MCIC: Multimodal Conversational Intent Classification for E-commerce Customer Service

Shaozu Yuan[1], Xin Shen[1,2], Yuming Zhao[1], Hang Liu[1],
Zhiling Yan[1], Ruixue Liu[1], Meng Chen[1]

[1] JD AI, Beijing, China
{yuanshaozu, zhaoyuming3, liuhang55,
yanzhiling, liuruixue, chenmeng20}@jd.com
[2] Australian National University, Canberra, Australia
u6498962@anu.edu.au

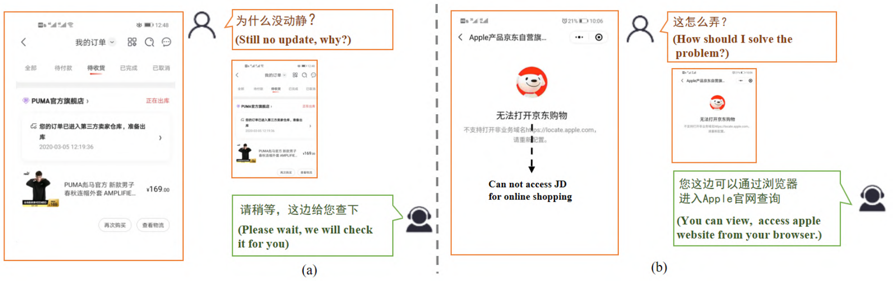
Abstract. Conversational intent classification (CIC) plays a significant role in dialogue understanding, and most previous works only focus on the text modality. Nevertheless, in real conversations of E-commerce customer service, users often send images (screenshots and photos) among the text, which makes multimodal CIC a challenging task for customer service systems. To understand the intent of a multimodal conversation, it is essential to understand the content of both text and images. In this paper, we construct a large-scale dataset for multimodal CIC in the Chinese E-commerce scenario, named MCIC, which contains more than 30,000 multimodal dialogues with image categories, OCR text (the text contained in images), and intent labels. To fuse visual and textual information effectively, we design two vision-language baselines to integrate either images or OCR text with the dialogue utterances. Experimental results verify that both the text and images are important for CIC in E-commerce customer service.

Keywords: Conversational Intent Classification · Multimodal Dataset.

1 Introduction

Conversational customer service has been widely deployed and achieved great success in recent years, which facilitates the development of dialogues that help users to achieve their goals. To better understand dialogues and serve for some downstream tasks, conversational intent classification (CIC) that aims at identifying the user intents behind their utterances has become increasingly important.

Previous works of CIC mainly based on the text modality [16,17,18,19], but E-commerce customer service naturally contains lots of multimodal conversations. Users usually leverage images to help illustrate their goals or supplement more information of the conversation background. As shown in Figure 1(a), it is difficult to infer the intent of *delivery delay* from the utterance "*Still no update, why?*" without the image that shows the customer paid for a coat and was waiting for delivery. Therefore, images are ubiquitous in such conversations and crucial to intent classification. Despite the importance, less attention has been devoted to multimodal CIC in real scenarios. One challenge for this task is the lack of a large-scale annotated dataset, since collecting multimodal dialogues and labelling intents for them are much more time-consuming and labor-intensive.



为什么没动静？
(Still no update, why?)

请稍等，这边给您查下
(Please wait, we will check it for you)

这怎么弄？
(How should I solve the problem?)

您这边可以通过浏览器
进入Apple官网查询
(You can view, access apple website from your browser.)

我的订单
全部
待付款
待收货
已完成
已取消
PUMA官方旗舰店 ›
正在出库
您的订单已进入第三方卖家仓库，准备出库
2020-03-05 12:19:36
PUMA彪马官方 新款男子春秋连帽外套 AMPLIFIE...
¥169.00
再次购买
查看物流
(a)
Apple产品京东自营旗...
无法打开京东购物
不支持打开非业务域名https://locate.apple.com，请重新配置。
Can not access JD for online shopping
(b)

Fig. 1. Examples of multimodal conversations in E-commerce customer service. User utterances are in orange and staff utterances are in green.

