SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

October 2022, Vol. 65 202104:1–202104:14
https://doi.org/10.1007/s11432-020-3307-3

Cognition-driven multimodal personality classification

Xiaoya GAO, Jingjing WANG*, Shoushan LI, Min ZHANG & Guodong ZHOU

School of Computer Science and Technology, Soochow University, Suzhou 215000*, China*

Received 29 October 2020/Revised 13 January 2021/Accepted 14 May 2021/Published online 27 September 2022

Abstract In this paper, we address a novel task, namely, cognition-driven multimodal personality classification (CMPC), aiming to infer personality traits (e.g., romantic, humorous, and gloomy) shown in real time by a human being from the perspective of cognitive psychology. Specifically, this task is motivated by a cognitive difference phenomenon that humans with different personality traits tend to give different personality-oriented textual descriptions when observing an image. In particular, to tackle the inherent noise challenges in this CMPC task, we propose a tailored reinforcement learning approach, namely, multi-agent SelectNet, aiming to integrate the opinion-word and image-region selection strategies to select informative opinion-word and image-region features for CMPC. To justify the effectiveness of our approach, we construct six kinds of multimodal personality classification datasets and conduct extensive experiments on the datasets. Experimental results demonstrate that our approach can significantly outperform other strong competitors, including the state-of-the-art unimodal and multimodal approaches.

Keywords cognitive psychology, personality classification, multimodal analysis, multi-agent reinforcement learning, agent-sharing mechanism

Citation Gao X Y, Wang J J, Li S S, et al. Cognition-driven multimodal personality classification. Sci China Inf Sci, 2022, 65(10): 202104, https://doi.org/10.1007/s11432-020-3307-3

1 Introduction

In the literature, existing studies on personality prediction [1–3] routinely focus on automatically predicting the personality of a human being based on the implicit and abstract Big Five [1] personality descriptors (i.e., conscientiousness, extraversion, agreeableness, neuroticism, and openness to experience) and treat personality prediction as a regression task of scoring in the range of $[0,1]$. In this study, we extend the research of personality prediction to an explicit and concrete classification scenario and propose a new personality prediction task, namely, cognition-driven multimodal personality classification (CMPC), which aims at predicting personality traits (e.g., romantic, gloomy, humorous, and aggressive) shown in real time by a human being from the perspective of cognitive psychology [4]. Compared to the traditional personality prediction task, this new task could play a crucial role in the development of emotional and empathetic virtual agents. For instance, the CMPC task potentially contributes to developing an interesting robot with a concrete humorous personality trait, just like the famous robot TARS in the popular movie Interstellar, but it is rather difficult for the traditional task to do this.

Specifically, our CMPC task is inspired by a cognitive difference phenomenon in terms of cognitive psychology [4] that humans with different personality traits tend to focus on different portions inside an image (i.e., different image regions) and give different personality-oriented textual expressions when observing an image. For instance, in Figure 1, a romantic person tends to focus on a whole flower covered by snow and gives a positive opinion “the flower looks like a dandelion.” By contrast, a gloomy person tends to focus on the withered flower and gives a negative opinion, i.e., “the flower has withered.” Inspired by the above cognitive difference phenomenon in human beings, one ideal solution for CMPC is to simultaneously select personality-relevant opinion words and image regions, discarding irrelevant or even noisy parts inside a text and an image, and then incorporate the selected opinion words and image

*Corresponding author (email: djingwang@suda.edu.cn)

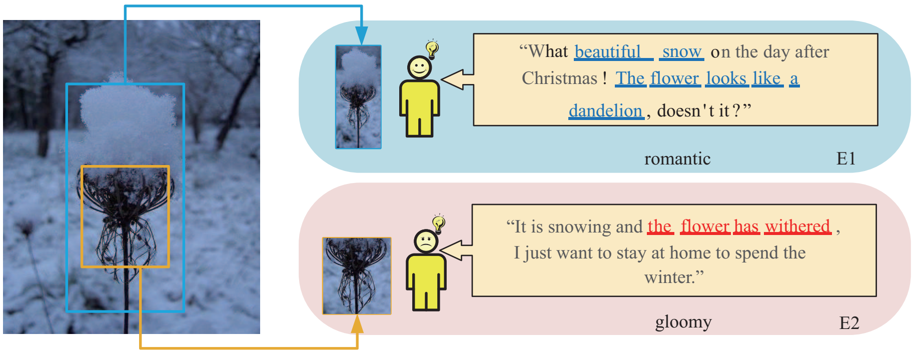

"What beautiful snow on the day after Christmas ! The flower looks like a dandelion , doesn't it ?"
romantic
E1
"It is snowing and the flower has withered , I just want to stay at home to spend the winter."
gloomy
E2

Figure 1 (Color online) Two descriptions, i.e., **E1** and **E2**, by two persons with different personalities (i.e., romantic and gloomy) after observing the same image.

regions for personality classification. In this solution, two major challenges exist, which are illustrated as follows.

On the one hand, selecting personality-relevant opinion words and discarding irrelevant or even noisy words is challenging. For instance, in **E1**, the opinion words "beautiful snow" and "the flower looks like a dandelion" contribute more than other words in implying the romantic personality and thus should be highlighted, whereas the phrase "doesn't it" without expressing opinions is more likely to be irrelevant or noisy and must be filtered because they make little contribution in implying the romantic personality. One straightforward approach to address this challenge is to employ the soft-attention mechanism as proposed by Wang et al. [5]. However, such a soft-attention mechanism has a shortcoming as the softmax function always assigns small but non-zero probabilities to irrelevant or noisy words, which may largely weaken attention weights given to the few truly discriminative opinion words. Alternatively, a better-behaved approach to CMPC should highlight discriminative opinion words and discard irrelevant and noisy ones for personality prediction during model training.

On the other hand, selecting personality-relevant image regions and discarding irrelevant or even noisy ones inside the image is challenging. For instance, in **E2**, the image region of the yellow box, i.e., the withered flower, apparently contributes more in implying the gloomy personality of the person. This finding is reasonable because gloomy persons are more likely to focus on negative portions inside an image. In this scenario, the image region of the blue box, i.e., the flower with the covering snow, might become the noise and should be discarded because snow is usually closely related to romance and may mislead the model into predicting the romantic personality. Therefore, a better-behaved approach to CMPC should discard those irrelevant and noisy image regions for personality prediction during model training.

In this paper, we propose a reinforcement learning-based approach, namely, multi-agent SelectNet (MASN), to simultaneously tackle the above two challenges. Specifically, we first leverage two pre-trained models, i.e., Bidirectional Encoder Representations from Transformers (BERT) [6] and residual network (ResNet) [7], to extract word and image-region features from a text and an image, respectively. Second, we propose two agents, i.e., opinion-word selector and image-region selector, to select personality-relevant opinion-word and image-region features for the CMPC task, respectively. Intuitively, the selected personality-relevant opinion-word features are beneficial for a good image-region selection and vice versa. Inspired by this intuition, we further propose an agent-sharing module to impose the two agents to collaborate with each other to boost the classification performance. Finally, we incorporate the representations of the selected opinion words and image regions for performing personality classification. The experimental evaluation demonstrates the impressive effectiveness of our MASN approach to CMPC over several strong baselines, including the strong unimodal approaches (e.g., BERT and ResNet) and state-of-the-art attention-based multimodal approaches.

