2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428322

# COGNITION-DRIVEN REAL-TIME PERSONALITY DETECTION VIA LANGUAGE-GUIDED CONTRASTIVE VISUAL ATTENTION

*Xiaoya Gao, Jingjing Wang**, *Shoushan Li, Guodong Zhou*

School of Computer Science and Technology, Soochow University, China
Email: xygao1221@stu.suda.edu.cn, {djingwang, lishoushan, gdzhou}@suda.edu.cn

## ABSTRACT

In this paper, we address a novel Cognition-driven Real-time Personality Detection (C-RPD) task, aiming to predict the personality trait (e.g., *romantic* and *humorous*) real-time shown by a human being from the perspective of cognitive psychology. Specifically, this task is motivated by a cognition difference phenomenon that humans with different personality traits tend to focus on different portions of the image and then give different personality-oriented language descriptions when observing an image. On this basis, we propose a new Language-guided Contrastive Visual Attention (L-CVA) approach to capture the above cognition difference information for addressing the C-RPD task. Experimental results on a real-world multimodal personality corpus verify the great advantage of our L-CVA approach to C-RPD over the state-of-the-art baselines. This justifies the importance of the cognition difference information to C-RPD and the effectiveness of our approach in capturing such information.

***Index Terms*****—** Cognitive Psychology, Personality Detection, Multimodal Analysis, Contrastive Visual Attention

## 1. INTRODUCTION

In the literature, dominant paradigms commonly cast personality prediction [1] as a regression task of scoring in the range of [0, 1] for each static and abstract Big-Five [1] personality descriptor (e.g., Conscientiousness and Extraversion). In this study, we extend this personality prediction research to an explicit and concrete classification scenario, and propose a new Cognition-driven Real-time Personality Detection (C-RPD) task for predicting the dynamic and concrete personality trait (e.g., *romantic* and *humorous*) real-time shown by a human being from the perspective of cognitive psychology [2]. Beyond the traditional personality prediction, the C-RPD task could play a more powerful role in the construction of personalized and empathetic robots. For instance, C-RPD potentially contributes to developing an intelligent robot real-time showing a concrete *humorous* personality trait, e.g., the robot *TARS* in the popular movie *Interstellar*, while this is rather difficult for the traditional regression-based task.

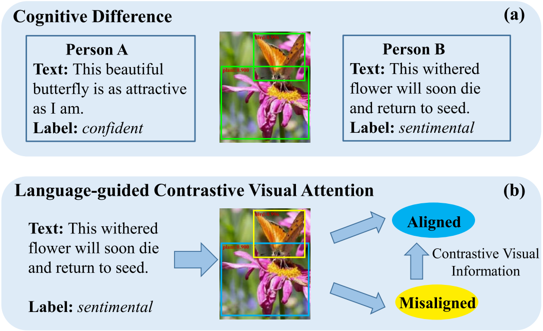
Cognitive Difference
(a)
Person A
Text: This beautiful butterfly is as attractive as I am.
Label: confident
Person B
Text: This withered flower will soon die and return to seed.
Label: sentimental
Language-guided Contrastive Visual Attention
(b)
Text: This withered flower will soon die and return to seed.
Label: sentimental
Aligned
Contrastive Visual Information
Misaligned

**Fig. 1**: An example for illustrating the cognitive difference phenomenon and the motivated C-RPD task.

In particular, our C-RPD task is motivated by a cognitive difference phenomenon in terms of cognitive psychology [2] that humans with different personality traits tend to pay attention to different portions inside the image (e.g., different image objects) and then give different personality-oriented language descriptions when observing an image. Take Fig. 1(a) as an example, a ***confident*** person **A** focuses on the beautiful butterfly resting on the flower and then expresses a positive language description "*...as attractive as I am.*", while a ***sentimental*** person **B** focuses on the little withered flower and then expresses a negative language description "*The withered flower...*". On this basis, a simple approach for C-RPD is to mine the language clues for modeling the cognitive difference information. Beyond the language information, we argue that the contrastive visual information (i.e., the cognitive behavior of focusing on some specific image-regions rather than the rest ones) is another crucial cognition difference information and could potentially assist the personality detection. In this way, two key challenges exist which are illustrated as follows.