In this paper, we collect a large-scale dataset for **M**ultimodal **C**onversational **I**ntent **C**lassification, named MCIC. MCIC consists of more than 30,000 dialogues that focus on after-sales topics between users and customer service staff and contain at least one image per session. Moreover, there are more than 200 intents in the dataset, which cover most user intents in the E-commerce scenario. Specifically, over 50% images are screenshots and 80% images has some text. We observe that text in images carries essential information for image understanding. As shown in Figure 1(b), the text extracted from the image "*can not access JD for online shopping*" reveals the situation in which the user meets difficulty when accessing JD online. To better assist intent classification with such images, we also apply an Optical Character Recognition (OCR) model to extract the text from images.

Moreover, we design two models based on the BERT architecture [28], VisualBERT and OCRBERT, to capture the interaction between user utterances and visual signals, i.e., images and OCR text. VisualBERT integrates the text and visual features extracted from ResNet [13] model to infer the intent, while OCRBERT captures the interaction between OCR text and dialogue utterances via BERT to obtain the intents. Since there is no pre-trained vision-language model released to the public in Chinese version, we use the single modality model BERT [28] as our baseline. Compared with BERT, both models achieve improvement under the automatic evaluation.

In short, our contribution is twofold: (1) We construct a large-scale dataset for multimodal conversational intent classification, with various annotated labels, including image categories, OCR text, and intents. (2) We design two BERT-based baselines to utilize multimodal information and conduct experiments to prove that it is necessary to integrate both visual and textual features into models for MCIC.

2 Related Work

For conversational intent classification (CIC), most datasets are constructed for task-oriented dialogues. MultiWOZ [16] is a large-scale multi-domain dataset that consists of around 10K crowd-sourced human-to-human dialogues with 13 intent types. MDC [17] consists of human-annotated conversational data in three domains (movie-ticket booking, restaurant reservation, and taxi booking) with 11 intents. SGD [18] dataset contains over 16K multi-domain conversations spanning 16 domains and 86 intents. CrossWOZ [19] is the first large-scale Chinese multi-domain Wizard-of-Oz dataset proposed recently, and has 6,012 dialogues covering 6 intents. E-IntentConv [1] contains real online E-commerce conversations between users and staff with diverse and complex intents. The above datasets are text-based, and their intents are labelled only grounded on pure-text conversations.

Dataset	Dialogue	Image	Intent	# Dialogues	# Average Turns	# Images	# Intents
MultiWOZ [16]	✓	✗	✓	8,438	13.7	0	13
MDC [17]	✓	✗	✓	10,087	7.5	0	11
SGD [18]	✓	✗	✓	16,142	20.4	0	86
CrossWOZ [19]	✓	✗	✓	5,012	16.9	0	6
E-IntentConv [1]	✓	✗	✓	1,134,487	20	0	289
Portraits of Politicians [20]	✗	✓	✓	0	0	1,124	9
Motivations [21]	✗	✓	✓	0	0	10,191	256
MDID [22]	✗	✓	✓	0	0	1,299	8
Intentonomy [23]	✗	✓	✓	0	0	14,455	28
SIMMC 2.0 [15]	✓	✓	✓	11,244	10.4	1,566	10
MCIC	✓	✓	✓	**30,716**	**8.4**	**30,716**	**212**

Table 1. The overview of related datasets for intent classification.

Some works also focus on intent recognition from images, which is called image intent classification. [20] defined 9 dimensions of persuasive intents of a politician implied through a photo. [21] collected a new dataset of people performing actions annotated with likely motivations. Intentonomy [23] comprises 14K images covering a wide range of everyday scenes. These images were manually annotated with 28 intent categories derived from a social psychology taxonomy. Image content may not enough to explain its meanings, then some researchers added the text (e.g., captions) to assist the image understanding. For example, [22] tried to extract intents from multimodal data like Instagram posts and proposed MDID dataset that consists of 1299 public Instagram posts with 8 intents.