2 Related work

Personality classification. In the literature, various studies have been devoted to personality classification in the natural language processing (NLP) field, with the difference in the types of features (i.e., text features, image features, or multimodal features). Particularly, Liu et al. [8] proposed a language-independent approach to extract character-level text features for predicting the personalities of texts in different languages. Yamada et al. [9] investigated the impact of user behavior features for personality classification on the basis of Twitter texts. Arnoux et al. [10] aimed to drastically reduce the tweet requirement for personality classification and proposed an approach that is applicable to most users on Twitter. Sun et al. [11] employed neural network models to extract text structures based on texts from online social networks for personality classification. Silva et al. [12] employed embedding-based approaches to classify personalities from the Facebook text. Pizzolli et al. [13] focused on extracting the features of utterances in theater scripts from literary texts. In addition, a few studies have used image information to perform personality classification. For instance, Liu et al. [14] analyzed how profile images vary with the personality of users. Ferwerda et al. [15] analyzed the effects of the visual and content features of Instagram pictures on personality prediction. Moubayed et al. [16] exploited facial features for personality classification. Xu et al. [17] employed a 2.5D hybrid personality computational model to gain a comprehensive understanding of one's personality traits.

Recently, with the development of multimodal analysis technologies, Kampman et al. [18] and Farnadi et al. [19] also adopted texts and images to predict personality traits. However, the following significant differences were noted. (1) We propose a new CMPC task, which is inspired by the cognitive difference of human beings. The images in the Big-Five datasets [1] adopted by the previous two papers only consist of facial features, which are not suitable for our new CMPC task that aims at leveraging the cognitive difference of humans to perform personality classification. (2) Our task aims to predict explicit personality traits (e.g., romantic and humorous) shown in real time by a human being instead of predicting the implicit and abstract Big-Five personality descriptors [1]. (3) Their approaches simply fuse the text and image features and fail to identify which word and image region contribute to the final personality prediction.

Different from all the above studies, we propose a new CMPC task inspired by the cognitive difference of human beings. To the best of our knowledge, this is the first attempt to address this new task. Moreover, this is the first study to perform personality classification from the perspective of cognitive psychology and consider the fine-grained opinion-word and image-region selections for personality classification.

Reinforcement learning. Recently, reinforcement learning approaches have been successfully applied to many NLP tasks. Specifically, Lei et al. [20] employed reinforcement learning to rationalize neural prediction for multi-aspect sentiment analysis. Guo [21] leveraged a Q-learning-based neural network to improve the performance of the text generation task. Huang et al. [22] proposed a hierarchically structured reinforcement learning approach to generate coherent multi-sentence stories for the visual storytelling task. Li et al. [23] employed deep reinforcement learning to improve the reward in the chatbot dialog task. Takanobu et al. [24] employed hierarchical reinforcement learning to model the relation extraction task. Wang et al. [25] presented a deep reinforcement learning-based model to incorporate the graph attention mechanism into knowledge graph reasoning. Zhang et al. [26] showed how to combine long short-term memory (LSTM) with policy gradient to obtain structured representations for the text classification task.

Recently, multi-agent reinforcement learning (MARL) has attracted increasing attention from NLPers. The key challenge in MARL is how to design an effective agent-sharing mechanism. In particular, Feng et al. [27] proposed a communication component between agents and aimed to solve a new multi-scenario ranking task. Gui et al. [28] also employed MARL with a critic-sharing strategy (i.e., centralized critic) between agents and aimed to filter tweets and images for the depression detection task, which is inspirational to our approach. Different from Gui et al. [28], we propose another new state-sharing strategy for agent sharing. To the best of our knowledge, this study is the first to integrate the state-sharing and reward-sharing strategies into the MARL architecture for better-behaved agent sharing in the CMPC task.

3 MASN model design

In this section, we formulate the CMPC task as an MARL problem [29] to address the two challenges mentioned in the Introduction section. Specifically, we propose a model, i.e., MASN, which integrates opinion-word selection and image-region selection strategies to tackle the data noise problem in the CMPC task. The basic motivation for using MARL lies in that although we do not have an explicit annotation of which opinion word and image region are discriminative for CMPC, we can measure their usefulness by the

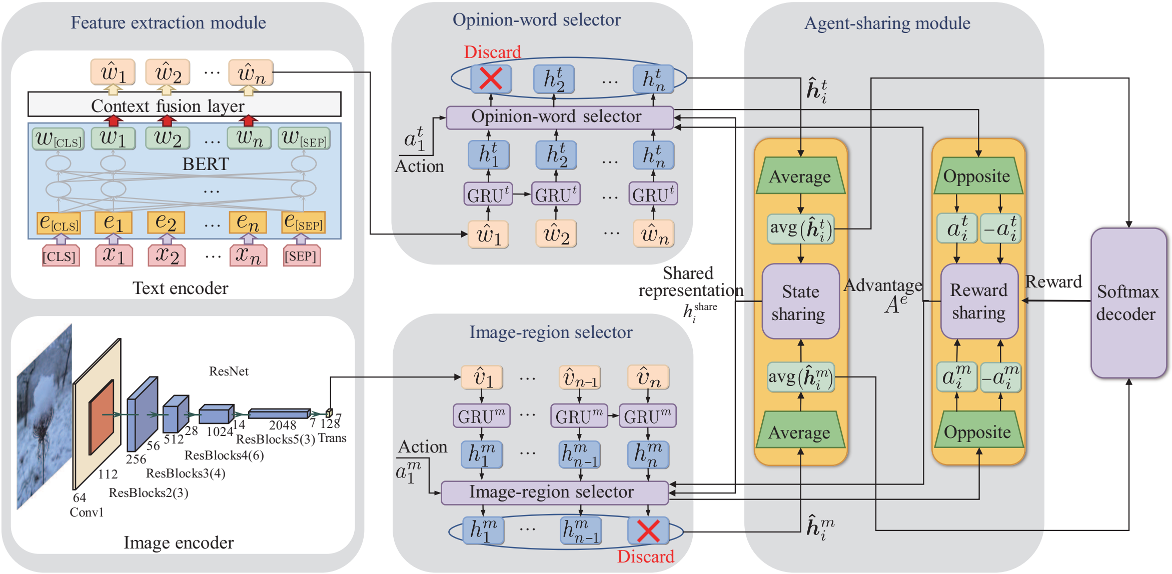

Feature extraction module
$\hat{w}_1$ $\hat{w}_2$ ··· $\hat{w}_n$
Context fusion layer
$w_{[CLS]}$ w_1 w_2 ··· w_n $w_{[SEP]}$
BERT
$e_{[CLS]}$ e_1 e_2 ··· e_n $e_{[SEP]}$
[CLS] x_1 x_2 ··· x_n [SEP]
Text encoder
ResNet
64
Conv1
112
ResBlocks2(3)
256
56
ResBlocks3(4)
512
28
1024
14
ResBlocks4(6)
2048
7
ResBlocks5(3)
7
128
Trans
Image encoder
Opinion-word selector
Discard
h_2^t ··· h_n^t
Opinion-word selector
a_1^t
Action
h_1^t h_2^t ··· h_n^t
GRU^t GRU^t ··· GRU^t
$\hat{w}_1$ $\hat{w}_2$ ··· $\hat{w}_n$
Shared representation
h_i^{share}
Image-region selector
$\hat{v}_1$ ··· $\hat{v}_{n\text{-}1}$ $\hat{v}_n$
GRU^m ··· GRU^m GRU^m
Action
a_1^m
h_1^m ··· $h_{n\text{-}1}^m$ h_n^m
Image-region selector
h_1^m ··· $h_{n\text{-}1}^m$
Discard
Agent-sharing module
$\hat{\boldsymbol{h}}_i^t$
Average
avg$(\hat{\boldsymbol{h}}_i^t)$
State sharing
avg$(\hat{\boldsymbol{h}}_i^m)$
Average
$\hat{\boldsymbol{h}}_i^m$
Advantage
A^e
Opposite
a_i^t $-a_i^t$
Reward sharing
a_i^m $-a_i^m$
Opposite
Reward
Softmax decoder

Figure 2 (Color online) Overall architecture of our proposed MASN approach to CMPC.

obtained reward signals from the target CMPC task. Concretely, we leverage two agents to perform word and image-region selections, respectively. The agents employ the trial-and-error-search strategy [30] to explore which word and image region are discriminative for the final personality classification and obtain the reward on the quality of personality inference from the personality classifier. Figure 2 shows the overall architecture of the MASN approach, which consists of five major components.

