Generating Personalized Behavioral Feedback for a Virtual Job Interview Training System Through Adversarial Learning

Alexander Heimerl[1](✉), Silvan Mertes[1], Tanja Schneeberger[2], Tobias Baur[1], Ailin Liu[1], Linda Becker[3], Nicolas Rohleder[3], Patrick Gebhard[2], and Elisabeth André[1]

[1] Lab for Human-Centered AI, Augsburg University, 86159 Augsburg, Germany
{alexander.heimerl,silvan.mertes,tobias.baur,elisabeth.andre}@uni-a.de
[2] German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus D3.2, Saarbrücken, Germany
{schneeberger,gebhard}@dfki.de
[3] Department of Psychology, Friedrich-Alexander University Erlangen-Nürnberg,
91054 Erlangen, Germany
{linda.becker,nicolas.rohleder}@fau.de

Abstract. Job interviews are usually high-stakes social situations where professional and behavioral skills are required for a satisfactory outcome. In order to increase the chances of recruitment technological approaches have emerged to generate meaningful feedback for job candidates. We extended an interactive virtual job interview training system with a Generative Adversarial Network (GAN)-based approach that first detects behavioral weaknesses and subsequently generates personalized feedback. To evaluate the usefulness of the generated feedback, we conducted a mixed-methods pilot study using mock-ups from the job interview training system. The overall study results indicate that the GAN-based generated behavioral feedback is helpful. Moreover, participants assessed that the feedback would improve their job interview performance.

Keywords: Job interview training · Generative adversarial networks · Counterfactual explanations · Engagement

1 Introduction

In stressful situations, such as job interviews, many people tend to show nervous and uncontrolled behaviours. This circumstance most often affects their performance in a negative way. Especially in job interviews, the goal is to convince a recruiter of ones fit in a company by actively engaging in the conversation. Recruiters hereby consciously or unconsciously evaluate the candidate's social cues. The amount of positive engagement a candidate shows towards the interviewer may play a central role in deciding whether the candidate is suitable.

Delroy et al. [7] found that active integration behaviors such as engagement, laughing, and humor led to better performance ratings and, therefore, to a higher chance of recruitment. In recent years, technology-based job interview training systems have been developed to improve the performance of candidates (e.g. [1,4,11]).

This paper presents a feedback extension to an existing job interview training environment that uses a socially interactive agent as a recruiter and an engagement recognition component to enable the virtual agent to react and adapt to the user's behavior, and emotions [2]. This training aims to help improve social skills that are pertinent to job interviews. The new feedback extension employs an eXplainable AI (XAI) method based on counterfactual reasoning for generating verbal feedback about observed social behavior.

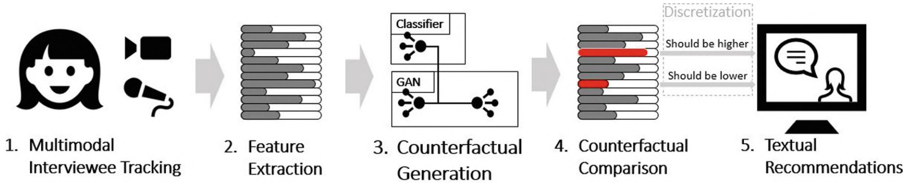
1. Multimodal Interviewee Tracking
2. Feature Extraction
Classifier
GAN
3. Counterfactual Generation
Discretization
Should be higher
Should be lower
4. Counterfactual Comparison
5. Textual Recommendations

Fig. 1. Job interview training system with GAN-generated recommendations.

The introduced feedback extension is based on a deep learning classifier predicting the user engagement in job interview situations that uses multimodal feature (e.g., gaze, body posture, or gestures) representations of the trainee as input. We exploit the concept of counterfactual explanations to show what the user would need to change to appear more engaged. Therefore, a GAN-driven counterfactual explanation model is trained that transforms those feature representations to corresponding counterfactual explanations, i.e., the feature representations are changed so that the user would have appeared engaged. The explanation generation compares the counterfactual feature vectors with the original feature vectors to derive textual recommendations automatically. Finally, they are presented to the trainee by a socially interactive agent in the role of a job interview coach. Figure 1 shows a schematic overview of our approach.

2 Recommendation Generation

The next sections offer an overview of the different components we implemented to generate behavioral recommendations that point out how the user should have behaved to appear more engaged.

Feature Extraction. In order to train a model for engagement recognition and recommendation generation, we modeled a high-level engagement feature set. The feature set consists of 18 metrics mapping facial behavior, body language and conversation dynamics.

Engagement Model. Based on the introduced feature set we trained a simple feedforward neural network with two dense layers for the recognition of low and high engagement on the NoXi database [3]. We decided on the NoXi corpus since it contains multi-modal multi-person interaction data and its transferability to social coaching scenarios. Moreover, the setup of the corpus allowed for both engaging as well as non-engaging interactions. The 10.5 h of data has been randomly split into training and test sets, so that no sample of the same participant is present in the training and the test set. The corresponding classifier achieved an accuracy of 70.5%.