For one thing, how to effectively capture the image-region information focused by cognition-driven humans is challenging. From the cognitive linguistic perspective [3], language is exactly the embodiment of cognition. Thus, an ideal approach is to leverage a language-guided visual attention mechanism to align the image-regions with the language, for capturing the

*Corresponding author. This work was supported by three NSFC grants, i.e., No.62006166, No.62076175 and No.62076176. This work was also supported by a project funded by China Postdoctoral Science Foundation No.2019M661930 and a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

attention tendency of humans with various kinds of personality traits. For another, how to effectively model the contrastive visual information is challenging. Prior visual attention models [4] are always specially-designed for capturing the visual information aligned with the language, which does not suit well our motivation of modeling the contrastive visual information for C-RPD. Still take Fig. 1 as an example, person **B** focuses on the negative blue-box image-region "*the withered flower*" rather than the positive yellow-box image-region "*the beautiful butterfly*", according to his/her language. Apparently, this contrastive visual information can powerfully contribute to understanding the personality trait of person **B** is *sentimental* instead of other *non-sentimental* traits, since *sentimental* persons are more likely to focus on negative portions within the image instead of the positive portions, e.g., "*beautiful butterfly*" usually associated with the romantic things. Thus, a well-behaved visual attention mechanism should not only highlight the aligned image-regions but also ignore the misaligned image-regions as much as possible for powerfully capturing this contrastive visual information.

In this paper, we propose a novel contrastive visual attention mechanism, namely Language-guided Contrastive Visual Attention (L-CVA), to tackle the above two challenges simultaneously. Specifically, we first leverage two pre-trained models, i.e., BERT [5] and Faster R-CNN [6], to encode the language and extract object features inside the image respectively. Second, we design a contrastive visual attention module to extract both the aligned and misaligned objects towards the language and compute the aligned and misaligned visual representation respectively. Third and finally, the aligned visual representation is used to maximize the softmax probabilities of the ground-truth personality labels, while the misaligned one is used to minimize the softmax probabilities of ground-truth personality labels, for highlighting the contrastive visual information. In addition, considering the massive personality trait categories, we further propose a ranking-aware loss to enlarge the predicted score variance of candidate labels during model training for boosting the classification performance. Experimentation demonstrates the great effectiveness of our L-CVA approach to the C-RPD task over a bunch of state-of-the-art unimodal and multimodal baselines.

## 2. RELATED WORK

**Personality Detection.** As an interdisciplinary research task, personality detection has been drawing ever-more attention in the multimedia analysis and artificial intelligence communities with a focus on extracting various types of features from either the text modality (e.g. social network texts [7]) or the image modality (e.g. the profile images [8] ). Compared with these studies on the single modality, the studies on multi-modality (e.g., both the text and image modalities) like [9] are much less and limited to treat personality prediction as a regression task based on the implicit Big-Five [1] personality descriptors. Unlike all the above studies, we propose a new C-RPD task to detect the explicit and concrete personality traits from the cognitive perspective. To our best knowledge, this is the first attempt to address this new task.

**Visual Attention.** Recently, visual attention mechanisms have been successfully leveraged to many CV and NLP tasks, e.g., named entity recognition [10], for aligning the fine-grained image-regions with the language. These studies routinely segment the whole image into multiple equal-sized regions averagely and then employ visual attention to align these regions with the language. However, it is rather difficult for such a vanilla visual attention mechanism to locate the completely semantic unit, e.g., "the *withered flower*" in Fig. 1. Inspired by recent visual grounding task [11], we aim to leverage the detected objects within the image as the basic image units for performing the alignment between objects and the language. Besides, close to our work, [12] and [13] employ the contrastive attention mechanism for the person re-identification and text summarization tasks, but they only focus on the single text or image modality. Different from the above studies, we focus on exploring the contrastive visual attention mechanism in a multi-modal scenario. To our best knowledge, this is the first attempt to incorporate both the contrastive attention and visual attention simultaneously.