- **Feature extraction module** contains two large-scale pre-trained encoders for encoding texts and images, respectively. Given an input, i.e., a text and an image pair, we use BERT [6] to encode the text (a word sequence) into a vector sequence and ResNet [7] to encode image regions into a spatial convolutional neural network (CNN) feature sequence.
- **Opinion-word selector** is an agent that aims to select discriminative opinion words, discarding the noisy ones for the CMPC task.
- **Image-region selector** is the other agent that aims to select discriminative image regions, discarding the noisy ones for the CMPC task.
- **Agent-sharing module** consists of two agent-sharing mechanisms, i.e., state sharing and reward sharing. It is designed to impose the two agents to accomplish the same goal, i.e., improving the performance of the CMPC task.
- **Softmax decoder** is designed to perform personality classification and provide reward signals for measuring the usefulness of selected opinion words and image regions.

3.1 Feature extraction module

Word encoder. As a large-scale pre-trained text encoding model, BERT [6] can be fine-tuned to create state-of-the-art models for a range of NLP tasks, e.g., text classification and natural language inference. In this study, we use the BERT-base (uncased) model, followed by a context fusion layer as the word encoder. Following Devlin et al. [6], given an input word sequence $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, the input word sequence is first processed to WordPiece embeddings [31] with a 30000-token vocabulary and padded to 512 tokens. Then, the marked "[CLS]" is added as the first special BERT token, and then WordPiece embeddings are summed with the corresponding segment and positional embeddings. On the basis, a multilayered bidirectional transformer encoder [32] is used to map an input word sequence into a sequence of word embedding vectors $\boldsymbol{w} = [w_1, w_2, \ldots, w_n]$, where $w_i \in \mathbb{R}^{d'}$. Finally, to further enhance the context-aware representation for each word, we follow the work by Shen et al. [33] in using a context fusion layer for computing the final word vector $\hat{w}_i \in \mathbb{R}^d$ as follows:

$$\hat{w}_i = W_w \cdot (w_i \oplus \text{pool}(\boldsymbol{w}) \oplus (w_i \odot \text{pool}(\boldsymbol{w}))), \tag{1}$$

where $W_w \in \mathbb{R}^{d \times 3d'}$, symbol $\oplus$ denotes the vector concatenation operation, symbol pool denotes the mean-pooling operation along the sequential axis, and $\odot$ denotes the element-wise multiplication.

Image-region encoder. As a large-scale pre-trained image encoding model, ResNet [7] has shown state-of-the-art performance on various computer version tasks, e.g., image captioning [34] and image classification [7]. In this study, we use ResNet to extract spatial CNN features from each image. Specifically, following Lu et al. [34], we first employ ResNet to extract the spatial CNN features of each image, where each spatial CNN feature is taken as the vector encoding for each image region [35][1]. The spatial CNN feature sequence is denoted as $\boldsymbol{v} = [v_1, v_2, \ldots, v_n]$ where $v_i \in \mathbb{R}^{2048}$. Then, we feed the spatial CNN features into a fully connected layer to obtain the final image-region vector sequence $\hat{\boldsymbol{v}} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_n]$, which is computed as follows:

$$\hat{v}_i = \mathrm{ReLU}(W_v v_i), \tag{2}$$

where $\hat{v}_i \in \mathbb{R}^d$ and $W_v \in \mathbb{R}^{d\times 2048}$. ReLU is the activation function.

3.2 Opinion-word and image-region selectors

As introduced in the previous sections, we employ two agents, i.e., opinion selector and image-region selector, to perform opinion selection and image-region selection, respectively, to discard noisy information, and further improve the classification performance of the CMPC task. In the following sections, we will introduce the two selectors in detail. Specifically, we formalize an opinion selector and image-region selector as agents operating in a partially observable world and optimize their policies using deep reinforcement learning.

Opinion-word selector. This is an agent implemented with the reinforcement learning algorithm, i.e., actor-critic policy gradient [36]. In the following, for clarity, we use the mark t to represent all symbols in the opinion-word selector.

In brief, given a word sequence $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$, the goal of the opinion-word selector is to make the decision of whether to select the word x_i or not. The decision follows a stochastic policy π, which is defined as a conditional probability distribution $\pi(\boldsymbol{a}^t|\cdot)$ over the action sequence $\boldsymbol{a}^t = \{a_1^t, \ldots, a_n^t\}$, $a_i^t \in \{0, 1\}$. Here, $a_i^t = 1$ indicates that the word x_i is selected, and $a_i^t = 0$ indicates that the word x_i is discarded.

Specifically, as recurrent neural network (RNN)-based sequential models (e.g., LSTM and gated recurrent unit (GRU)) have been shown to effectively model the action sequence of reinforcement learning [26, 28], we leverage an action-aware GRU model, denoted as GRU^t, to encode the word vector $\hat{w}_i$ from BERT. In this action-aware GRU^t, the hidden state $h_i^t \in \mathbb{R}^d$ of the word x_i at the i-th time step is computed as follows:

$$h_i^t = \begin{cases} \mathrm{GRU}^t(h_{i-1}^t, \hat{w}_i), & a_i^t = 1, \\ h_{i-1}^t, & a_i^t = 0, \end{cases} \tag{3}$$

where $a_i^t = 1$ indicates that the word x_i is selected and the hidden state h_i^t is appended in a selected subset $\hat{\boldsymbol{h}}_i^t = \{h_i^t\}$ and $a_i^t = 0$ indicates that the word x_i is not selected and the hidden state h_i^t of the current time step i is directly copied from the previous time step $i-1$. In this way, the action-aware GRU^t can focus on personality-relevant opinion words and filter irrelevant parts to obtain a purified text representation.

In addition, for the opinion-word selector, the state and action, as the main components in the reinforcement learning algorithm of the actor-critic policy gradient, are designed as follows:

• **State.** The state should provide adequate information for deciding whether to select a word or not. Thus, at the i-th time step, the state $s_i^t \in \mathbb{R}^{4d}$ is defined as $s_i^t = h_{i-1}^t \oplus \hat{w}_i \oplus h_{i-1}^{\mathrm{share}}$, where $\oplus$ denotes the concatenate operation, and $h_{i-1}^{\mathrm{share}} \in \mathbb{R}^{2d}$ is the shared representation passed from the agent-sharing module (to be introduced later in the agent-sharing module), which is used to facilitate the cooperation of the two agents, i.e., word and image-region selectors, for achieving the same goal of boosting the CMPC performance.

• **Action.** The action $a_i^t \in \{0, 1\}$ is sampled with the conditional probability $\pi(a_i^t|s_i^t;\theta^t)$, which could be casted as a binary classification problem. Thus, we adopt a logistic function to define this conditional probability as follows:

$$a_i^t \sim \pi(a_i^t|s_i^t;\theta^t) = a_i^t\sigma(W^t s_i^t + b^t) + (1 - a_i^t)(1 - \sigma(W^t s_i^t + b^t)), \tag{4}$$

[1] In our experiments, we tried adopting objects detected by some promising object detection models (e.g., Faster-RCNN [35]) as image regions for performing the image-region selection. However, we found that this process would hurt the performance. Detailed comparison results and analysis are given in Table 3 and Section 5, respectively.

where $\theta^t = \{W^t \in \mathbb{R}^{1\times 4d}, b^t \in \mathbb{R}\}$ denotes the trainable parameters inside the opinion-word selector. The sign $\sim$ denotes the sampling operation, and σ is the sigmoid activation function.

