Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Multimodal Trait Scoring for Video Interviews Using Neural Models with Handcrafted Features and Trait-Attention

Taichi KITAJIMA* **& Masaki UTO***
The University of Electro-Communications, Japan
*{kitajima, uto}@ai.lab.uec.ac.jp

Abstract: Interview examinations are widely used in various educational assessments to evaluate students' interpersonal skills. Automated interview scoring methods that predict scores from video recordings of interviews using artificial intelligence technologies have recently attracted considerable attention. The primary limitations of traditional methods are twofold. First, they depend solely on either handcrafted or neural features. Second, although traditional methods are typically designed as trait-scoring models, they overlook inter-trait correlations. To address these limitations, this study proposes a trait-scoring model for interview examinations that predicts multiple trait scores by incorporating inter-trait correlations and combining handcrafted features with neural features derived from pre-trained language and computer vision models.

Keywords: Automated interview scoring, trait-scoring, large language models, computer vision, multimodal data analysis

1. Introduction

Interview examinations are widely used in various educational assessments to assess students' interpersonal skills, which are difficult to evaluate through objective tests. However, manual evaluation poses significant challenges, including dependency on rater characteristics and substantial time and cost requirements. To address these issues, automated scoring of interview examinations has been gaining increasing attention (Hemamou et al., 2023; Naim et al., 2015). Automated interview scoring aims to predict scores by using artificial intelligence technology to analyze multimodal data, including video, speech audio, and dialogue text recorded during interviews. Such systems can offer consistent and timely evaluations to students, thereby enhancing both the efficiency and fairness of evaluations, and holding significant potential for educational applications.

Traditional automated scoring methods for interview evaluations have relied primarily on handcrafted features (Naim et al., 2015). However, the handcrafted features might not fully capture the complex and valuable information inherent in multimodal data. In contrast, recent approaches have employed deep learning techniques to automatically extract features that are effective for automated scoring (Hemamou et al., 2023). These methods utilize either neural features or handcrafted features for each modality. However, neural and handcrafted features are often complementary (Ridley et al., 2021), suggesting that incorporating both types of features for each modality has the potential to improve interview scoring accuracy. Additionally, interview examinations typically involve trait scoring based on multiple evaluation criteria. However, conventional methods are designed to predict each trait score independently.

To address these challenges, the present study proposes a trait-scoring model for interview examinations that leverages inter-trait correlations and integrates hybrid input from both handcrafted and neural features across all modalities. The effectiveness of the proposed method is demonstrated through an evaluation using a real-world interview dataset.

2. Proposed Method

Letting A denote the audio data, V the video data, T the dialogue text data, and H the handcrafted features derived from $\{A, V, T\}$, our task is to construct the function $\boldsymbol{S} \leftarrow f_{\boldsymbol{\theta}}(A, V, T, H)$ that predicts the trait scores $\boldsymbol{S} = \{s_1, \ldots, s_K\}$, where K represents the number of traits, s_k is the score for the k-th trait, and $\boldsymbol{\theta}$ denotes the parameters. This study assumes that each interview follows a standardized format, in which all interviewees are asked the same questions in the same order. Accordingly, each set of modality data A, V, and T includes the interviewee's responses to each of the Q questions, denoted as A_q, V_q, and T_q, where each element corresponds to the modality data for the response to the q-th question.

Our proposed model is outlined in Figure 1. It first constructs neural features from the data $\{A_q, V_q, T_q \mid q \in \{1, \ldots, Q\}\}$ by extracting embedding vectors using pre-trained neural models. Specifically, Sentence-BERT extracts textual neural features E_q^T from dialogue text data T_q, while VideoLLaMA-2 extracts neural audio-visual features E_q^V from video data V_q and audio data A_q. Additionally, GhostFaceNets are used to extract facial neural features E_q^F from video data V_q, and ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Networks) extracts neural speech features E_q^S from speech audio data A_q. For each question q, the neural features E_q^T, E_q^V, E_q^F, and E_q^S are each passed through a modality-specific fully connected layer. The resulting outputs are concatenated and further processed by two additional fully connected layers to form an embedding vector $\boldsymbol{z}_q$, where these layers share parameters across all questions. Subsequently, attention pooling is applied to the sequence of $\boldsymbol{z}_q$ to construct a neural feature for the k-th trait, defined as $\boldsymbol{h}_k = \sum_{q=1}^{Q} softmax\left(\boldsymbol{V}_k \cdot \tanh\left(\boldsymbol{W}_k \cdot z_q + \boldsymbol{b}_k\right)\right) \boldsymbol{z}_q$, where $\boldsymbol{W}_k$, $\boldsymbol{V}_k$, and $\boldsymbol{b}_k$ are trait-specific weights and bias parameters. Next, the handcrafted features H, derived from previous work (Naim et al., 2015), are concatenated with $\boldsymbol{h}_k$ to form a hybrid feature vector $\boldsymbol{r}_k$. Based on the hybrid feature $\boldsymbol{r}_k$, the trait-attention (Ridley et al., 2021) is applied to construct an inter-trait correlation-aware vector $\boldsymbol{x}_k$, defined as $\boldsymbol{x}_k = \sum_{k'=1}^{K} I(k' \neq k)[\exp(\boldsymbol{r}_k \cdot \boldsymbol{r}_{k'})/C]\, \boldsymbol{r}_{k'}$, where $I(\cdot)$ is the indicator function that returns 1 when its argument is true and 0 otherwise, and $C = \sum_{k'=1}^{K} I(k' \neq k) \exp(\boldsymbol{r}_k \cdot \boldsymbol{r}_{k'})$. Finally, the concatenated vector $\tilde{\boldsymbol{x}}_k = [\boldsymbol{x}_k, \boldsymbol{r}_k]$ is fed into a trait-specific two-layer fully connected network with sigmoid activation to produce the normalized scores for the k-th trait. The model is trained by minimizing the mean squared error loss through back-propagation after normalizing the gold-standard scores into the range $[0, 1]$. During inference, predicted scores can be linearly rescaled back to the original score scale.