## 3. LANGUAGE-GUIDED CONTRASTIVE VISUAL ATTENTION DESIGN

Given an input text-image pair $(T_i, I_i)$ generated by a human being, our C-RPD task aims at leveraging this pair to detect the real-time personality trait of this human. On this basis, we propose a language-guided contrastive visual attention (L-CVA) approach to C-RPD. Fig. 2 shows the framework of L-CVA, which consists of three major components: **1)** a feature extraction module, which uses a BERT [5] model to encode the text $T_i$ (a word sequence) into a vector sequence and a Faster R-CNN [6] model to encode the detected objects inside the image $I_i$ into a spatial CNN feature sequence, when given the input text-image pair $(T_i, I_i)$; **2)** a contrastive visual attention module, which accommodates two types of attention models, i.e., basic visual attention and opponent visual attention, to extract the visual representation of both the aligned and misaligned objects towards the language; **3)** a contrast-aware decoder, which integrates the contrast-aware losses and a ranking-aware loss for better personality detection.

### 3.1. Feature Extraction Module

**Language Features.** Given an input text $T_i$, we employ the promising BERT-base (uncased) model to encode the text (https://github.com/google-research/bert). Following Devlin et al. [5], given a sentence S, the input word sequence $\mathrm{S} = \{\mathrm{s}_1, \mathrm{s}_2, ..., \mathrm{s}_\mathrm{n}\}$ is first processed with WordPiece [14], positional and segment embeddings, and then added with

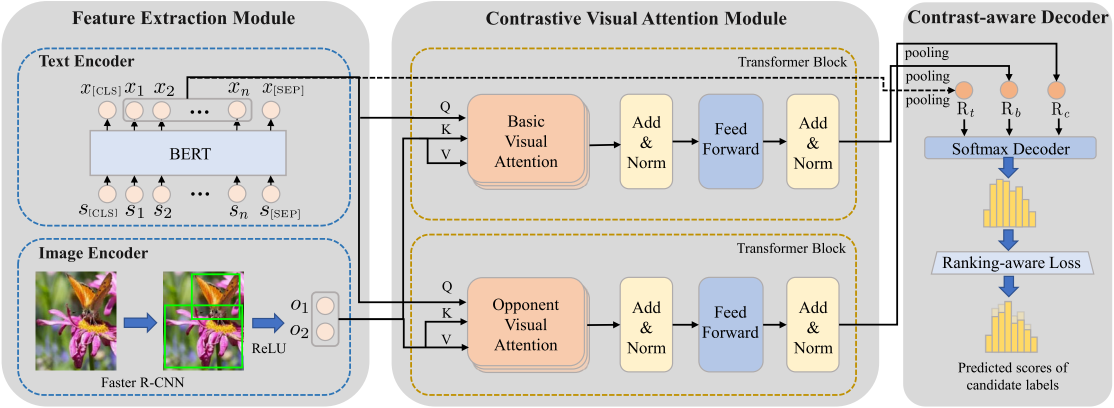

**Fig. 2**: Overall architecture of our proposed Language-guided Contrastive Attention (L-CVA) approach to C-RPD.

mark "[CLS]" and "[SEP]" as the first and the last special BERT tokens respectively. Lastly, this new word sequence is fed into a multi-layer bidirectional Transformer encoder [15] to obtain a sequence of word embedding vectors, i.e., $\mathrm{X} = [x_1, x_2, ..., x_n]$, where $x_i \in \mathbb{R}^d$.

**Vision Features.** Given an input image $I_i$, we employ the pre-trained object detection model, i.e., Faster R-CNN with ResNet-101 [16], to extract the objects within the image. For each input image, the objects with the confidence scores above 0.6 (fine-tuned with the development set) are selected as the objects of the image. For each object, we extract the output of the penultimate layer inside ResNet as the object vector and obtain a vector sequence of objects as $\hat{\mathrm{V}} = [\hat{v}_1, \hat{v}_2, ..., \hat{v}_m]$, where $\hat{v}_i \in \mathbb{R}^{2048}$. Then, we feed the object vector sequence into a fully connected layer so as to obtain the final object vector sequence as $\mathrm{O} = [o_1, o_2, ..., o_m]$, where $o_i = \mathrm{ReLU}(\mathrm{W}_o \hat{v}_i)$ and $o_i \in \mathbb{R}^d$. Here, $\mathrm{W}_o \in \mathbb{R}^{d\times 2048}$ is the trainable parameter.

### 3.2. Contrastive Visual Attention Module

To model the contrastive visual information, our contrastive visual attention module has two attention models, i.e., basic visual attention and opponent visual attention, for computing the aligned and misaligned visual representation respectively.