Image-region selector. This is an agent that is also implemented with the reinforcement learning algorithm, i.e., actor-critic policy gradient [36]. In the following, for clarity, we use the mark m to represent all symbols in the image-region selector.

In brief, given an image-region vector sequence $\hat{\boldsymbol{v}} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_n]$, the goal of the image-region selector is to make the decision of whether to select the image region[2] $\hat{v}_i$ or not. The decision follows a stochastic policy π, which is defined as a conditional probability distribution $\pi(\boldsymbol{a}^m|\cdot)$ over the action sequence $\boldsymbol{a}^m = \{a_1^m, \ldots, a_n^m\}$, $a_i^m \in \{0,1\}$. Here, $a_i^m = 1$ indicates that the image region $\hat{v}_i$ is selected, and $a_i^m = 0$ indicates that the image region $\hat{v}_i$ is discarded.

Specifically, similar to the opinion-word selector, another action-aware GRU model, denoted as GRU^m, is leveraged to encode the image-region vector $\hat{v}_i$ from ResNet. In GRU^m, the hidden state $h_i^m \in \mathbb{R}^d$ of the image region $\hat{v}_i$ at the i-th time step is computed as follows:

$$h_i^m = \begin{cases} \mathrm{GRU}^m(h_{i-1}^m, \hat{v}_i), & a_i^m = 1, \\ h_{i-1}^m, & a_i^m = 0, \end{cases} \tag{5}$$

where $a_i^m = 1$ indicates that the image region $\hat{v}_i$ is selected and the hidden state h_i^m is appended in a selected subset $\hat{\boldsymbol{h}}_i^m = \{h_i^m\}$ and $a_i^m = 0$ indicates that the image region $\hat{v}_i$ is not selected and the hidden state h_i^m of the current time step i is directly copied from the previous time step $i-1$.

In addition, the state and action for the image-region selector are defined as follows:

- **State.** The state should provide adequate information for deciding whether to select an image region or not. Thus, at the i-th time step, the state $s_i^m \in \mathbb{R}^{4d}$ is defined as $s_i^m = h_{i-1}^m \oplus \hat{v}_i \oplus h_{i-1}^{\mathrm{share}}$, where $h_{i-1}^{\mathrm{share}} \in \mathbb{R}^{2d}$ is the shared representation passed from the agent-sharing module for facilitating the cooperation of the two agents.
- **Action.** The action $a_i^m \in \{0,1\}$ is sampled with the conditional probability $\pi(a_i^m|s_i^m;\theta^m)$. Similar to the opinion-word selector, this conditional probability is defined as follows:

$$a_i^m \sim \pi(a_i^m|s_i^m;\theta^m) = a_i^m \sigma(W^m s_i^m + b^m) + (1-a_i^m)(1-\sigma(W^m s_i^m + b^m)), \tag{6}$$

where $\theta^m = \{W^m \in \mathbb{R}^{1\times 4d}, b^m \in \mathbb{R}\}$ denotes the trainable parameters inside the image-region selector.

3.3 Agent-sharing module

This module is designed to enable the two agents (i.e., opinion-word selector and image-region selector) to collaborate to improve the classification performance of the CMPC task. Because we adopt the actor-critic policy gradient [36] to implement the two agents, we propose two agent-sharing mechanisms, i.e., state sharing and reward sharing, for guiding the agents to communicate with each other.

State sharing. The goal of the state-sharing mechanism is to share the state representation to make the two actors (i.e., the two selectors in our MASN approach) help each other. For instance, in Figure 1, if the opinion-word selector precisely selects the opinion words "the flower has withered," it can provide strong signals to enable the image-region selector to select the withered flower inside the image. Specifically, we enable the two selectors to communicate with each other by sharing a representation h_i^{share} in each state, and the shared representation $h_i^{\mathrm{share}} \in \mathbb{R}^{2d}$ is computed by concatenating the average representations of the selected opinion words and image regions at the i-th time step, i.e., $h_i^{\mathrm{share}} = \sigma(\mathrm{avg}(\hat{\boldsymbol{h}}_i^t) \oplus \mathrm{avg}(\hat{\boldsymbol{h}}_i^m))$. In the experiments, this state-sharing mechanism combined with the reward-sharing mechanism can further improve the performance of the CMPC task.

Reward sharing. The goal of the reward-sharing mechanism is to utilize one shared Q-value function for estimating the future overall rewards obtained by both agents to make the agents collaborate with each other for better performance. Specifically, we employ a critic network [36] to estimate the shared Q-value function $Q(\hat{\boldsymbol{h}}_i, a_i; \theta^c)$ for joint action $\boldsymbol{a}$ depending on the central state $\hat{\boldsymbol{h}}_i$, where $\hat{\boldsymbol{h}}_i = \hat{\boldsymbol{h}}_i^t \oplus \hat{\boldsymbol{h}}_i^m$ and $\boldsymbol{a} = \boldsymbol{a}^t \oplus \boldsymbol{a}^m$. This Q-value function can approximate the expected future total rewards based on the reward signals provided by the softmax decoder (to be introduced thereafter). However, the shared Q-value function typically generates only global rewards, which cannot effectively deduce each selector's own contribution and thus harms the performance, as proposed by Gui et al. [28]. To alleviate this

[2] For clarity, we directly adopt the vector symbol $\hat{v}_i$ to represent each image region instead of using a new symbol.

issue, following Gui et al. [28], we adopt an opposite action operation to compute different advantages to different selectors. Concretely, for each selector $e \in \{t, m\}$, the corresponding advantage $A^e(\hat{\boldsymbol{h}}_i, a_i)$ at the i-th time step is computed as follows:

$$A^e(\hat{\boldsymbol{h}}_i, a_i) = Q(\hat{\boldsymbol{h}}_i, (a_i^e, a_i^{-e})) - Q(\hat{\boldsymbol{h}}_i, (-a_i^e, a_i^{-e})), \tag{7}$$

where $-a_i^e$ denotes that the selector takes an opposite action and a_i^{-e} denotes that the action is sampled by the other selector.

3.4 Softmax decoder

The aim of using a softmax decoder lies twofold, i.e., providing reward and performing personality classification, which are formulated as follows.

Providing reward. During the training process, the softmax decoder is used to provide classification probabilities as the reward signals (to be presented in (9)) for guiding the agents to select discriminative opinion words and image regions.

Personality classification. During the classification process, the softmax decoder is used to perform the CMPC task. Specifically, we feed the joint representation of the selected opinion words and image regions at the final time step n, i.e., h_n^{share}, to a softmax layer to predict the probability of label $\hat{y} \in [0, 1]$, i.e., $p_\theta(\hat{y}|h_n^{\text{share}}) = \text{softmax}(W h_n^{\text{share}} + b)$. Here, $\theta = \{W, b\}$ is the trainable parameter. Then, the label with the highest probability stands for the predicted label for the sample. If no words and image regions are finally selected, the final representation for performing personality classification will be initialized with zero vectors, and MASN will predict a label with this representation.

3.5 Model training

The parameters in MASN contain three parts: (1) θ of GRU^t, GRU^m and softmax decoder; (2) θ^c of Q-value function $Q(\hat{\boldsymbol{h}}_i, a_i; \theta^c)$; and (3) θ^e (i.e., θ^t and θ^m) of opinion-word and image-region selectors. This study adopts backpropagation to optimize θ and actor-critic policy gradient [36] to optimize θ^c of the Q-value function (in the critic network) and θ^e of two selectors (in the actor network).

For θ, the objective of learning θ is to minimize the cross-entropy loss as follows:

$$J(\theta) = \mathbb{E}_{(\mathcal{S},y)\sim\mathcal{C}}[-\log p_\theta(y|h_n^{\text{share}})] + \frac{\delta}{2}\|\theta\|_2^2, \tag{8}$$

where $(\mathcal{S}, y)$ denotes a sample from the corpus $\mathcal{C}$, y is the ground-truth label, and δ is a L_2 regularization.