Different from all the datasets above, our dataset combines conversations and images to fulfill multimodal CIC. The most similar work to ours is SIMMC 2.0 [15] that labels task-oriented dialogues with 10 intents, while our dataset consists of large-scale dialogues with 200+ intents in E-commerce scenario, and the annotation of intents fully takes visual information into consideration. We summarize these mentioned datasets in Table 1.

3 Dataset Construction

3.1 Data Collection and Pre-Processing

We sample 500,000 conversations between users and customer service staff from **JD.com**[3], which is a leading online shopping platform that sells over tens of thousands of brands and over 40.2 million items. To protect the users' privacy and remove the invalid textual conversation, the data is processed by following three steps: (1) To guarantee each conversation session contains at least one image, we first remove pure-text conversations. (2) For the purpose of protecting an individual's privacy while maintaining the integrity of the conversation, we replace sensitive text, like user name, user ID, address and telephone number, with special tokens <NAME>, <ID>, <ADDRESS> and <TEL> respectively. (3) To filter user-sensitive images, such as some screenshots with personal name, telephone number and home address, we extract textual information from the image and detect if it contains sensitive information with regular expression. Then the conversations with the detected image will be removed from the dataset.

3.2 Data Annotation

After pre-processing, we obtain 30,716 valid conversations. To promote relevant research and make the dataset more valuable, we exploit both automatic and manual methods to annotate the dataset with different labels, including intents, image categories, and OCR text from images.

Intent Annotation. In the context of E-commerce, the communications between users and customer service staff are involved in a multimodal setting, where users tend to apply both images and text to express their intents and goals. These intents, which covers 212 types, are indirect and diverse. Ten crowdsourcing annotators are hired to not only understand the meaning of the text and image in a session but also select the right intent from intent candidates. We observe that the image intent is usually consistent with the intent of its surrounding text. To reduce the workload of annotators, we train a text-based intent classifier with 700,000 labelled context to provide intent candidates for each user utterance. The annotator could choose an proper intent from the candidates, or infer the intent from the image and text if the actual one is not included in those candidates. The intent selected the most is regarded as the final result.

[3] https://www.jd.com/

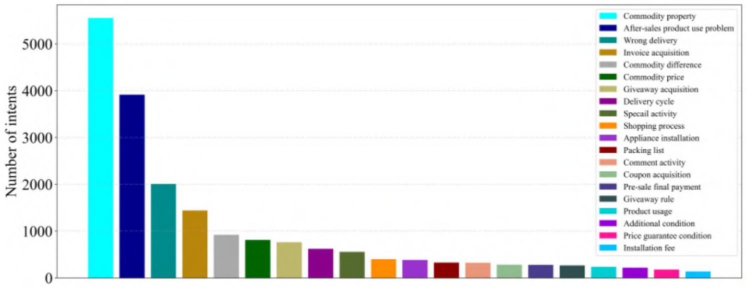


Fig. 2. The intent distribution of MCIC dataset.

Image Category Identification. Considering that the image categories may be beneficial for MCIC and a reliable CNN feature extractor requires images with category labels for training, we also provide image categories in our dataset. We divide all the images into 26 types, such as app screenshots and commodity pictures, and invite three professional customer service staff to annotate the image categories.

Optical Character Recognition. OCR is a well-studied task in the literature, thus we choose two open-source models that have been utilized in many scenarios to extract the text information from images. Specifically, we first use EAST [8] model to detect the text blocks, and then the identified blocks are fed into RCNN [10] to obtain text in images. To ensure the quality of extracted text, the output texts under the threshold of 0.5 are dropped.

3.3 Data Statistics and Demonstration

Total dialogue sessions	30,716
Total utterances	844,661
Average utterances per session	27.5
Total turns	257,024
Average turns per session	8.4
Max turns	218
Min turns	1

Table 2. Statistics of conversations.