**Basic Visual Attention.** To highlight the image-regions focused by a human, we leverage a basic visual attention [10] to capture the aligned image-regions towards the language and obtain the aligned visual representation. Different from Lu et al. [10], we leverage objects as the basic semantic units of the image to align with the language. Moreover, we replace their basic soft-attention with the transformer block [15]. Specifically, inspired by Tsai et al. [17], we use a cross-modal transformer to learn to align the objects with the language. Given the word embedding sequence X and object vector sequence O, we define the *Queries* as $\mathrm{Q} = \mathrm{XW}_Q$, *Keys* as $\mathrm{K} = \mathrm{OW}_K$, and *Values* as $\mathrm{V} = \mathrm{OW}_V$, where $\mathrm{W}_Q \in \mathbb{R}^{d\times d}$, $\mathrm{W}_K \in \mathbb{R}^{d\times d}$ and $\mathrm{W}_V \in \mathbb{R}^{d\times d}$ are trainable parameters. Then, the visual attention for the $i$-th head is computed as:

$$\mathrm{selfAtt}_i(\mathrm{X}, \mathrm{O}) = \mathrm{softmax}(\frac{\mathrm{QK}^\top}{\sqrt{d}})\mathrm{V} \tag{1}$$

where $\sqrt{d}$ is the scaling factor. Then, all outputs of $h$ heads are concatenated as follows:

$$\mathrm{selfAtt(X{,}O)} = [\mathrm{selfAtt}_1(\mathrm{X{,}O}), ..., \mathrm{selfAtt}_h(\mathrm{X{,}O})]\mathrm{W}_h \tag{2}$$

where $\mathrm{W}_h \in \mathbb{R}^{hd\times d}$ and $[,]$ is the concatenate operation. Lastly, a residual position-wise feed-forward layer and a pooling operation are used to compute the final aligned visual representation $\mathrm{R}_b \in \mathbb{R}^d$ as:

$$\mathrm{R}_b = \mathrm{PL}\big(\mathrm{LN}\big(\mathrm{X} + \mathrm{FFN}\big(\mathrm{LN}\big(\mathrm{X} + \mathrm{selfAtt}(\mathrm{X}, \mathrm{O})\big)\big)\big)\big) \tag{3}$$

where $\mathrm{LN}(\cdot)$ means layer normalization. $\mathrm{FFN}(\cdot)$ is feed-forward network and $\mathrm{PL}(\cdot)$ is the mean-pooling operation.

**Opponent Visual Attention.** The opponent visual attention aims to capture the objects misaligned with language and then compute the misaligned visual representation, in order to modeling the contrast with the aligned visual representation in the followed contrast-aware decoder. Specifically, given a word embedding matrix X and the image object vector sequence O, we define *Queries*, *Keys* and *Values* like the basic visual attention. Then, the opponent visual attention $\mathrm{oppoAtt}$ is computed by modifying Eq. (1) in a basic transformer block, i.e.,

$$\mathrm{oppoAtt}_i(\mathrm{X}, \mathrm{O}) = \frac{1}{m-1}\big(\mathbb{1} - \mathrm{softmax}(\frac{\mathrm{QK}^\top}{\sqrt{d}})\big)\mathrm{V} \tag{4}$$

where $\mathbb{1} \in \mathbb{R}^{n\times m}$ is a unit matrix whose values are all 1 and the operation $\mathbb{1} - \mathrm{softmax}(\cdot)$ is applied to calculating opponent attention weights. $\frac{1}{m-1}$ is used to maintain the sum of the opponent attention weights for all $m$ objects is 1. Note that, the weight parameters of Eq. (1) and (4) are shared for better learning the discrimination between the aligned and misaligned visual representations. Finally, we can compute the misaligned visual representation $\mathrm{R}_c \in \mathbb{R}^d$ as follows:

$$\mathrm{R}_c = \mathrm{PL}\big(\mathrm{LN}\big(\mathrm{X} + \mathrm{FFN}\big(\mathrm{LN}\big(\mathrm{X} + \mathrm{oppoAtt}(\mathrm{X}, \mathrm{O})\big)\big)\big)\big) \tag{5}$$