For θ^c of the Q-value function $Q(\hat{\boldsymbol{h}}_i, a_i; \theta^c)$, we optimize them by minimizing the mean squared error loss $J(\theta^c)$ by following Yeung et al. [37].

$$J(\theta^c) = \mathbb{E}\left[r_i + \gamma \max_{a_{i+1}} Q(\hat{\boldsymbol{h}}_{i+1}, a_{i+1}) - Q(\hat{\boldsymbol{h}}_i, a_i)\right], \tag{9}$$

where symbol $r_i = p_\theta(y|h_i^{\text{share}})$ is the reward provided by the softmax decoder. Concretely, we first feed the shared representation h_i^{share} of the selected opinion words and image regions at the i-th time step to the softmax decoder and then regard the classification probability as the reward. γ is the discount factor.

For θ^e, $e \in \{t, m\}$ of the two selectors, we optimize them with the policy gradient [30, 38]. The policy gradient w.r.t. θ^e is computed by differentiating the maximized expected reward $J(\theta^e)$ as follows:

$$\nabla_{\theta^e} J(\theta^e) = \mathbb{E}_{\pi^e}\left[\sum_{i=1}^{n} A^e(\hat{\boldsymbol{h}}_i, a_i) \nabla_{\theta^e} \log \pi^e(a_i^e|s_i^e)\right], \tag{10}$$

where the advantage $A^e(\hat{\boldsymbol{h}}_i, a_i)$ is computed according to (7).

Furthermore, during model training, θ^c and θ^e are not updated in the early stage, and thus the two selectors select all words and image regions in the text and image, respectively. When θ is optimized until the loss over development set does not significantly decrease, we then begin to simultaneously optimize θ, θ^c, and θ^e.

Table 1 Statistics of the corpus (including six different datasets) for CMPC, where "#pairs of text and image" denotes the number of "text + image" pairs and "#words/text" denotes the number of words (average per text)

Datasets	Five binary classification tasks					All traits
	Romantic	Calm	Scornful	Gloomy	Aggressive	
#pairs of text and image	1862	1878	1848	1908	1888	201795
#words/text	16.3	17.2	18.5	15.1	22.4	19.5

4 Experimentation

In this section, we first illustrate the experimental settings and then systematically evaluate the performance of our proposed MASN approach toward the new CMPC task.

4.1 Experimental settings

Data settings. We construct the corpus for CMPC from the PERSONALITY-CAPTIONS[3] dataset released by Shuster et al. [39]. The PERSONALITY-CAPTIONS dataset contains 201795 text-image pairs and 215 personality traits, where each pair is labeled with one trait. With this dataset, we conduct two kinds of experiments for a thorough comparison study. On the one hand, we directly adopt the above PERSONALITY-CAPTIONS dataset with all traits (i.e., all traits) in Table 1 for performing a multi-category classification task. On the other hand, we randomly select five personality traits and construct five sub-datasets for performing five binary classification tasks (i.e., romantic vs. non-romantic, calm vs. non-calm, scornful vs. non-scornful, gloomy vs. non-gloomy, and aggressive vs. non-aggressive). Specifically, for each binary classification task, we first select one random personality trait as the positive category (e.g., romantic) and regard the rest 214 personality traits as the negative category (e.g., non-romantic). Second, we pick all text-image pairs labeled with the positive category as the positive samples and randomly pick the text-image pairs with the remaining 214 personality traits as the negative samples. The selected positive and negative samples are balanced. Then, we adopt the same training/dev/test settings by following Shuster et al. [39]. The detailed statistics of the six datasets are shown in Table 1.

Implementation details. In all our experiments, we fine-tune BERT and update the word vectors for performing the CMPC task. The parameters of BERT-base (uncased) follow [6], and the parameters of ResNet follow [34]. For the image-region encoder of our MASN approach, the dimension of the ResNet outputs is $2048 \times 7 \times 7$. All the other hyperparameters are fine-tuned according to the development set. Specifically, we set the dimensions of the GRU hidden states to 256 and initialize all weights of the other layers using the Glorot uniform initializer [40]. In addition, we adopt the Adam optimizer [41] with an initial learning rate of 0.001 for cross-entropy and mean squared error training and adopt the Adam optimizer with a learning rate of 0.0001 for the training of all policy gradients. In addition, the discount factor γ in (9) is 0.85, the regularization weight of parameters is 10^{-5}, the dropout rate is 0.5, and the batch size is 32.

Evaluation metrics. The performance is evaluated with the accuracy (Acc.) and macro-F1 (F1). Here, F1 is the average F-score for all categories, where each F-score is calculated as the F-score $= \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$. Moreover, t-test is used to evaluate the significance of the performance difference between the two approaches by following Yang et al. [42].

Baselines. For comparison, we implement multiple state-of-the-art approaches to CMPC as baselines: (1) Char-RNN [8], a state-of-the-art textual personality classification approach, which employs an RNN to extract language-independent features (2) HS-LSTM [26], a text representation approach with reinforcement learning (3) VGG [43], which uses the VGG-19 model to extract image features for personality classification without considering textual information (4) ResNet [7], an image classification approach that uses the ResNet model to extract image features for personality classification (5) SIFT+SVM [44], which leverages the traditional feature engineering SIFT to extract image features and employs an SVM classifier for addressing personality classification (6) BERT [6], a state-of-the-art textual encoding model, which is used to perform personality classification by learning text representation (7) DAN [45], a multimodal approach to solve the visual question-answering problem with an

[3] This dataset released by Shuster et al. [39] is annotated by a large number of crowd workers. Concretely, given an image, each annotator is required to imitate a human with a specific personality trait (e.g., romantic) and then write the corresponding caption for the image. Shuster et al. [39] focused on generating personality-oriented image captioning with this dataset. Different from them, this study aims to use texts and images for performing the personality classification task.

attention mechanism for text and image denoising (8) FMN [18], a state-of-the-art multimodal approach for personality classification. In our implementation, we only use the text and image as the inputs. (9) CoATT [46], a multimodal approach to named entity recognition with an attention mechanism for text and image denoising (10) UDMF [19], a multimodal approach to predict the age, gender, and personality traits of social media users. In our implementation, we only use the network to predict personality traits. (11) COMMA [28], a state-of-the-art reinforcement learning-based approach to multimodal depression detection. In our implementation, we use this approach to perform opinion-word and image-region selections. (12) BERT+ResNet, a straightforward multimodal approach that simply concatenates textual and visual features extracted by BERT and ResNet for addressing personality classification (13) ViLT [47], a state-of-the-art cross-modal pre-trained model, which shows promising performance on various downstream cross-modal tasks and outperforms a dozen of strong cross-modal pre-trained models, such as ERNIE-VIL [48] and ImageBERT [49]. In our experiment, we employ ViLT[4)] as encoders to extract the textual and visual features of each input text-image pair and then perform personality classification. (14) MASN (text), our proposed approach that leverages only the opinion-word selector to perform opinion-word selection without considering image information (15) MASN (image), our proposed approach that leverages only the image-region selector to perform image-region selection without considering text information (16) MASN (random), a variant of our approach, which does not leverage BERT as the word encoder but randomly initializes word embeddings, following Gui et al. [28]. For a fair comparison, all the above multimodal approaches (except BERT+ResNet and ViLT) randomly initialize the word embeddings and adopt ResNet as the image encoder, like our MASN (random) approach.