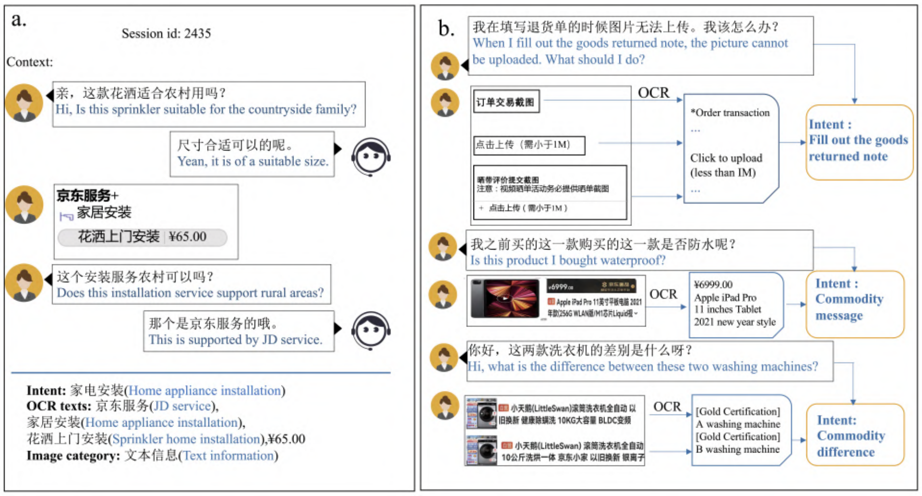
Total images	30,716
Total OCR texts	678,503
Average OCR texts	26.0
Max OCR texts	275
Min OCR texts	0
Images with OCR texts	26,123
Ratio of images with OCR texts	85.05%

Table 3. Statistics of visual information.

In Table 2, it can been seen that MCIC dataset includes 30,716 multi-turn conversation sessions and 844,661 utterances[4]. The number of turns[5] for a session ranges from 1 to 218, and the average number of turns per session is 8.4. In Table 3, the number of extracted OCR texts ranges from 0 (no extracted Chinese characters from the image) to 275, and 85.05% images have extracted text. The OCR text is provided as auxiliary information to facilitate the performance of MCIC.

[4] An image in a session is regarded as an utterance in our multimodal dataset.

[5] A "turn" in a conversation is marked by one back-and-forth interaction: the user speaks and the staff follows, or vice-versa.



Session id: 2435

Context:

亲，这款花洒适合农村用吗？
Hi, Is this sprinkler suitable for the countryside family?

尺寸合适可以的呢。
Yean, it is of a suitable size.

京东服务+
家居安装
花洒上门安装 ¥65.00

这个安装服务农村可以吗？
Does this installation service support rural areas?

那个是京东服务的哦。
This is supported by JD service.

Intent: 家电安装(Home appliance installation)
OCR texts: 京东服务(JD service),
家居安装(Home appliance installation),
花洒上门安装(Sprinkler home installation),¥65.00
Image category: 文本信息(Text information)

我在填写退货单的时候图片无法上传。我该怎么办？
When I fill out the goods returned note, the picture cannot be uploaded. What should I do?

我之前买的这一款购买的这一款是否防水呢？
Is this product I bought waterproof?

你好，这两款洗衣机的差别是什么呀？
Hi, what is the difference between these two washing machines?

Fig. 3. Overview of MCIC Dataset. (a) A sample from MCIC dataset with intent, OCR text, and image category. (b) Three summarised samples with different user intents.

Figure 2 illustrates the distribution of Top-20 intents with the most occurrences. It can be seen that the Top-5 intents are: commodity property, after-sales product use problem, wrong delivery, invoice acquisition, and commodity difference. For online shopping, people often ask about the attributes of a commodity in the form of text and images. Compared with simply using text to ask, the usage of images can not only reduce the communication time but also make it easier for the customer service staff to understand user intents.

We further show a sample of MCIC dataset in Figure 3(a), which has five items including session id, context, intent, OCR text, and image category. The context is multimodal dialogue (the image is saved as URL link) between the user and customer service staff[6]. Each session id is unique to distinguish different examples. The other three items are annotations. The intent, annotated by annotators according to the whole context, is the target of MCIC task. The OCR text and image category are auxiliary information of the image.

