INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Dialogue Situation Recognition for Everyday Conversation Using Multimodal Information

Yuya Chiba[1] *and Ryuichiro Higashinaka*[2]

[1]NTT Communication Science Laboratories, Japan
[2]Graduate School of Informatics, Nagoya University, Japan
yuuya.chiba.ax@hco.ntt.co.jp, higashinaka@i.nagoya-u.ac.jp

Abstract

In recent years, dialogue systems have been applied to daily living. Such systems should be able to associate conversations with dialogue situations, such as a place where a dialogue occurs and the relationship between participants. In this study, we propose a dialogue situation recognition method that understands the perspective of dialogue scenes. The target dialogue situations contain dialogue styles, places, activities, and relations between participants. We used the Corpus of Everyday Japanese Conversation (CEJC), which records natural everyday conversations in various situations for experiments. We experimentally verified the effectiveness of our proposed method using multimodal information for situation recognition.

1. Introduction

Dialogue systems for daily living have been developed in recent years [1]. Such systems, which are used by unspecific users in various situations, must work adaptively with dialogue situations. For example, if a system can understand a situation, it can precisely recognize a user's speech by changing the ASR models based on the environment. In addition, a social conversation system can adjust its utterance style based on the relationship shared by participants. Several studies have developed systems by investigating dialogue situations. Bohus et al. [2] built a system that can refine the locations of participants by recognizing their positional relations. Utami and Bickmore [3] developed a counseling dialogue system that adapts to dialogues between intimate pairs like couples. Dialogues about surrounding environments have also been studied for spoken dialogue systems within vehicles [4]. In addition, research to conduct the dialogues about the given scenes has been actively studied [5–7]. These studies suggest that recognition and adaptation to a situation are crucial functions for future dialogue systems. However, since previous studies generally assumed that conventional systems work under limited and specific dialogue situations, research has failed to focus on the recognition of everyday dialogue situations.

On the other hand, many researchers have investigated acoustic and visual scene recognition and proposed numerous approaches based on deep learning. In acoustic scene recognition, many studies described successful results in the challenges of the Detection and Classification of Acoustic Scenes and Events (DCASE) [8, 9]. For example, Convolutional Neural Networks (CNNs) [10, 11] and a combination of CNN and Long Short-Term Memory (LSTM) were used for classification. In the area of computer vision, the MIT Indoor 67 [12], SUN 397 [13], and Places [14] databases were used as benchmarks for visual scene understanding. Similar to acoustic scene recognition, CNN-based models have been created for visual scene recognition [15–19]. The above studies aimed to estimate scenes from acoustic and visual cues. Although some of the scenes contained human interaction, dialogue behavior was not the main target of recognition.

Figure 1: *Examples of Dialogue Data of CEJC with Dialogue Situation Labels: Images are anonymized for publication.*

Unlike conventional studies, we propose a dialogue situation recognition method that understands the perspectives of dialogue scenes including the activity of and the relationship between dialogue participants. The target dialogue situations include the following elements: a conversational style, the place where the conversation took place, the kind of activity conducted while talking, and the relationship between participants. Since humans estimate such dialogue situations from audio, visual, and linguistic contexts, we propose a recognition model using multimodal information. For our experiments, we used the Corpus of Everyday Japanese Conversation (CEJC), which records conversations that naturally occur in daily life. First, we analyze the relevance between the situations to investigate an effective learning method. Then, we evaluate the performance of our proposed model by recognition experiments.

2. Corpus of Everyday Japanese Conversation

2.1. Overview

Figure 1 shows examples of the Corpus of Everyday Japanese Conversation (CEJC) [20, 21] that we used for this study. CEJC is a unique corpus that records by video cameras and IC recorders conversations that are embedded in naturally occurring activities in daily life. We used CEJC's monitor edition,

http://dx.doi.org/10.21437/Interspeech.2021-171

Table 1: *Dialogue situation labels.*

Situations	Labels
Style	Meeting, Discussion, Chat
Place	School/Workspace, Indoor, Outdoor, Other_facilities, Restaurant, Home
Activity	Leisure activities, Work, Social life, Social life with meals, Rest, Studying, Housework, Social participation, Transportation, Personal care, Meals
Relation	Social relationships, Friends, Family

which contains 50 hours of conversations. This edition includes the data of everyday conversations recorded by 40 informants selected considering the gender and age balances. Since the balance of the recorded situations was somewhat biased, it was adjusted so that the distribution became as close as possible to that of the actual conversations [20].