4.2 Experimental results

Table 2 shows the performance[5)] of different approaches to CMPC. From Table 2, we can see that:

Unimodality performance. When only using text modality, (1) the reinforcement learning-based approach HS-LSTM performs better than Char-RNN. This demonstrates the effectiveness of using a proper reinforcement learning approach to learn the text representation for CMPC. (2) The large-scale pre-trained BERT approach performs better than HS-LSTM. This indicates the appropriateness of leveraging the large-scale pre-trained BERT as the word encoder for the CMPC task. (3) Our approach MASN (text) with opinion-word selection performs slightly better than BERT. This encourages us to perform an opinion-word selection for CMPC.

When only using image modality, the following aspects should be considered: (1) Neural network approaches could perform better than the traditional feature extraction approaches, i.e., SIFT+SVM. This confirms the powerful representation ability of neural networks for images. (2) The state-of-the-art image classification approaches VGG and RESNET perform better than a random performance. In particular, in the romantic dataset, RESNET achieves 60.4% in terms of accuracy, which is 10.4% better than that of the random. This encourages us to consider the image information for the personality classification task and indicates the appropriateness of using ResNet as the image-region encoder. Moreover, RESNET performs consistently better than VGG. This indicates that it is a better choice to leverage ResNet as the image-region encoder. (3) Our approach MASN (image) with image-region selection can perform consistently better than RESNET. This encourages us to perform an image-region selection for CMPC.

Multimodality performance. When using text and image modalities, DAN, UDMF, and COMMA perform better than most of the above unimodal approaches (except two BERT-based approaches, i.e., BERT and MASN (text)). This confirms the helpfulness of considering the image information in the CMPC task. In comparison, our approach MASN (random) significantly outperforms (p-value < 0.05) all the above multimodal approaches in terms of Acc. and F1. Among all the approaches, our approach MASN performs best and even significantly outperforms (p-value < 0.01) the strong baseline BERT in terms of Acc. and F1. These results encourage us to perform opinion-word and image-region selections for CMPC. In addition, the cross-modal pre-trained approach ViLT performs better than BERT+ResNet but still performs much worse than our MASN approach. These results show that the pre-trained cross-modal can obtain more high-quality textual and visual features by integrating multimodal alignment information and potentially contributes to our CMPC task. Accordingly, it inspires us to combine the cross-modal pre-trained model with the reinforcement learning mechanism to further boost the performance of our task in the future.

[4)] https://github.com/dandelin/ViLT.

[5)] For the detailed results w.r.t. the macro-precision and macro-recall, please see Tables A1 and A2 inside Appendix A.

Table 2 Performance comparison of various approaches to CMPC, where Unimodality denotes that the input of the approaches is unimodality (either text or image) and Multimodality denotes that the input consists of a text and an image. Top-n denotes the accuracy of the top n discovered personality traits with the highest probabilities

		Five binary classification tasks										All traits					
	Approaches	Romantic		Calm		Scornful		Gloomy		Aggressive		Top-1		Top-5		Top-10	
		F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
Unimodality (text)	Char-RNN [8]	68.7	68.9	61.9	62.2	60.1	60.2	58.6	58.7	60.9	61.0	5.1	6.5	18.9	20.0	30.2	30.1
	HS-LSTM [26]	69.5	69.8	62.7	63.3	61.7	62.3	60.9	61.4	62.6	62.9	5.6	6.9	20.0	21.0	31.7	31.6
	BERT [6]	79.2	79.3	72.9	73.3	75.4	75.5	71.8	72.0	76.7	76.8	7.2	9.5	27.5	29.2	41.0	41.3
	MASN (text)	80.2	80.2	73.9	74.4	76.5	76.5	73.0	73.2	78.3	78.4	9.4	10.9	28.9	29.6	41.6	41.6
Unimodality (image)	SIFT+SVM [44]	54.5	54.7	52.2	52.2	43.6	43.9	48.3	48.9	46.5	46.6	0.2	0.5	1.2	1.9	3.1	4.2
	VGG [43]	57.1	57.9	53.7	55.6	55.3	56.4	54.3	57.3	52.5	52.6	0.4	1.3	1.9	2.2	4.2	4.4
	ResNet [7]	59.2	60.4	52.4	56.7	56.9	57.1	56.4	59.2	55.0	55.4	0.7	1.9	2.0	2.3	4.5	4.5
	MASN (image)	61.3	62.5	57.2	58.9	58.6	59.2	58.5	61.0	55.2	58.6	1.1	2.3	2.5	2.6	4.8	5.1
Multimodality (text+image)	DAN [45]	73.5	73.6	64.4	64.5	63.7	63.8	62.3	62.8	65.0	66.4	6.8	8.5	23.0	24.3	35.2	35.5
	FMN [18]	71.7	71.8	62.8	63.3	63.5	65.3	54.8	56.1	59.8	60.3	4.8	6.2	19.2	20.1	30.8	30.6
	CoATT [46]	69.7	69.8	65.5	65.6	69.1	69.4	61.4	61.6	63.6	63.8	6.5	8.1	22.5	23.9	33.4	33.9
	UDMF [19]	71.7	71.7	65.6	65.6	66.3	66.3	61.0	61.0	64.4	64.7	7.2	8.4	23.5	24.5	35.0	35.2
	COMMA [28]	73.2	73.4	66.7	66.7	65.6	67.3	63.2	64.6	65.0	65.5	6.7	8.2	23.6	24.1	34.7	34.9
	BERT+ResNet	78.3	78.3	72.1	72.8	74.9	75.0	68.2	68.3	76.5	76.7	8.9	10.7	28.2	29.2	41.2	41.4
	ViLT [47]	79.2	79.2	74.4	74.4	78.6	78.6	71.9	72.0	79.3	79.3	9.4	11.1	28.9	29.8	41.4	41.6
	MASN (random)	76.4	76.4	68.9	68.9	73.4	73.5	69.4	69.5	71.3	71.6	7.5	9.0	24.4	25.4	35.9	36.2
	MASN	**83.0**	**83.0**	**76.3**	**76.7**	**79.6**	**79.6**	**74.3**	**74.4**	**81.0**	**81.2**	**10.7**	**12.1**	**30.4**	**31.3**	**43.1**	**44.4**

Table 3 Ablation study of our proposed MASN approach on different personality datasets

Approaches	Five binary classification tasks									
	Romantic		Calm		Scornful		Gloomy		Aggressive	
	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
MASN	**83.0**	**83.0**	**76.3**	**76.7**	**79.6**	**79.6**	**74.3**	**74.4**	**81.0**	**81.2**
w/o state-sharing	74.4	74.5	68.6	70.0	74.4	74.5	71.7	72.0	75.8	75.9
w/o reward-sharing	78.2	78.3	71.2	72.2	75.6	75.7	72.5	72.6	77.6	77.6
w/o opinion-word selection	79.1	80.1	70.2	71.1	76.2	76.5	70.5	70.7	79.2	79.3
w/o image-region selection	80.2	80.2	72.5	73.3	75.5	75.5	69.3	69.5	78.4	78.5
Using objects as image-regions	81.1	81.1	72.7	73.3	76.4	76.5	70.0	70.3	79.3	79.3
Using soft-attention as selectors	80.0	80.2	70.2	71.1	72.3	72.4	71.9	72.0	80.1	80.2