Counterfactual Features. In a next step, to be able to give recommendations on how the user should have behaved to appear more engaged, we apply a counterfactual explanation generation algorithm, i.e., we aim to modify the input feature vectors that were classified as *low engaged* in a way that the classifier would change it's decision to *high engaged.* As described in Sect. 1, the recommendations that we aim for can be seen as counterfactual explanations for the engagement model presented in Sect. 2. To generate these counterfactual feature vectors, we used an adversarial learning approach. In prior work, Mertes et al. [5] presented their *GANterfactual* architecture, which is an adversarial approach to transforming original samples to counterfactual samples that are classified in a different way by a specific decision system to be explained. For our system, we built a network architecture adapted from the GANterfactual framework, which was originally implemented for generating counterfactual explanations in the image domain. The use of the GANterfactual framework has multiple benefits for the recommendation quality: Firstly, the cycle-consistency loss that is an integral part of the underlying adversarial architecture forces that the learned transformation is minimal, i.e., only relevant features are changed. In the context of recommendation generation, this implies that the generated behavioral recommendations are highly personalized. Secondly, the adversarial loss component that is part of every GAN architecture leads to highly realistic results. Thus, recommendations are not drawn from highly exaggerated or oversimplified feature vectors. Thirdly, the counterfactual loss introduced by Mertes et al. enforces that the counterfactual explanations (in our case, the behavioral recommendations), are valid. For technical details of our modifications to the original GANterfactual framework, please refer to our implementation.[1] For the GAN-training, we relied on the NOXI dataset, which we also used for training the engagement classifier. Thus, the adversarial framework learns to convert feature vectors that show low engagement to feature vectors that show high engagement.

Textual Recommendations. After generating the counterfactual feature vectors we compare them to the original feature vectors that represent the shown nonverbal behavior. Depending on the demanded detail of feedback we return the features that had undergone the greatest value transformation and convert them into textual feedback. For this purpose, we discretize the features based

[1] Our implementation is available at https://github.com/hcmlab/FeatureFactual.

on a defined textual template. For example, the feature representing the overall activity of the head gets translated into “try to keep your attention on your interlocutor” or “try to use more nonverbal feedback” depending on the present feature value. The generated feedback is provided verbally to the user by the virtual coach inside the job interview training environment.

3 Evaluation

Pilot Study. The present pilot study’s goal was to get preliminary insights about the assessment of a possible job interview training applying GAN driven recommendations. We gathered data from 12 volunteering student participants (7 female, 5 male). Participants’ age was between 21 and 29 years ($M = 23.83$, $SD = 2.66$). The participants were presented with videos of our job interview training system applied to a multi-modal job interview role-play dataset [10]. Participants were asked to imagine that they were the trainees using the training to practice a job interview. Next, participants filled in questionnaires about Demographics (age, sex), Usefulness (MeCUE [6]), Transfer motivation (four items adapted from [9]) and Feedback Quality (“I felt the feedback was accurate.”, “I would have given similar feedback.”, “I feel like the feedback is helpful.”, “I don’t think the computer can give me accurate feedback.”). Then, a semi-structured interview was held, which covered five areas: 1) general impression, 2) other possible use-cases, 3) suggestions for improvement, 4) intention for further use, and 5) added value.

In the three questionnaires, the following descriptive data was found: Usefulness ($M = 4.72$, $SD = 1.17$); Transfer motivation ($M = 4.92$, $SD = .94$); Feedback Quality ($M = 4.60$, $SD = 1.26$). The answers gathered in the semi-structured interview were analyzed and categorized for each of the five areas separately. Regarding the *General impression*, the majority of participants mentioned that the recommendations were useful (6)/feasible or comprehensible (2). As *other possible use-cases* participants named training to improve communication skills in general (8) and for more specific groups, like patients with anxiety disorders or people with social phobias. Participants mentioned seven times that they would like to have more specific recommendations, e.g. “The agent could say something like: Nonverbal feedback is nodding, for example.”. *Intention for further use* was indicated by 9 participants. The *added value* of the training was for most of the participants that the recommendations are given directly on a specific behavior shown in a specific situation during the job interview.

Recommendation Generation. In order to verify the validity of our approach, we examined whether the counterfactuals generated by the GAN are modifying the features that the incorporated engagement classifier identified as important for the classification of low and high engagement. For this evaluation, we used five sessions of the multi-modal job interview role-play dataset [10] that have also been used in the pilot study and extracted the importance scores of every feature in regard to the model’s classification with LIME [8]. We calculated the pearson correlation between the absolute value change of how much each feature has been modified by the counterfactual transformation and the importance score of every feature, see Table 1. High correlation scores indicate that the counterfactual feature transformation is in line with the corresponding importance of the feature. Seven features showed a strong positive correlation (GZ_DR, AM_CR, HD_TH, DIST_RW, YROT_LE, SDX_HD, SDXROT_HD), six features had a moderate positive correlation (HD_AC, YROT_RE, XROT_RE, TN_HD, CONT_MOV, EN_HA) and two features presented with a low positive correlation (DIST_LW, XROT_LE). Moreover, FO_RW had a strong negative correlation, VAL_F showed a moderate negative correlation and FO_LW had a weak negative correlation. Moreover, we conducted a computational evaluation to investigate how well the generated counterfactual features change the decision of the engagement classifier. We found that 96.49% of the generated counterfactual feature vectors led to a different decision of the engagement model as the original input features.