The audio data were recorded by IC recorders (Sony ICD-SX734) for individual participants and a central IC recorder (Sony ICD-SX1000) placed at the center of the dialogue scenes. For video recording, the Panasonic HX-A500 was used for the outdoor and moving situations, and a spherical camera (Kodak PIXPRO SP360 4K) and two portable video cameras (GoPro Hero3+) were used for other situations. In particular, the single GoPro Hero3+ was employed for many recordings. The corpus contains transcriptions with detailed annotations with speaker labels and the starting and ending times of the utterances.

2.2. Labels of Dialogue Situations

Many kinds of metadata were attached to each record. As labels of dialogue situations, we used, from the metadata, a conversational style (Style), the place where the conversation took place (Place), the kind of activity conducted while talking (Activity), and the relationship between the participants (Relation). Since the number of samples of some labels was small, we combined some of the labels. For the labels under "Place," we combined "School" and "Workplace" into "School/Workplace." The labels related to "Indoor" were integrated into a single category. The labels related to "Facility" other than "Restaurant" were combined into "Other_facilities." For labels under "Activity," we combined the following: "Leisure activities" and "Leisure activities with transportation," "Housework" and "Housework with meals," and "Social participation" and "Social participation with meals." For the labels under "Relation," those except for "Family" and "Friends" were merged into "Social relationships." The definitive dialogue situation labels are summarized in Table 1.

2.3. Analysis of Relevance Among Situation Labels

The target dialogue situations are considered interrelated. The dialogue situation recognition model can be trained effectively by capturing such relations. Here, we calculated the affinity scores [22] among labels to investigate the relevance among the situations. Affinity score $A(X, Y)$ between labels X and Y is calculated as:

$$A(X,Y) = \frac{P(X,Y)}{P(X) + P(Y) - P(X,Y)}. \quad (1)$$

The higher an affinity score is, the more likely that X and Y will co-occur and not appear separately. Figure 2 shows a heatmap

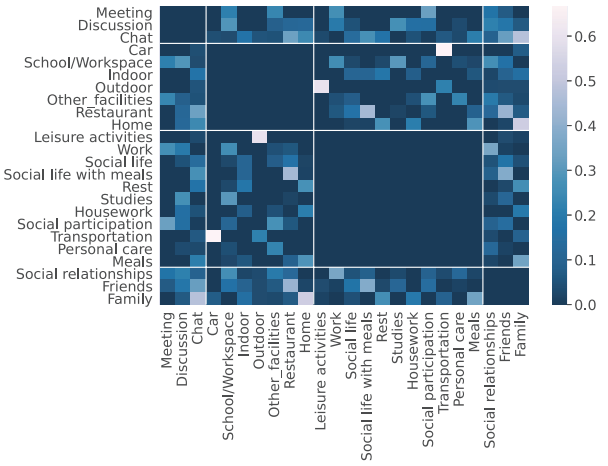

Figure 2: *Affinity matrix: lighter color represents higher affinity and darker color represents lower affinity.*

of the affinity scores between labels. A lighter color represents higher affinity, and a darker color represents lower affinity. We found relationships between the situations; higher affinity scores were observed between "Car" and "Transportation," and between "Family" and "Chat." These are intuitive results considering our daily conversations. We also found other reasonable relations. For example, "Chat" tends to co-occur with "Restaurant" and "Friend," and "Work" tends to co-occur with "Meeting" and "Social relationships."

These results suggest that considering the relevance between dialogue situations is effective for training a situation recognition model. Therefore, we conducted multi-task learning to capture the relationships between situations in addition to employing multimodal features in our experiments.