### 3.3. Contrast-aware Decoder

Once obtained the language representation $\mathrm{R}_t = \mathrm{PL}(\mathrm{X})$ ($\mathrm{R}_t \in \mathbb{R}^d$), the aligned visual representation $\mathrm{R}_b$ and the misaligned visual representation $\mathrm{R}_c$ of the input text-image pair $P_i = (T_i, I_i)$, we employ two softmax operations to compute the prediction probabilities of the aligned and misaligned visual representations for the ground-truth label $y_i$ as follows:

$$p_b(y_i|P_i) = \mathrm{softmax}([\mathrm{R}_t, \mathrm{R}_b]\mathrm{W}_r + b_r) \tag{6}$$

$$p_c(y_i|P_i) = \mathrm{softmax}(\mathrm{R}_c\mathrm{W}_c + b_c) \tag{7}$$

where $\mathrm{W}_r \in \mathbb{R}^{2d\times 215}, \mathrm{W}_c \in \mathbb{R}^{d\times 215}, b_r \in \mathbb{R}^{215}, b_c \in \mathbb{R}^{215}$ are trainable parameters. Subsequently, to capture the contrastive visual information, $p_b(y_i|P_i)$ is used to maximize the probability of aligned visual representation $\mathrm{R}_b$ (concatenated with the language representation $\mathrm{R}_t$) for ground-truth label $y_i$, while $p_c(y_i|P_i)$ is used to minimize the probability of misaligned visual representation $\mathrm{R}_c$ for $y_i$. Both the former and the latter are formulated as two contrast-aware loss functions in the first and second term of Eq. (8) respectively. In addition, since the number of personality categories is relatively massive, up to 215 traits, we design a ranking-aware loss function, aiming at enlarging the probability of the ground-truth label and reduce the top-$J$ highest probabilities of other negative labels (see the third term in Eq. (8)) simultaneously, for boosting the classification performance. Finally, the objective function $\mathcal{L}$ for C-RPD is defined by combining the contrast-aware and ranking-aware losses as:

$$\mathcal{L} = -\sum_{i=1}^{M}\big(\mathrm{log}p_b(y_i|P_i) - \mathrm{log}p_c(y_i|P_i) + \sum_{j=1}^{J}\log(1-p_b(y^*_{i,j}|P_i))\big) \tag{8}$$

where $y_i$ is the ground-truth label for the $i$-th text-image pair $P_i$. $y^*_{i,j}$ is the predicted label that has the $j$-th highest probability among the other negative labels for the $i$-th sample. $M$ is the number of training text-image pairs. In our approach, we set $J = 5$ which is fine-tuned with the development set.

## 4. EXPERIMENTATION

### 4.1. Experimental Settings

**Data Settings.** We construct the dataset for the C-RPD task from the PERSONALITY-CAPTIONS data [22] (Shuster et al. [22] focus on generating personality-oriented image captioning. Different from them, this paper aims to use both text and image for performing the task of personality detection.). The C-RPD dataset contains 201795 text-image pairs and 215 personality traits. Each pair is labeled with one trait. In this work, we then treat the C-RPD task as a multi-class classification problem of 215 categories. Besides, this dataset is divided into training/dev/test by following the ratio setting by Shuster et al. [22] and the number of samples for all personality trait labels is balanced in our C-RPD dataset.

**Implementation Details.** In all our experiments, BERT-base (uncased) is fine-tuned with our C-RPD dataset and the parameters of this BERT model are following Devlin et al. [5] with the dimension 768 for all word vectors. For the objects detected by Faster R-CNN [6], the final dimension of the object vector is also set to be 768. In the contrastive visual attention module, the number of heads in basic visual attention and opponent visual attention is 8. All other hyper-parameters are fine-tuned according to the development set. Specifically, we initialize all weights of other layers by the Glorot uniform initializer [23]. We adopt the Adam optimizer [24] with an initial learning rate of $10^{-4}$ and set the regularization weight of parameters as $10^{-5}$. In addition, the dropout rate is 0.5 and the batch size is 32. To facilitate further research, our work is released via github (https://github.com/Elegdawnce/L-CVA).

**Evaluation Metrics.** The performance is evaluated with *Accuracy* (Acc.) and *Macro-F1* (F1). Here, F1 is calculated as $F = \frac{2PR}{P+R}$, where $P$ and $R$ are averaged on the precision/recall of all categories. Moreover, $t$-test [25] is used to evaluate the significance of the performance difference.