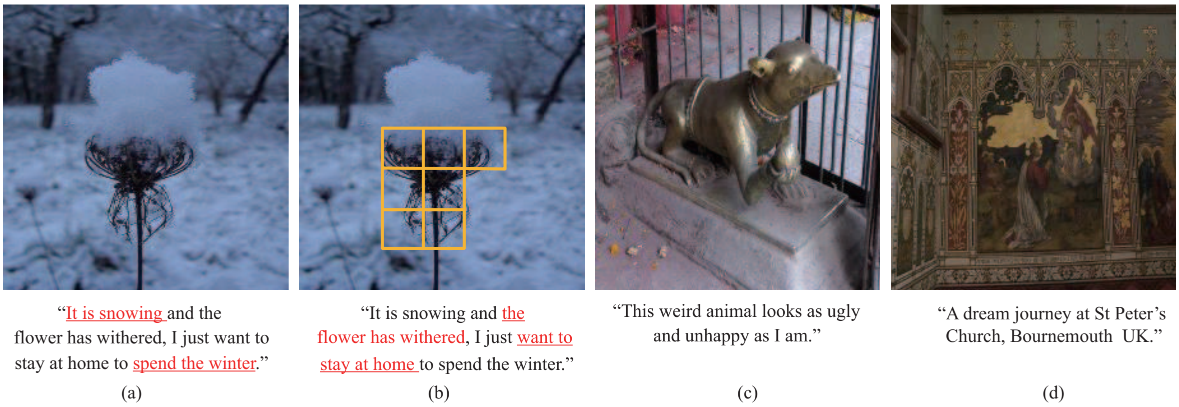
	All traits					
	Top-1		Top-5		Top-10	
	F1	Acc.	F1	Acc.	F1	Acc.
MASN	**10.7**	**12.1**	**30.4**	**31.3**	**43.1**	**44.4**
w/o state-sharing	10.0	11.5	29.8	30.7	42.5	42.8
w/o reward-sharing	9.8	11.3	29.2	30.0	42.3	42.4
w/o opinion-word selection	10.6	12.0	30.1	31.1	**43.1**	43.4
w/o image-region selection	10.2	11.8	29.7	30.7	42.2	42.6
Using objects as image-regions	10.0	11.4	29.4	30.5	42.2	42.7
Using soft-attention as selectors	9.7	11.2	28.6	29.7	41.8	42.1

Contribution of each key component. We conducted an ablation study to evaluate the contribution of each key component in our proposed MASN approach. From Table 3, we present the following observations: (1) Incorporating the state-sharing mechanism into the agent-sharing module can improve Acc. by an average of 3.85% in six different datasets. This justifies the effectiveness of our proposed state-sharing mechanism for the CMPC task. (2) Incorporating the reward-sharing mechanism into the agent-sharing module can improve Acc. by an average of 2.83%. This confirms the effectiveness of the reward-sharing mechanism for the CMPC task. (3) Performing opinion-word selection on the text and discarding noisy ones can improve Acc. by an average of 2.31%. This demonstrates the helpfulness of performing opinion-word selection for CMPC, encouraging us to perform opinion-word selection for CMPC. (4) Performing image-region selection on the image and discarding noisy ones can improve Acc. by an average of 2.58%. This further encourages us to perform image-region selection for the CMPC task. (5) In our MASN approach, leveraging objects detected by Faster-RCNN[6] as image regions for performing selection will reduce Acc. by an average of 2.20% compared with directly treating the spatial CNN features extracted by ResNet as image regions. This is mainly because, although the promising object detection models (e.g., Faster-RCNN) have achieved impressive progress in some domain-specific scenarios, they are always domain sensitive and might perform rather poorly on another domain, as proposed by Zheng et al. [50], resulting in the error propagation issue. For example, in our task, the statistical analysis shows that 15.1% of images do not have any objects detected by Faster-RCNN, which cannot suit well with our motivation of performing fine-grained image-region selection and thus hurt the classification performance. (6) In our MASN approach, using the soft-attention mechanism instead of the reinforcement learning mechanism as opinion-word and image-region selectors will reduce Acc. by an average of 2.98%. This again justifies the effectiveness of performing opinion-word and image-region selections for CMPC.

5 Analysis and discussion

Qualitative analysis. To get a better understanding of our MASN approach and validate that our approach can select informative opinion words and image regions for the CMPC task, we provide a qualitative analysis of our MASN approach on the development set of the gloomy dataset. Specifically, in Figure 3, we visualize the selected opinion words and image regions using the two approaches, i.e., MASN (text) and MASN. Specifically, we can observe the following from this figure: (1) For the gloomy sample in Figure 3(a), MASN (text) fails to select the opinion words "the flower has withered" and predicts the

[6] https://github.com/jwyang/faster-rcnn.pytorch.



"It is snowing and the flower has withered, I just want to stay at home to spend the winter."

(a)

"It is snowing and the flower has withered, I just want to stay at home to spend the winter."

(b)

"This weird animal looks as ugly and unhappy as I am."

(c)

"A dream journey at St Peter's Church, Bournemouth UK."

(d)

Figure 3 (Color online) (a) MASN (text); (b) MASN; (c) true label: miserable, prediction: gloomy; (d) true label: romantic; prediction: non-romantic. (a) and (b) are the visualization of the selected opinion words and image regions using the two approaches, where the red color denotes the word that has been selected, the yellow box denotes the image region that has been selected, and the other colors denote the deletion operation. (c) and (d) are the error cases with the ground-truth personality traits and corresponding predicted labels.

sample as "romantic" according to selected words; e.g., "It is snowing" and "spend the winter." This is due to the fact that snow is usually associated with romance and beauty and then misleads the model into giving a wrong label. (2) In Figure 3(b), when incorporating the image information, MASN can not only effectively select the opinion words "the flower has withered" but also precisely capture the discriminative image-region, i.e., the withered flower, and thus give the correct gloomy prediction.

Remaining challenges. Although the experimental results are impressive, some challenges were not addressed by our MASN approach, which can be considered to potentially boost the performance of CMPC in the future. To investigate the shortcomings of our MASN approach, we randomly select and analyze 100 error cases in the experiments, which can be roughly categorized into two main types. (1) The first type of error is due to the fuzzy boundary among personality labels, such as miserable and gloomy, which is the main reason why the top-1 performance on the All Traits dataset is relatively low. For example, in Figure 3(c), our MASN approach predicts a non-gloomy sample as gloomy, but the true label of this sample is miserable. Accordingly, we can incorporate the correlation information among similar traits to improve classification performance in future works. An easy solution to remedy the above issue is to normalize the dataset by grouping some similar traits. For example, the traits miserable and gloomy can be treated as the same trait. The corresponding results on this newly reconstructed dataset are shown in Appendix B. (2) The second type of error is due to the requirement of reasoning with external knowledge. For example, in Figure 3(d), the MASN approach predicts the sample as non-romantic, but the true label is romantic. Accordingly, we utilize the external knowledge base (e.g., ConceptNet) for capturing romantic elements, such as the famous painting inside the image and the famous building "St Peter's Church" in the text in future works.

6 Conclusions

In this study, we address a new CMPC task, aiming at leveraging the cognitive difference phenomenon of human beings to predict their personality traits shown real time. In particular, we propose an MASN approach to address this CMPC task. The main idea of our proposed approach is to incorporate the knowledge of opinion words and fine-grained image regions for the CMPC task. Specifically, our approach takes advantage of two opinion-word and image-region selectors to perform opinion-word selection and image-region selection for the CMPC task. Detailed experiments justify that opinion-word and image-region selections are effective, and the proposed MASN approach significantly outperforms several strong baselines.

In our future works, we would like to solve other challenges in CMPC, such as incorporating the external ConceptNet knowledge base to perform reasoning and address various issues exposed by error analysis and combining the large-scale cross-modal pre-training model (e.g., ERNIE-VIL [48] and ViLT [47]) with the reinforcement learning mechanism to further boost the performance. Furthermore, we would like to apply our MASN approach to other psychological analysis tasks, such as multimodal emotion analysis and multimodal anxiety detection.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62006166, 62076175, 62076176), China Postdoctoral Science Foundation (Grant No. 2019M661930), and Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). We thank our anonymous reviewers for their helpful comments.