In Figure 3(b), we present three typical samples whose intents are “fill out the goods”, “commodity message” and “commodity difference”, respectively. As shown in these samples, it is challenging to understand user intents with textual information solely, which indicates the necessity of combining both textual and visual information for conversation understanding. Moreover, we also highlight the important OCR text from the images, and it shows that these texts contribute to the MCIC task.

[6] Because of the space limitation, we only show part of context in the figure.

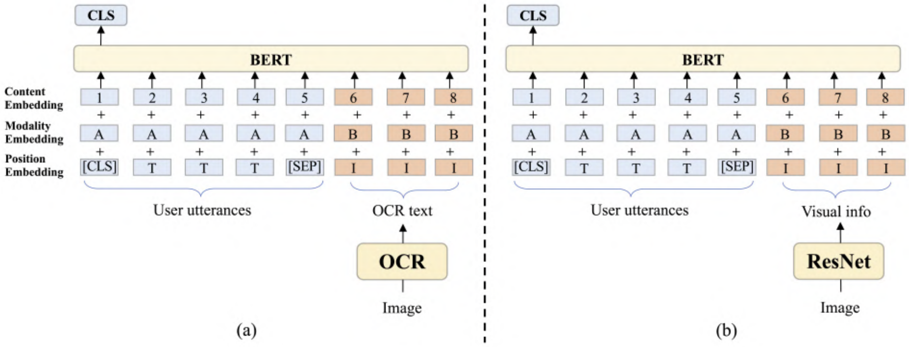


Fig. 4. Overview of our two proposed baselines, (a) OCRBERT and (b) VisualBERT. We utilize different colors to denote multimodal inputs in the embedding layer.

4 Framework

Since existing open-source vision-language pretrained models are in English version, there is no suitable model for our Chinese MCIC task. Previous work [27] has shown that BERT [28] can not only be beneficial for NLP-related tasks but also facilitate the model performance in multimodal applications. Therefore, we propose two BERT-based baselines to accommodate multimodal inputs and evaluate their performance of intent classification on MCIC dataset. Figure 4 illustrates the architecture of our proposed models and we introduce model details in the following parts. **Note that** the user utterances in model input are 2 surrounding user utterances of the input image, one before and one after the image (marked with red boxes in Figure 3(a)).

4.1 Input Embedding

The original BERT model is designed for language modeling, to adopt it multimodal inputs, we modify the embedding layer before feeding them to the model. There are three kinds of embeddings involved in our model: **Position Embedding** is the index of the token in the flattened multimodal sequence (user utterances are followed by the visual information), which is the same as that of BERT. **Modality Embedding**, used to distinguish different modalities, takes two possible values: “A” for the user utterances, and “B” for the input image. **Content Embedding** is composed of utterance tokens and image features.

4.2 Backbone

Formally, the given multimodal features x_T (text features) and x_I (image features) are first fed into N self-attention layers to obtain more interactive features as follows:

$$
\begin{aligned}
e_{ij} &= (W_Q[x_T;x_I])(W_K[x_T;x_I])^T/\sqrt{d}, \\
a_{ij} &= \exp(e_{ij})/\sum_r \exp(e_{ir}), \\
c &= \sum_j a_{ij} W_V[x_T;x_I],
\end{aligned}
\tag{1}
$$

where W_Q,W_K, W_V are learnable parameters, a_{ij} represents the attention weight, c is attentive feature, and d is the dimension of W_Q.

To classify the intent, we experimentally adopt the representation z of the special token "[CLS]", and feed it into a fully-connected layer with a softmax activation function:

$$
\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K, \tag{2}
$$

where K is the number of intents and $\sigma(z_i)$ represents the predicted score of each intent.