3. Multimodal Dialogue Situation Recognition Model

Figure 3 shows an overview of our proposed dialogue situation recognition model, which takes audio signals, video data, and utterance transcriptions as input. First, the audio, visual, and linguistic inputs are converted to embedding vector sequences by ResNet50 [23], VGGish [11], and BERT [24]. Each embedding vector sequence is sent to individual Uni-directional Gated Recurrent Units (Uni-GRUs), which output representation vectors at every time step. The outputs of the Uni-GRUs at the final step are concatenated and input to the output layer corresponding to each situation. Finally, the output layer outputs the prediction results. In this study, the dialogue situations of "Style," "Place," "Activity," and "Relation" are dealt with as individual tasks. Since analysis of the affinity scores revealed that the dialogue situations are interrelated, we introduced multi-task learning to train the network. In multi-task learning, the output of the Uni-GRUs is connected to four output layers.

In this study, we assumed that the system observes a dialogue scene for a short time and estimates its dialogue situation. Therefore, the inputs are segments that were cropped from the original recording. We used the embedding vectors of the utterances as linguistic features. Since the segments generally contain multiple utterances, we input the utterance vectors in the order of the starting times to the language Uni-GRU. The parameters of ResNet50, VGGish, and BERT were frozen in

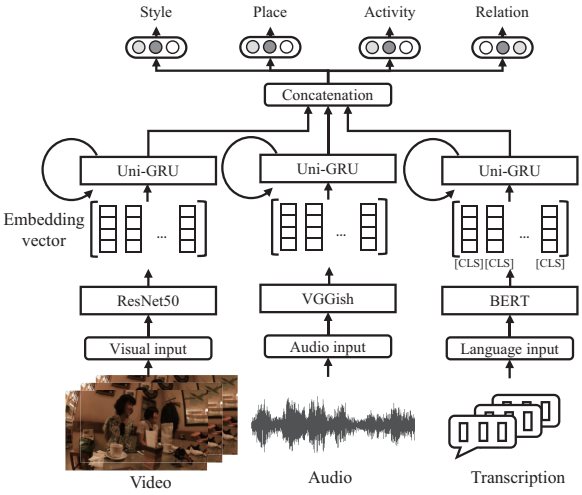

Figure 3: *Overview of dialogue situation recognition model: Proposed model takes video, audio, and text inputs and outputs prediction results of conversational style, place, activity, and relationship between participants. Network is trained by multi-task learning.*

this paper. When conducting single-task learning for comparison, only the final layer that corresponds to the target task was connected to the model.

4. Experiments

4.1. Setup of Experimental Data

The monitor edition of CEJC contains 126 records. We separated them into training, development, and test sets. We made the ratio of labels as similar as possible between sets. The amounts of records of the training, development, and test sets were 76, 25, and 25, respectively. Records derived from the same dialogue were assigned to the same set.

Each record was divided into segments of 30 seconds. The audio signals of the segments were converted to monaural and input to the VGGish to obtain 128-dimensional embedding vector sequences. The video data were converted to 2,048-dimensional vector sequences by ResNet50. In this paper, we used audio data recorded by the central IC recorder (Sony ICD-SX1000) placed in the center of the dialogue scene. Here, the condition of the recording equipment differed from informant to informant. If the record lacked the audio data of the central IC recorder, we used the mixed-down audio of the individual IC recorders. We used video clips from the Panasonic HX-A500 for the outdoor data and video clips from the single GoPro Hero3+ for the other data. We gave a zero-vector for the records that did not have any video data. The audio-visual features were resampled every second. The utterances were segmented using MeCab[1] with the JUMAN dictionary[2], and converted to 768-dimensional vectors by BERT[3]. The embedding vectors of the CLS tokens were input to the network by starting times. The utterances straddling the segments were discarded. The drop rate of the utterances was 2.88% when separated into segments of 30 seconds.

Finally, the amounts of samples of the training, development, and test sets were 3,371, 1,537, and 1,176. Here, we identically labeled the segments as those of the original record.

4.2. Conditions of Training Network

The network was trained to minimize the weighted cross-entropy loss. The losses in each label were multiplied by a weight that is proportional to the inverse of the sample size. The number of Uni-GRU layers was 1. The optimization method was Adam with a learning rate of 0.0005. The minibatch size was 32 and the maximum number of epochs was 100. We investigated the classification performance while changing the number of nodes of the hidden layers as 16, 32, 64, and 128.

The condition that yielded the best macro average of F1-score for the validation set was used for the definitive evaluation. In the following section, we show the recognition results for the test set.