**Baselines.** For thorough comparison, we implement the following unimodal and multimodal approaches to C-RPD as baselines: **1) Char-RNN** [7], a textual personality detection approach with a char-level RNN to extract language-agnostic features. **2) Doc-LSTM** [18] is a neural network approach to document-level sentiment classification with a gated LSTM to learn text representation. **3) VGG** [19], an image classification approach with a VGG-19 model to extract image features for personality detection. **4) ResNet** [16], an image classification approach with a ResNet-101 model to extract image features for personality detection. **5) BERT** [5], one state-of-the-art textual encoding models to learn text representation for personality detection. **6) CoATT** [4], a state-of-the-art multimodal approach to named entity recognition with an attention mechanism for the text-image alignment. **7) MulT** [17], a state-of-the-art attention-based multimodal approach for multimodal sentiment analysis. **8) FMN** [9], one state-of-the-art multimodal approach for personality detection. **9) UDMF** [20], another state-of-the-art multimodal approach to infer personality traits of social media users. **10) MEMI** [21], another state-of-the-art attention-based multimodal approach for personality detection. In our implementation of all the above multimodal approaches, we only use text and image modalities as its inputs.

### 4.2. Experimental Results

Table 1 shows the performance comparison of various approaches to the C-RPD task. From this table, we can see that: **1) Unimodality Performance.** When only using the text modality, the large-scale pre-trained BERT approach performs better than Char-RNN and Doc-LSTM. This indicates the appropriateness of leveraging the large-scale pre-trained BERT to mine the language information for C-RPD. When

**Table 1**: Performance comparison of various approaches to C-RPD, where **Unimodality** denotes the input is unimodality (either text or image) while **Multimodality** denotes the input consists of both the text and image. Top$N$ denotes the accuracy of the top $N$ discovered personality traits with the highest probabilities.

| | **Approaches** | TOP1 | | TOP5 | | TOP10 | | TOP15 | | TOP20 | | TOP25 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. |
| **Unimodality (Text)** | Char-RNN [7] | 5.1 | 6.5 | 19.5 | 20.5 | 30.0 | 30.2 | 38.0 | 37.5 | 44.3 | 43.3 | 49.1 | 47.8 |
| | Doc-LSTM [18] | 5.4 | 6.8 | 21.0 | 21.8 | 32.6 | 32.6 | 40.4 | 40.0 | 46.2 | 45.5 | 51.2 | 50.4 |
| | BERT [5] | 7.2 | 9.5 | 27.6 | 29.3 | 41.8 | 42.7 | 50.0 | 50.3 | 56.5 | 56.6 | 61.6 | 61.6 |
| **Unimodality (Image)** | VGG [19] | 0.4 | 1.3 | 1.9 | 2.2 | 4.2 | 4.4 | 6.5 | 6.4 | 8.8 | 8.7 | 11.2 | 10.9 |
| | ResNet [16] | 0.7 | 1.9 | 2.0 | 2.3 | 4.5 | 4.5 | 7.0 | 6.6 | 10.7 | 9.0 | 13.2 | 11.2 |
| **Multimodality (Text + Image)** | CoATT [4] | 5.7 | 7.5 | 20.6 | 20.8 | 31.9 | 32.3 | 40.1 | 40.4 | 46.3 | 46.3 | 51.5 | 51.2 |
| | MulT [17] | 6.2 | 7.9 | 21.7 | 22.6 | 33.1 | 33.1 | 41.3 | 40.9 | 48.0 | 47.3 | 52.1 | 51.7 |
| | FMN [9] | 4.8 | 6.0 | 18.2 | 19.4 | 30.0 | 30.0 | 37.7 | 37.0 | 44.1 | 43.1 | 49.5 | 48.1 |
| | UDMF [20] | 6.7 | 8.2 | 22.9 | 23.9 | 34.5 | 34.2 | 42.6 | 42.3 | 48.8 | 48.3 | 54.3 | 53.7 |
| | MEMI [21] | 5.3 | 6.8 | 19.2 | 20.6 | 30.6 | 31.1 | 39.0 | 39.0 | 45.0 | 44.7 | 50.5 | 49.9 |
| | **L-CVA** | **11.5** | **13.0** | **31.3** | **32.3** | **44.2** | **44.3** | **52.3** | **52.3** | **58.1** | **58.0** | **62.8** | **62.7** |