4.3 OCRBERT

OCRBERT explicitly incorporates OCR text from the images to promote the multimodal intent classification. We adopt pretrained BERT that has been already pretrained on a large-scale Chinese corpus to help obtain contextualized character representations. For model input, we directly concatenate "[CLS]" with the characters T in user utterances and the OCR text I extracted from the image, where the text characters and the OCR characters are split by a special token "[SEP]" and different OCR texts are distinguished with "|". Finally, the final hidden state of the "[CLS]" token is used for intent classification.

4.4 VisualBERT

VisualBERT is designed to capture the interaction between textual and visual information. VisualBERT is also based on BERT whose self-attention layers can automatically learn different levels of alignment between two modalities. The visual features of an image are extracted from a pretrained ResNet [13] model. Here, we design two variants by exploiting different visual features. VisualBERT(*) utilizes the original 8×8×2048 visual features, while VisualBERT adopts the pooling feature with 2048 dimensions. To learn a cross-modality relationship between the image features and linguistic tokens, their embeddings are fed into a multi-layer bidirectional transformer encoder. Similarly, the final hidden state of the special token "[CLS]" is used for intent classification.

5 Experiments

In this section, we introduce experimental settings, experimental results, and the results of case study and attention visualization.

5.1 Experimental Settings

We split the MCIC dataset with 27,000 samples for training, 2,000 for validation, and 1,716 for test in our experiments. In order to assess the performance of different models, we use "Accuracy" as the quantitative metric for automatic evaluation.

Images in MCIC dataset are first resized to 256×256 pixels. Then we extract visual representations with ResNet [13] pretrained with image category labels. Specifically, we extract two kinds of visual features: the local feature and the global feature. For the former, we split the original image into 64 regions, and the local feature is extracted from the last convolution layer. The global feature, as the global representation of the whole image, is extracted from the pooling layer. The local feature is of 8×8×2048 dimensions, while the global feature is of 2048 dimensions. To avoid the model being interfered by the noise contained in the OCR text, the OCR text with low predicted score and that in relatively small regions are discarded.

During training stage, the number of attention heads in BERT is 12, and the batch size is set to 16 for both VisualBERT and OCRBERT. Besides, Dropout is employed with the rate of 0.1, and the maximum length of the multimodal inputs for VisualBERT and OCRBERT is set to 512. All models are fine-tuned for 10 epochs by the Adam Optimizer [11] with an initial learning rate of 5e-5, and are implemented based on PyTorch with 4 Tesla P40 GPUs.

Model	BERT	VisualBERT	VisualBERT(*)	OCRBERT
Accuracy	85.66%	85.87%	86.03%	87.41%

Table 4. The performance of our proposed models and BERT. Note that VisualBERT(*) utilizes the 8x8x2048 visual features, while VisualBERT adopts the pooling feature with 2048 dimensions.

5.2 Experimental Results

We choose textual BERT as the baseline to evaluate the effectiveness of the multimodal input and our proposed models. Table 4 reports the accuracy results on MCIC dataset.

Both OCRBERT and VisualBERT outperform the BERT, which reveals the necessity of fusing multimodal information. The VisualBERT(*) that exploits local features shows superiority compared with VisualBERT that utilize global features. We argue that the reason for this is that VisualBERT(*) takes better advantage of visual features, and it selectively focuses on important regions of an image by deep interaction of multiple self-attention layers.

Interestingly, OCRBERT achieves the best result with 87.41% accuracy score. OCRBERT improves the accuracy by 1.75% compared with BERT, and 1.38% compared with VisualBERT(*). This indicates that the OCR text is more effective for multimodal conversational intent classification in E-commerce customer service.