4.3. Conditions of Multi-Task Learning

The network was trained to minimize the weighted sum of the task losses in multi-task learning. We dynamically determined the weights of the tasks by Dynamic Weighted Average (DWA) [25]. In this method, the weight of task k was determined as follows:

$$\lambda_k(t) = \frac{K \exp\left(w_k(t-1)/T\right)}{\sum_i \exp\left(w_i(t-1)/T\right)}, w_k(t-1) = \frac{\mathcal{L}_k(t-1)}{\mathcal{L}_k(t-2)} \tag{2}$$

Here, $w_k(\cdot)$ is the relative descending rate obtained as the ratio of $\mathcal{L}_k(t-1)$ and $\mathcal{L}_k(t-2)$, which are the losses of task k at the $t-1$-th and $t-2$-th trials. T is a temperature that controls the softness of the task weighting. We employed $T = 2$, which is identical to the conventional study [25]. K is a coefficient that ensures $\sum_i \lambda_i = K$. In this paper, the task corresponds to the individual dialogue situation. Therefore, we set K to the number of kinds of situations ($K = 4$). We processed the calculation of the relative descending rate every epoch and initialized $w_k(t) = 1.0$ for $t = 1, 2$.

In multi-task learning, we selected the hyper-parameters based on the macro average F1-scores of the four tasks for the validation set.

4.4. Human Evaluation

To investigate the upper-bound of the recognition results, we conducted a human evaluation for the test set. Five segments for each record were selected randomly to reduce the evaluation cost. The number of samples was 125 (5 video clips $\times$ 25 records). An expert watched the video clips in random order, and selected an appropriate label from the list shown in Table 1 for each dialogue situation. The expert judged the situations only from the information observed in the target video clip.

5. Results of Situation Recognition

Table 2 shows the recognition results. The table shows the macro averages of precision, recall, and F1-score for each task. "Average" in the table is the overall results representing the averages of the four tasks. A, V, and L are the results when using the acoustic, visual, and linguistic information, respectively. "Chance" represents the chance-level results. In addition, we showed the results of the human evaluation as "Human." Here "F1" is not always equal to the harmonic mean of the precision

[1] http://mecab.sourceforge.jp

[2] https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN

[3] https://alaginrc.nict.go.jp/nict-bert/index.html

Table 2: *Results of situation recognition experiments: Table shows macro average of classification results for each task. Average is macro averages of four tasks. A, V, and L represent acoustic, visual, and linguistic features, respectively. ST and MT are single- and multi-task learning results. Pre., Rec., and F1 are precision, recall, and F1-score. Chance shows the chance-level results and Human shows human evaluation results.*

		Style			Place			Activity			Relation			Average		
Mod.	Cond.	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Chance	–	0.209	0.333	0.257	0.053	0.143	0.077	0.017	0.091	0.028	0.158	0.333	0.214	0.109	0.225	0.144
A	ST	0.411	0.402	0.403	0.485	0.559	0.492	0.234	0.251	0.232	0.508	0.520	0.511	0.410	0.433	0.410
V	ST	0.466	0.438	0.443	0.444	0.525	0.431	0.283	0.337	0.291	0.511	0.495	0.474	0.426	0.449	0.410
L	ST	0.497	**0.545**	0.498	0.209	0.218	0.191	0.151	0.159	0.128	0.515	0.484	0.481	0.343	0.352	0.325
A + V	ST	0.470	0.398	0.376	**0.548**	0.558	0.493	0.398	**0.440**	0.381	0.451	0.463	0.450	0.467	0.465	0.425
A + L	ST	0.439	0.415	0.422	0.528	**0.606**	0.530	0.215	0.229	0.211	0.523	0.526	0.510	0.426	0.444	0.418
L + V	ST	0.524	0.505	0.509	0.486	0.537	0.466	0.310	0.337	0.273	0.551	0.549	0.542	0.468	0.482	0.448
A + L + V	ST	0.539	0.425	0.440	0.538	0.596	**0.532**	0.328	0.418	0.344	0.527	0.509	0.503	0.483	0.487	0.455
A + L + V	MT	**0.546**	0.513	**0.519**	0.520	0.558	0.488	**0.458**	0.428	**0.390**	**0.573**	**0.554**	**0.548**	**0.524**	**0.513**	**0.486**
Human	–	0.844	0.669	0.700	0.64	0.667	0.642	0.617	0.552	0.500	0.848	0.823	0.824	0.737	0.678	0.667

and recall in the table since these scores are calculated as the macro average of the labels.