**Table 2**: Effectiveness study of our proposed L-CVA approach.

| **Approaches** | TOP1 | | TOP5 | | TOP10 | | TOP15 | | TOP20 | | TOP25 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. |
| **L-CVA** | **11.5** | **13.0** | **31.3** | **32.3** | **44.2** | **44.3** | **52.3** | **52.3** | **58.1** | **58.0** | **62.8** | **62.7** |
| w/o misaligned visual representation | 9.9 | 11.4 | 29.6 | 30.3 | 42.8 | 42.8 | 50.8 | 50.5 | 56.9 | 56.6 | 61.5 | 61.1 |
| w/o aligned visual representation | 9.2 | 10.7 | 28.9 | 29.6 | 41.5 | 41.4 | 50.2 | 49.7 | 56.3 | 55.8 | 61.0 | 60.5 |
| w/o language representation | 1.5 | 2.6 | 3.2 | 3.0 | 7.5 | 5.9 | 11.7 | 9.0 | 15.6 | 11.7 | 19.3 | 14.4 |
| w/o ranking-aware loss function | 10.7 | 12.0 | 30.5 | 31.5 | 43.4 | 44.0 | 51.6 | 51.8 | 57.3 | 57.4 | 61.9 | 61.9 |
| using spatial features as image-regions | 9.9 | 11.0 | 28.6 | 29.5 | 41.5 | 42.1 | 49.7 | 50.1 | 56.2 | 56.4 | 60.9 | 61.0 |

only using the image modality, both the state-of-the-art image classification approaches VGG and ResNet perform slightly better than a random performance. This indicates the difficulty in detecting personality using single image modality and encourages us to incorporate both language and vision information for C-RPD. **2) Multimodality Performance.** When using both the text and image modalities, UDMF and MulT perform better than most of the above unimodal approaches (except BERT). This confirms the helpfulness of considering the multimodal information in the C-RPD task. In comparison, our approach L-CVA significantly outperforms ($p$-value $< 0.01$) all the above unimodal and multimodal approaches in terms of both Acc. and F1. These results encourage us to capture the cognitive difference information (i.e., both the language information and the contrastive visual information) for the C-RPD task. It should be mentioned that the Top1 performance is relatively low, which is mainly due to the fuzzy boundary between the personality traits, e.g., *sentimental* and *gloomy*, motivating us to incorporate the correlation among similar traits to further boost the performance in the future.

## 5. ANALYSIS AND DISCUSSION

**Effectiveness Analysis.** We conduct an effectiveness study to comprehensively evaluate our L-CVA approach in Table 2. From this table, we can see that: **1)** Minimizing the probabilities of the misaligned representation in the second term of Eq. (8) can improve the Top1 Acc. by 1.6%. This justifies the effectiveness of using opponent visual attention to mine the misaligned object features and encourage us to consider the contrastive visual information for C-RPD. **2)** Integrating the aligned visual representation in Eq. (6) can improve the Top1 Acc. by 2.3%. This justifies the effectiveness of using basic visual attention to capture the image-regions focused by humans and encourages us to consider the aligned object features for C-RPD. **3)** Using the language information can improve the Top1 Acc. by 10.4%, and using only aligned and misaligned visual representation can improve the Acc. by an average of 1.85% in contrast to the single image modality baseline ResNet. This indicates the importance of

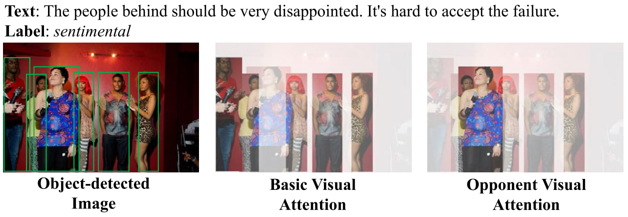
**Text**: The people behind should be very disappointed. It's hard to accept the failure.
**Label**: *sentimental*
Object-detected Image
Basic Visual Attention
Opponent Visual Attention

**Fig. 3**: Visualization of the outputs of contrastive visual attention in our L-CVA approach to C-RPD. The darker object means this object is more important.

the language information and again verifies the helpfulness of integrating the additional contrastive visual information to C-RPD. **4)** Incorporating the ranking-aware loss function in Eq. (8) can improve the Top1 Acc. by 1.0%. This justifies the effectiveness of using ranking-aware loss function to remedy the issue of the massive labels. **5)** Employing the objects as the image-regions for capturing the contrastive visual attention will improve the Top1 Acc. by 2.0%, compared with directly treating the spatial CNN features extracted by ResNet as image-regions. This justifies the effectiveness of using objects to perform the alignment with language.