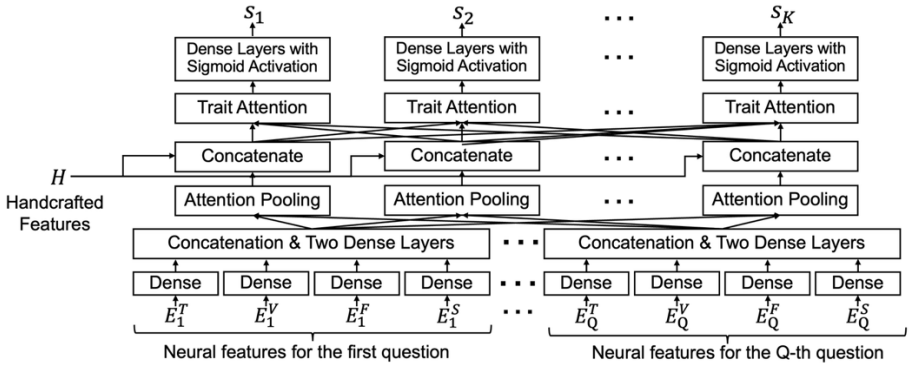

Figure 1. Architecture of the Proposed Model.

3. Evaluation Experiment

To evaluate the effectiveness of the proposed model, we conducted experiments using a real-world dataset comprising 138 interview videos of MIT students from a previous study (Naim et al., 2015). Each interview includes five standardized questions posed to all participants, with responses scored by nine human evaluators across the 16 traits, using a Likert scale ranging from 1 to 7. The average scores from nine raters were used as the gold standard.

Using this dataset, our experiment evaluated score-prediction accuracy, using 10-fold cross-validation, with the correlation coefficient as the evaluation metric. The proposed method was compared against 1) the support vector regression (SVR) and Lasso regression models, which use handcrafted features H as input, as proposed by Naim et al. (2015), 2) variants of the proposed model that exclude one of the neural features (E_q^T, E_q^V, E_q^F, or E_q^S) or the handcrafted features H from the input, and 3) the proposed model without the trait-attention mechanism. We also evaluated the accuracy of individual human raters' scoring as the average correlation between each rater's scores and the gold standard scores.

Table 1 presents the results, where $TrAtt$ denotes the trait-attention. Looking at the averaged performance, the proposed method achieves the highest accuracy. Additionally, a paired t-test between the proposed method and each of the others revealed that the proposed method significantly outperformed all others at the 5% significance level. These results indicate that integrating neural and handcrafted features across all modalities, together with accounting for inter-trait correlations, significantly improves scoring accuracy compared with the comparative models and even individual human raters.

Table 1. Experimental Results

Trait	Proposed model	Naim et al.		Proposed model without						Human rater
		SVR	Lasso	E_q^T	E_q^V	E_q^F	E_q^S	H	$TrAtt$	
1	**.702**	.650	.647	.671	.680	.676	.665	.562	.690	.616
2	**.710**	.652	.675	.643	.653	.671	.674	.540	.673	.617
3	**.739**	.669	.638	.647	.636	.694	.655	.465	.688	.707
4	**.792**	.790	.787	.776	.780	.789	.788	.556	.778	.661
5	.399	**.409**	.311	.406	.324	.323	.342	.352	.360	.600
6	.658	.656	**.710**	.612	.658	.673	.659	.540	.635	.642
7	.705	.731	**.755**	.661	.681	.726	.693	.551	.688	.661
8	**.528**	.473	.373	.491	.409	.502	.489	.404	.490	.355
9	.636	.595	**.739**	.627	.612	.589	.597	.447	.589	.628
10	.520	.501	.494	.537	**.541**	.484	.527	.199	.526	.389
11	**.513**	.477	.488	.469	.474	.487	.488	.332	.484	.444
12	.512	.458	.439	**.575**	.529	.436	.555	.229	.511	.513
13	**.631**	.587	.522	.567	.606	.595	.573	.469	**.631**	.525
14	**.594**	.588	.462	.549	.549	.543	.536	.439	.576	.510
15	.480	.399	.369	.480	.450	.352	**.494**	.300	.381	.494
16	**.636**	.618	.564	.624	.596	.591	.600	.551	.618	.591
Avg.	**.610**	.578	.561	.583	.574	.571	.583	.433	.582	.560

4. Conclusion

This study proposed a novel trait-scoring model for interview examinations and demonstrated its effectiveness. Future directions include extending the model to free-form interviews instead of standardized interview format, conducting comprehensive ablation studies and experiments with larger corpora, discussions of interpretability and endemic bias.

References

Hemamou, L., Guillon, A., Martin, J.-C., & Clavel, C. (2023). Multimodal hierarchical attention neural network: Looking for candidates behaviour which impact recruiter's decision. *IEEE Transactions on Affective Computing*, *14* (2), 969–985.

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. *Proceedings of 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 1–6.

Ridley, R., He, L., Dai, X., Huang, S., & Chen, J. (2021). Automated cross prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 13745–13753.