First, we found that the performance of the examined models surpassed "Chance." These results show that the network was appropriately trained to estimate the dialogue situations from the acoustic, visual, and linguistic cues. From the comparison between modalities, we found that the acoustic and visual information contributed to the "Place" and "Activity" estimation. In particular, the "Place" result coincided with the findings in conventional acoustic and visual scene analysis, suggesting that location can be estimated using visual and acoustic cues. In contrast, the linguistic information effectively estimated "Style" and "Relation." This result is appropriate because humans usually tailor such linguistic choices as phrasing and vocabulary based on the conversational style or the relationship to the person with whom they are speaking. For example, a casual speaking style is chosen for chat-talks with friends and a more formal tone for a discussion with a boss. "Relation" can also be estimated from acoustic and visual cues. This result suggests that the positions of participants were affected by their relationships.

In addition, the performance improved by combining multimodal information. In the single-task learning condition, the best overall result was obtained when using all the modalities; the average F1-score was 0.455. These results suggest that our proposed model robustly estimates dialogue situations using multimodal information. Besides, it is considered that the relevance between the situations can also be captured to some extent even with simple feature fusion. Multi-task learning further improved the performance of the proposed model; we obtained an average F1-score of 0.486. Here we conducted a McNemar test with Bonferroni correction for the wins and losses between the multi-task learning result and other multimodal results. This step is equivalent to comparing accuracy. For "Style," we observed significant differences ($p < 0.05$) among all combinations. For "Place," a significant difference was observed in the comparison to L+V and A+L+V (ST). For "Activity," significant differences were observed in the comparison to the bimodal results. For "Relation," significant differences were observed in the comparison to A+V. These results indicate that multi-task learning effectively improved the performance of "Style" in particular.

However, such definitive results failed to reach the levels achieved by human evaluations. For example, the evaluator successfully distinguished dialogue styles from the conversation content, even in 30-second movie clips, indicating that the proposed model insufficiently represented conversational information. For example, the humans seemingly distinguished among relationships by inferring participants' occupations or age. To correctly classify "Relation," the situation recognition model must capture wider dialogue contexts that include various personal attributes of participants. Besides, not only global visual information but also such local information as participant behaviors are effective for recognition. In future studies, we will improve our method's performance by incorporating the knowledge of recent action recognition [26], acoustic event detection [27], and age estimation [28] methods.

On the other hand, the result of the human evaluations was 0.667 in the F1-score, which was lower than our assumption. One reason for the low score is the difficulty of the classification of "Activity" and "Place." As shown in Table 1, there are overlaps and ambiguities in the class labels of CEJC. The evaluator pointed out that she was unsure which "Activity" to classify a meeting at a restaurant. In addition, "School/Workspace," "Indoor," and "Home" tended to be confusing. Therefore, it is necessary to re-examine the taxonomy of the labels.

6. Conclusion

This paper proposed a comprehensive dialogue situation recognition method using multimodal information. We used CEJC, which is a large-scale corpus of everyday Japanese conversation, for our experiments. From the analysis of its metadata, we clarified that dialogue situations are mutually related. Recognition experiments showed that performance was improved using multimodal information. Multi-task learning that focused on the relevance between situations further improved the recognition results. We obtained an F1-score of 0.486. Although this result surpassed chance-level results, it did not reach the human evaluation results.

In future studies, we will examine the features related to the action [26] and the attribution [28, 29] of the participants. In addition, we will also incorporate the proposed model into a response generation model (e.g., [30]) to achieve situation-aware response generation.

7. Acknowledgement

Funding was provided by Grants-in-Aid for Scientific Research (Grant No. JP19H05692). We thank the National Institute for Japanese Language and Linguistics for letting us use the CEJC corpus.

8. References

[1] A. Ram, R. Prasad, C. Khatri *et al.*, "Conversational AI: The science behind the Alexa Prize," *arXiv preprint arXiv:1801.03604*, pp. 1–18, 2018.

[2] D. Bohus, S. Andrist, and E. Horvitz, "A study in scene shaping: Adjusting f-formations in the wild," in *Proc. AAAI Fall Symposium*, 2017, pp. 1–7.