5.3 Case Study

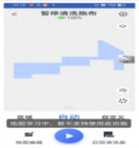
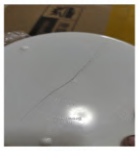
INPUT	OCR	BERT	OBERT	VBERT	GT
 暂停清洗拖布 100% 区域 自动 自定义 地图学习中，暂不支持使用此功能 这样了，我需要新建一个地图吗？ (It shows this message, should I create a new map?)	暂停清洗拖布 (stop clean the mop) 地图学习中，暂不支持使用此功能	其他 (Other)	售后商品使用问题 (use of after-sale goods)	物流无更新 (Logistics information not updated)	售后商品使用问题 (use of after-sale goods)
 收到这样了 (It is like this when I receive the express.)	None	货已收到 (goods received)	属性咨询 (Property consulting)	损坏 (Broken)	损坏 (Broken)

Fig. 5. The results of case study. OBERT, VBERT, and GT denote OCRBERT, VisualBERT, and the ground-truth label, respectively. We translate the most remarkable OCR text due to space limitation.

The upper case in Figure 5 shows a typical example in which the OCR text can boost the model performance of intent classification. The OCR text "*stop cleaning the map*" in the image, which reveals that the cleaning robot does not work, is beneficial for OCRBERT to identify the intent "*use of after-sale goods*". Conversely, without the assistance of OCR text, BERT and VisualBERT fail to capture the real intent.

The lower case shows that visual information also plays an essential role in the multimodal intent classification task. Since there is no OCR text information provided in the image, it is challenging to understand "*the goods is broken*" only from the text "*It is like this when I receive the express*". The visual information benefits VisualBERT to infer the right intent compared with OCRBERT and BERT.

5.4 Visualization

In order to study why OCRBERT shows more competitive performance, we visualize the self-attention weights of OCRBERT. We find that the OCR text is directly associated with the intent, which can help the model understand user utterances better. As shown in Figure 6, the attention weights imply that the OCR texts "*coupon*" and "*not available*" are more important to infer the intent "*coupon is not available*", which are implicit cues that are not contained in the input text.

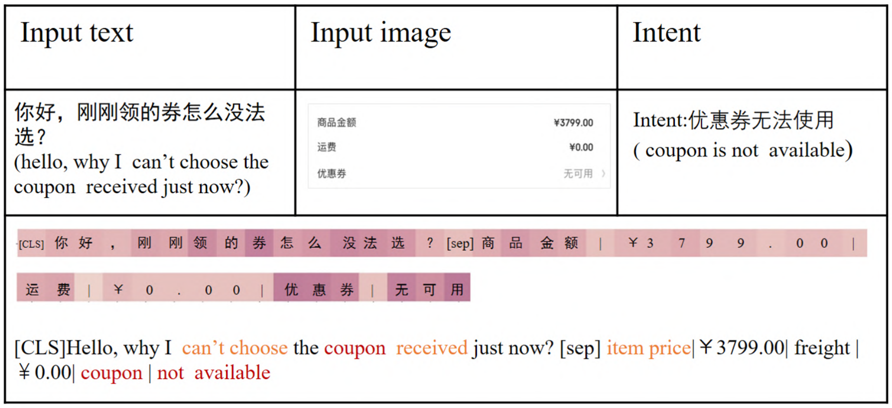

Input text
Input image
Intent
你好，刚刚领的券怎么没法选？
(hello, why I can’t choose the coupon received just now?)
商品金额 ¥3799.00
运费 ¥0.00
优惠券 无可用
Intent:优惠券无法使用
(coupon is not available)
[CLS] 你 好 ， 刚 刚 领 的 券 怎 么 没 法 选 ？ [sep] 商 品 金 额 | ￥ 3 7 9 9 . 0 0 |
运 费 | ￥ 0 . 0 0 | 优 惠 券 | 无 可 用
[CLS]Hello, why I can’t choose the coupon received just now? [sep] item price|￥3799.00| freight |
￥0.00| coupon | not available

Fig. 6. The visualization result of attention weights from OCRBERT to infer the intent “coupon is not available”. The darker color denotes that the character is more important for intent classification.

6 Conclusions

In this paper, we construct a large-scale dataset for multimodal conversational intent classification (MCIC), with different annotated labels, including image categories, OCR text, and intents. To promote relevant research, we design two BERT-based baselines to integrate multimodal input with deep interaction and verify the effectiveness of these models on our dataset. Users of this dataset are encouraged to explore more complicated architectures and learn a joint representation of dialogue text, images, and OCR text. And the dataset will be further enriched, including increasing the numbers of dialogues and and variety of image categories, in the future.