**Visualization Analysis.** In order to get a better understanding of our L-CVA approach and validate the effectiveness of our proposed contrastive visual attention mechanism, we provide a visualization analysis on the development set. Specifically, in Fig. 3, we visualize the basic visual attention and opponent visual attention outputs respectively. From this figure, we can see that: **1)** Basic Visual Attention succeeds to align the language "*... the people behind ...*" with the image objects in which the humans stand behind the woman who wears blue clothes. **2)** Opponent Visual Attention succeeds to capture the cheerful woman who wins the champion. This indicates that our L-CVA approach can effectively model both the aligned and misaligned visual features for precisely understanding the contrastive visual information (i.e., focusing on the negative persons rather than the positive persons) and correctly predicting the *sentimental* personality trait.

## 6. CONCLUSION

In this paper, we propose a new C-RPD task, aiming at leveraging the cognitive difference phenomenon of human beings to detect their real-time personality traits. In particular, we propose a Language-guided Contrastive Visual Attention (L-CVA) approach to address this C-RPD task. The main idea of this proposed approach is to capture not only the language information but also the contrastive visual information for addressing C-RPD. Detailed experiments show that our proposed L-CVA approach significantly outperforms the state-of-the-art baselines to C-RPD. In our future work, we would like to solve other challenges in C-RPD, such as incorporating the external ConceptNet knowledge base to perform better reasoning and leveraging the large-scale cross-modal pre-training model (e.g., ImageBERT [26]) to further boost the performance. Besides, we would like to apply our L-CVA approach to other psychological analysis tasks, e.g., multimodal emotion analysis and multimodal anxiety detection.

## 7. REFERENCES

[1] L. Goldberg, "An alternative "description of personality": The big-five factor structure," *Pers Soc Psy.*, 1990.

[2] C. S. Carver, M. F. Scheier, and et al., "Control theory: A useful conceptual framework for personality–social, clinical, and health psychology.," *Psy. bulletin*, 1982.

[3] W. Croft and D. A. Cruse, *Cognitive linguistics*, Cambridge University Press, 2004.

[4] Q. Zhang and et al., "Adaptive co-attention network for named entity recognition in tweets," in *AAAI*, 2018.

[5] J. Devlin, M. Chang, and et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[6] S. Ren, "Faster R-CNN: towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[7] F. Liu, S. Nowson, and et al., "A language-independent and compositional model for personality trait recognition from short texts," in *EACL*, 2017.

[8] L. Liu and et al., "Analyzing personality through social media profile picture choice," in *AAAI*, 2016.

[9] O. Kampman, E. J. Barezi, and et al., "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," in *ACL*, 2018.

[10] D. Lu and et al., "Visual attention model for name tagging in multimodal social media," in *ACL*, 2018.

[11] G. A. Sigurdsson and et al., "Visual grounding in video for unsupervised word translation," in *CVPR*, 2020.

[12] C. Song and et al., "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018.

[13] X.Duan and et al., "Contrastive attention mechanism for abstractive sentence summarization," in *EMNLP*, 2019.

[14] Y. Wu, M. Schuster, and et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, 2016.

[15] A. Vaswani, "Attention is all you need," in *NIPS*, 2017.

[16] K. He, X. Zhang, and et al., "Deep residual learning for image recognition," in *CVPR*, 2016.

[17] Y. H. Tsai and et al., "Multimodal transformer for unaligned multimodal language sequences," in *ACL*, 2019.

[18] D. Tang, B. Qin, and et al., "Document modeling with gated recurrent neural network for sentiment classification," in *EMNLP*, 2015.

[19] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[20] G. Farnadi, J. Tang, and et al., "User profiling through deep multimodal fusion," in *WSDM*, 2018.

[21] L.Wu, "Speaker personality recognition with multimodal explicit many2many interactions," in *ICME*, 2020.

[22] K. Shuster, S. Humeau, and et al., "Engaging image captioning via personality," in *CVPR*, 2019.

[23] X. Glorot, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[25] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *SIGIR*, 1999.

[26] D. Qi, "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," *CoRR*, 2020.