Supporting information Appendixes A and B. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Goldberg L R. An alternative "description of personality": the big-five factor structure. J Personal Social Psychol, 1990, 59: 1216–1229

2 Ríssola E A, Bahrainian S A, Crestani F. Personality recognition in conversations using capsule neural networks. In: Proceedings of Web Intelligence, Thessaloniki, 2019. 180–187

3 Li Y N, Wan J, Miao Q G, et al. CR-Net: a deep classification-regression network for multimodal apparent personality analysis. Int J Comput Vis, 2020, 128: 2763–2780

4 Carver C S, Scheier M F. Control theory: a useful conceptual framework for personality-social, clinical, and health psychology. Psychol Bull, 1982, 92: 111–135

5 Wang J J, Li J, Li S S, et al. Aspect sentiment classification with both word-level and clause-level attention networks. In: Proceedings of International Joint Conference on Artificial Intelligence, Stockholm, 2018. 4439–4445

6 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirec-tional transformers for language understanding. In: Proceedings of North American Chapter of the Association for Computational Linguistics, Minneapolis, 2019. 4171–4186

7 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778

8 Liu F, Nowson S, Perez J. A language-independent and compositional model for personality trait recognition from short texts. In: Proceedings of European Chapter of the Association for Computational Linguistics, Valencia, 2017. 754–764

9 Yamada K, Sasano R, Takeda K. Incorporating textual information on user behavior for personality prediction. In: Proceedings of Association for Computational Linguistics, Florence, 2019. 177–182

10 Arnoux P H, Xu A B, Boyette N, et al. 25 Tweets to know you: a new model to predict personality with social media. In: Proceedings of International Conference on Web and Social Media, Montréal, 2017. 472–475

11 Sun X G, Liu B, Cao J X, et al. Who am I? Personality detection based on deep learning for texts. In: Proceedings of International Conference on Communications, Kansas City, 2018. 1–6

12 da Silva B B C, Paraboni I. Personality recognition from facebook text. In: Proceedings of the Portuguese Language, Canela, 2018. 107–114

13 Pizzolli D, Strapparava C. Personality traits recognition in literary texts. In: Proceedings of Storytelling Workshop, 2019. 107–111

14 Liu L Q, Preotiuc-Pietro D, Samani Z R, et al. Analyzing personality through social media profile picture choice. In: Proceedings of International Conference on Web and Social Media, Cologne, 2016. 211–220

15 Ferwerda B, Tkalcic M. Predicting users' personality from instagram pictures: using visual and/or content features? In: Proceedings of User Modeling, Adaptation and Personalization, Singapore, 2018. 157–161

16 Moubayed N A, Vazquez-Alvarez Y, McKay A, et al. Face-based automatic personality perception. In: Proceedings of ACM-MM, Orlando, 2014. 1153–1156

17 Xu J, Tian W J, Fan Y Y, et al. Personality trait prediction based on 2.5D face feature model. In: Proceedings of Cloud Computing and Security, Haikou, 2018. 611–623

18 Kampman O, Barezi E J, Bertero D, et al. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In: Proceedings of Association for Computational Linguistics, Melbourne, 2018. 606–611

19 Farnadi G, Tang J, de Cock M, et al. User profiling through deep multimodal fusion. In: Proceedings of Web Search and Data Mining, Marina Del Rey, 2018. 171–179

20 Lei T, Barzilay R, Jaakkola T S. Rationalizing neural predictions. In: Proceedings of Empirical Methods in Natural Language Processing, Austin, 2016. 107–117

21 Guo H Y. Generating text with deep reinforcement learning. 2015. ArXiv:1510.09202

22 Huang Q Y, Gan Z, Celikyilmaz A, et al. Hierarchically structured reinforcement learning for topically coherent visual story generation. In: Proceedings of Association for the Advance of Artificial Intelligence, Honolulu, 2019. 8465–8472

23 Li J W, Monroe W, Ritter A, et al. Deep reinforcement learning for dialogue generation. In: Proceedings of Empirical Methods in Natural Language Processing, Austin, 2016. 1192–1202

24 Takanobu R, Zhang T Y, Liu J X, et al. A hierarchical framework for relation extraction with reinforcement learning. In: Proceedings of Association for the Advance of Artificial Intelligence, Honolulu, 2019. 7072–7079

25 Wang H, Li S Y, Pan R, et al. Incorporating graph attention mechanism into knowledge graph reasoning based on deep reinforcement learning. In: Proceedings of Empirical Methods in Natural Language Processing, Hong Kong, 2019. 2623–2631

26 Zhang T Y, Huang M L, Zhao L. Learning structured representation for text classification via reinforcement learning. In: Proceedings of Association for the Advance of Artificial Intelligence, New Orleans, 2018. 6053–6060

27 Feng J, Li H, Huang M L, et al. Learning to collaborate: multi-scenario ranking via multi-agent reinforcement learning. In: Proceedings of World Wide Web, Lyon, 2018. 1939–1948

28 Gui T, Zhu L, Zhang Q, et al. Cooperative multimodal approach to depression detection in twitter. In: Proceedings of Association for the Advance of Artificial Intelligence, Honolulu, 2019. 110–117

29 Littman M L. Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of International Conference of Machine Learning, New Brunswick, 1994. 157–163

30 Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation. In: Proceedings of Neural Information Processing Systems, Denver, 1999. 1057–1063

31 Wu Y H, Schuster M, Chen Z F, et al. Google's neural machine translation system: bridging the gap between human and machine translation. 2016. ArXiv:1609.08144

32 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Neural Information Processing Systems, Long Beach, 2017. 6000–6010

33 Shen T, Zhou T Y, Long G D, et al. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. In: Proceedings of International Joint Conference on Artificial Intelligence, Stockholm, 2018. 4345–4352

34 Lu J S, Xiong C M, Parikh D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of Computer Vision and Pattern Recognition, Honolulu, 2017. 3242–3250

35 Ren S Q, He K M, Girshick R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of Neural Information Processing Systems, Montreal, 2015. 91–99

36 Sutton R S, Barto A G. Reinforcement learning: an introduction. IEEE Trans Neural Netw, 1998, 9: 1054–1054

37 Yeung S, Ramanathan V, Russakovsky O, et al. Learning to learn from noisy web videos. In: Proceedings of Computer Vision and Pattern Recognition, Honolulu, 2017. 7455–7463

38 Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn, 1992, 8: 229–256

39 Shuster K, Humeau S, Hu H X, et al. Engaging image captioning via personality. In: Proceedings of Computer Vision and Pattern Recognition, Long Beach, 2019. 12516–12526

40 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of Artificial Intelligence and Statistics, Chia Laguna Resort, 2010. 249–256

41 Kingma D P, Ba J. ADAM: a method for stochastic optimization. In: Proceedings of International Conference on Learning Representations, San Diego, 2015

42 Yang Y M, Liu X. A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, 1999. 42–49

43 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations, San Diego, 2015

44 Olgun M, Onarcan A O, Özkan K, et al. Wheat grain classification by using dense SIFT features with SVM classifier. Comput Electron Agr, 2016, 122: 185–190

45 Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching. In: Proceedings of Computer Vision and Pattern Recognition, Honolulu, 2017. 2156–2164

46 Zhang Q, Fu J L, Liu X Y, et al. Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of Association for the Advance of Artificial Intelligence, New Orleans, 2018. 5674–5681

47 Kim W, Son B, Kim I. ViLT: vision-and-Language transformer without convolution or region supervision. 2021. ArXiv:2102.03334

48 Yu F, Tang J J, Yin W C, et al. ERNIE-ViL: knowledge enhanced vision-language representations through scene graph. 2020. ArXiv:2006.16934

49 Qi D, Su L, Song J, et al. ImageBERT: cross-modal pre-training with large-scale weak-supervised image-text data. 2020. ArXiv:2001.07966

50 Zheng Y T, Huang D, Liu S T, et al. Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of Computer Vision and Pattern Recognition, Seattle, 2020. 13763–13772