[3] D. Utami and T. Bickmore, "Collaborative user responses in multiparty interaction with a couples counselor robot," in *Proc. HRI*, 2019, pp. 294–303.

[4] T. Misu, "Situated reference resolution using visual saliency and crowdsourcing-based priors for a spoken dialog system within vehicles," *Computer Speech & Language*, vol. 48, pp. 1–14, 2018.

[5] H. Alamri, V. Cartillier, A. Das *et al.*, "Audio visual scene-aware dialog," in *Proc. CVPR*, 2019, pp. 7558–7567.

[6] S. Kumar, E. Okur, S. Sahay, J. Huang, and L. Nachman, "Leveraging topics and audio features with multimodal attention for audio visual scene-aware dialog," *arXiv preprint arXiv:1912.10131*, 2019.

[7] X. Lin, G. Bertasius, J. Wang, S.-F. Chang, D. Parikh, and L. Torresani, "Vx2text: End-to-end learning of video-based text generation from multimodal inputs," *arXiv preprint arXiv:2101.12059*, pp. 1–14, 2021.

[8] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Proc. IEEE AASP*, 2013, pp. 1–4.

[9] H. Zhu, C. Ren, J. Wang, S. Li, L. Wang, and L. Yang, "DCASE 2019 challenge task1 technical report," *Tech. Rep. DCASE2019 Challenge*, pp. 1–4, 2019.

[10] L. Yang, X. Chen, and L. Tao, "Acoustic scene classification using multi-scale features," in *Proc. DCASE*, 2018, pp. 29–33.

[11] S. Hershey, S. Chaudhuri, D. Ellis *et al.*, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, 2017, pp. 131–135.

[12] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. CVPR*, 2009, pp. 413–420.

[13] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. CVPR*, 2010, pp. 3485–3492.

[14] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[15] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognition*, vol. 102, pp. 1–18, 2020.

[16] A. Matei, A. Glavan, and E. Talavera, "Deep learning for scene recognition from visual data: a survey," in *International Conference on Hybrid Artificial Intelligence Systems*, 2020, pp. 763–773.

[17] D. Linsley, D. Shiebler, S. Eberhardt, and T. Serre, "Learning what and where to attend," *arXiv preprint arXiv:1805.08819*, pp. 1–21, 2018.

[18] S. Jiang, G. Chen, X. Song, and L. Liu, "Deep patch representations with shared codebook for scene classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1s, pp. 1–17, 2019.

[19] H. Seong, J. Hyun, and E. Kim, "FOSNet: An end-to-end trainable deep neural network for scene recognition," *IEEE Access*, vol. 8, pp. 82 066–82 077, 2020.

[20] H. Koiso, Y. Den, Y. Iseki, W. Kashino, Y. Kawabata, K. Nishikawa, Y. Tanaka, and Y. Usuda, "Construction of the corpus of everyday Japanese conversation: An interim report," in *Proc. LREC*, 2018, pp. 4259–4264.

[21] H. Koiso, H. Amatani, Y. Iseki *et al.*, "Design, evaluation, and preliminary analysis of the monitor version of the corpus of everyday Japanese conversation," *NINJAL Research Papers (in Japanese)*, no. 18, pp. 17–33, 2020.

[22] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. VLDB*, vol. 1215, 1994, pp. 487–499.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[25] S. Liu, E. Johns, and A. Davison, "End-to-end multi-task learning with attention," in *Proc. CVPR*, 2019, pp. 1871–1880.

[26] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. CVPR*, 2017, pp. 6299–6308.

[27] K. Imoto and N. Ono, "Acoustic topic model for scene analysis with intermittently missing observations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 367–382, 2018.

[28] M. Xia, X. Zhang, L. Weng, Y. Xu *et al.*, "Multi-stage feature constraints learning for age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2417–2428, 2020.

[29] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, and C. Yan, "Recurrent attention model for pedestrian attribute recognition," in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 9275–9282.

[30] E. Smith, M. Williamson, K. Shuster, J. Weston, and Y.-L. Boureau, "Can you put it all together: Evaluating conversational agents' ability to blend skills," *arXiv preprint arXiv:2004.08449*, pp. 1–10, 2020.