Table 1. Pearson correlation between the absolute change of the feature values and the LIME classification relevance scores for every feature. The features are from left to right: *Valence Face, Gaze behavior, Head activity, Arms crossed, Head touch, X distance of left/right wrist and hip, Y rotation left/right elbow, Y distance of left/right wrist and hip, X rotation left/right elbow, Standard deviation head movement in X axis, Standard deviation Head X rotation, Turn hold, Continuous movement, Gesticulation.*

Feature	VAL_F	GZ_DR	HD_AC	AM_CR	HD_TH	DST_LW	DST_RW	YR_LE	YR_RE	FO_LW	FO_LW	XR_LE	XR_RE	SDX_H	SDXR_H	TN_HO	CNT_MV	EN_HA
r	−0.59	0.98	0.63	0.80	0.93	0.22	0.74	0.77	0.62	−0.04	−0.77	0.29	0.64	0.74	0.87	0.43	0.48	0.35

4 Discussion and Conclusion

We introduce a novel approach for generating textual nonverbal behavior recommendations in job interview training environments. In a pilot study, we presented the approach to participants. The results indicate that such training could be helpful to prepare for job interviews successfully. The recommendations given by the system were found to be helpful and comprehensible, and transferable to other use cases. Moreover, most participants noted that the proposed approach adds additional value to the training by giving recommendations directly on a specific behavior in a specific situation. Part of the underlying training system automatically extracts situations that could be improved and displays them alongside the recommendation presented by the virtual coach. Moreover, we examined the validity of our GAN-driven recommendation generation approach by calculating the Pearson correlation coefficient between the absolute changes of the feature values after counterfactual transformation and the importance of the features the classifier attributed to them regarding the classification result. We showed that most of the features (15 out of 18 features) had a moderate to strong correlation, which emphasizes the validity of the proposed approach. Only the two features corresponding to the relative position and movement of the left wrist and the feature representing the flexion of the left elbow presented a weak correlation. Further, we also investigated how well the generated counterfactual features can change the decision of the engagement classifier. Overall, 96.49% of the counterfactual feature vectors led to a different decision of the engagement classifier as the original input features. This indicates that our GAN-driven approach enables to generate recommendations that, when being adopted, are consistently leading to a perception of high engagement. The computational evaluation, as well as the user study, indicate that the generated recommendations are valid and helpful in the context of job interview coaching scenarios.

Acknowledgments. This work presents and discusses results in the context of the research project ForDigitHealth. The project is part of the Bavarian Research Association on Healthy Use of Digital Technologies and Media (ForDigitHealth), funded by the Bavarian Ministry of Science and Arts. Further, the work described in this paper has been partially supported by the BMBF under 16SV8688 within the MITHOS project, and by the BMBF under 16SV8493 within the AVASAG project.

References

1. Baur, T., Damian, I., Gebhard, P., Porayska-Pomsta, K., Andre, E.: A job interview simulation: social cue-based interaction with a virtual character (2013)
2. Baur, T., et al.: Context-aware automated analysis and annotation of social human-agent interactions. ACM Trans. Interact. Intell. Syst. **5**(2) (2015)
3. Cafaro, A., et al.: The noxi database: multimodal recordings of mediated novice-expert interactions. In: ICMI 2017, November 2017
4. Hoque, E., Courgeon, M., Claude Martin, J., Mutlu, B., Picard, R.W.: Mach: my automated conversation coach. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (2013)
5. Mertes, S., Huber, T., Weitz, K., Heimerl, A., André, E.: Ganterfactual-counterfactual explanations for medical non-experts using generative adversarial learning. Front. Artif. Intell. **5** (2022)
6. Minge, M., Riedel, L.: meCUE-Ein modularer Fragebogen zur Erfassung des Nutzungserlebens. In: Mensch and Computer 2013-Tagungsband, pp. 89–98. Oldenbourg Wissenschaftsverlag (2013)
7. Paulhus, D.L., Westlake, B.G., Calvez, S.S., Harms, P.D.: Self-presentation style in job interviews: the role of personality and culture. J. Appl. Soc. Psychol. **43**(10), 2042–2059 (2013)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: "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, San Francisco, CA, USA, 13–17 August 2016, pp. 1135–1144 (2016)
9. Rowold, J., Hochholdinger, S., Schaper, N.: Evaluation und Transfersicherung betrieblicher Trainings: Modelle. Methoden und Befunde, Hogrefe (2008)
10. Schneeberger, T., Scholtes, M., Hilpert, B., Langer, M., Gebhard, P.: Can social agents elicit shame as humans do? In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 164–170. IEEE (2019)
11. Takeuchi, N., Koda, T.: Initial assessment of job interview training system using multimodal behavior analysis. In: Proceedings of the 9th International Conference on Human-Agent Interaction. HAI 2021, pp. 407–411. Association for Computing Machinery, New York (2021)