J. 2019. Robust audio adversarial example for a physical attack. In *IJCAI*, 5334–5341.

Yakura, H.; et al. 2018a. FocusMusicRecommender: A system for recommending music to listen to while Working. In *ACM IUI*, 7–17.

Yakura, H.; et al. 2018b. Malware analysis of imaged binary samples by convolutional neural network with attention mechanism. In *ACM CODASPY*, 127–134.

Yakura, H.; et al. 2019. Neural malware analysis with attention mechanism. *Comput. Security*, 87.

Yakura, H.; et al. 2020. Generate (non-software) bugs to fool classifiers. In *AAAI*, 1070–1078.

Yakura, H.; et al. 2021. Tool- and domain-agnostic parameterization of style transfer effects leveraging pretrained perceptual metrics. In *IJCAI*, 1208–1216.