References

1. Liu, R.,Chen, M,.Liu, H.,Shen, L.,Song, Y.,He, X.: Enhancing Multi-turn Dialogue Modeling with Intent Information for E-Commerce Customer Service. In: Proceedings of NLPCC 2020 (2020)
2. Chen, M., Liu, R., Shen, L., Yuan, S., Zhou, J., Wu, Y., He, X., Zhou, B.: The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. In: Proceedings of LREC 2022 (2020)
3. Liao, L., Ma, Y., He, X., Hong, R., Chua, T.: Knowledge-aware multimodal dialogue systems. In: Proceedings of ACM MM 2018 (2018)
4. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of CVPR 2017 (2017)
5. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: Proceedings of ACL 2019 (2019)
6. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of ICCV 2015 (2015)

7. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. In: Proceedings of CVPR 2019 (2019)
8. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: an efficient and accurate scene text detector. In: Proceedings of CVPR 2017 (2017)
9. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of CVPR 2016 (2016)
10. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11),2298–2304 (2016)
11. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proceedings of ICLR 2015 (2015)
12. Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G., Vanderwende, L.: Image-grounded conversations: Multimodal context for natural question and response generation. In: Proceedings of IJCNLP 2017 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR 2016 (2016)
14. Shuster, K., Humeau, S., Bordes, A., Weston, J.: Image Chat: Engaging Grounded Conversations. In: Proceedings of ACL 2020 (2020)
15. Kottur, S., Moon, S., Geramifard, A., Damavandi, B.: SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In: Proceedings of EMNLP 2021 (2021)
16. Budzianowski, P., Wen, T., Tseng, B., Casanueva, I., Ultes, S., Ramadan, O., Gasic, M.: MultiWOZ–A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In: Proceedings of EMNLP 2018 (2018)
17. Li, X., Wang, Y., Sun, S., Panda, S., Liu, J., Gao, J.: Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. Journal: arXiv preprint arXiv:1807.11125 (2018)
18. Rastogi, A., Zang, X., Sunkara, S., Gupta, R., Khaitan, P.: Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05),8689–8696 (2020)
19. Zhu, Q., Huang, K., Zhang, Z., Zhu, X., Huang, M.: Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. TACL. **8**,281–295 (2020)
20. Joo, J., Li, W., Steen, F., Zhu, S.: Visual persuasion: Inferring communicative intents of images. In: Proceedings of CVPR 2014 (2014)
21. Vondrick, C., Oktay, D., Pirsiavash, H., Torralba, A.: Predicting motivations of actions by leveraging text. In: Proceedings of CVPR 2016 (2016)
22. Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., Divakaran, A.: Integrating text and image: Determining multimodal document intent in instagram posts. In: Proceedings of IJCNLP 2019 (2019)
23. Jia, M., Wu, Z., Reiter, A., Cardie, C., Belongie, S., Lim, S.: Intentonomy: a Dataset and Study towards Human Intent Understanding. In: Proceedings of CVPR 2021 (2021)
24. Saha, A., Khapra, M., Sankaranarayanan, K.: Towards building large scale multimodal domain-aware conversation systems. In: Proceedings of ACL 2018 (2018)
25. Farhadi, A., Hejrati, M., Sadeghi, M., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: Proceedings of ECCV 2010 (2010)
26. Zhao, N., Li, H., Wu, Y., He, X., Zhou, B.: The JDDC 2.0 Corpus: A Large-Scale Multimodal Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service. Journal: arXiv preprint arXiv:2109.12913 (2021)

27. Rahman, W., Hasan, M., Zadeh, A., Morency, L., Hoque, Mohammed E.: M-bert: Injecting multimodal information in the bert structure. Journal: arXiv preprint arXiv:1908.05787 (2019)
28. Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019 (2019)