Arnab Bhattacharya · Janice Lee Mong Li ·
Divyakant Agrawal · P. Krishna Reddy ·
Mukesh Mohania · Anirban Mondal ·
Vikram Goyal · Rage Uday Kiran (Eds.)

LNCS 13247

Database Systems for Advanced Applications

27th International Conference, DASFAA 2022
Virtual Event, April 11–14, 2022
Proceedings, Part III

3 Part III

Springer

Contents – Part III

Text and Image Processing

Emotion-Aware Multimodal Pre-training for Image-Grounded Emotional Response Generation

Zhiliang Tian[1], Zhihua Wen[2], Zhenghao Wu[1], Yiping Song[3], Jintao Tang[3], Dongsheng Li[2](✉), and Nevin L. Zhang[1]

[1] The Hong Kong University of Science and Technology, Sai Kung, Hong Kong SAR, China
{ztianac,lzhang}@cse.ust.hk, zwubq@connect.ust.hk

[2] Science and Technology on Parallel and Distributed Laboratory, National University of Defense Technology, Changsha, Hunan, China
{zhwen,dsli}@nudt.edu.cn

[3] National University of Defense Technology, Changsha, Hunan, China
{songyiping,tangjintao}@nudt.edu.cn

Abstract. Face-to-face communication leads to better interactions between speakers than text-to-text conversations since the speakers can capture both textual and visual signals. Image-grounded emotional response generation (IgERG) tasks requires chatbots to generate a response with the understanding of both textual contexts and speakers' emotions in visual signals. Pre-training models enhance many NLP and CV tasks and image-text pre-training also helps multimodal tasks. However, existing image-text pre-training methods typically pre-train on images by recognizing or modeling objects, but ignore the emotions expressed in the images. In this paper, we propose several pre-training tasks in a unified framework that not only captures emotions from images but also learns to incorporate the emotion into text generation. The pre-training involves single-modal learning to strengthen the ability to understand images and generate texts. It also involves cross-modal learning to enhance interactions between images and texts. The experiments verify our method in appropriateness, informativeness, and emotion consistency.

Keywords: Multimodal · Conversation · Emotion · Pre-training · Generation

1 Introduction

Most conversation systems [21,45,59] lead a text-to-text dialog between users and chatbots. However, most people prefer face-to-face communication due to the accessibility of the speaker's visual signals, like facial expressions and body language. After analysing those signals, chatbots can garner speakers' emotional

Z. Tian and Z. Wen–The two authors contributed equally to this work.

A. Bhattacharya et al. (Eds.): DASFAA 2022, LNCS 13247, pp. 3–19, 2022.
https://doi.org/10.1007/978-3-031-00129-1_1

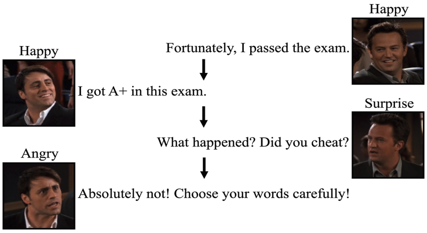
Happy
Fortunately, I passed the exam.
Happy
I got A+ in this exam.
Surprise
What happened? Did you cheat?
Angry
Absolutely not! Choose your words carefully!

Fig. 1. An example of IgERG. Images hint speakers' emotions of in multi-turn dialogs.

states and make empathetic responses. Image-grounded emotional response generation (IgERG) [17] generates a response to context utterances while understanding both the users' textual contexts and emotional states reflected in images, where images contain users' facial expressions and gestures. IgERG has several practical applications, including AI therapist capturing a counselee's emotions via facial expressions, and AI baby-sitter conversing with kids who are moody and unstable (Fig. 1).

Deep learning achieves impressive performances on conversation systems but relies highly on large-scale corpora [44,56]. IgERG's training samples are hard to collect and limited since each sample contains conversations and the aligned speakers' visual signals. Pre-training models [9] help many NLP tasks without large-scale corpus by pre-training on large-scale unsupervised corpora.

Researchers propose multimodal pre-training for tasks involving both textual and visual data. To facilitate image understanding, they apply pre-training tasks of object classification [20] and object region modeling [5]. To ensure image-text alignment, they pre-train them with image captioning [16], text-conditioned object classification [23,51], and image-text entity matching [22]. Those tasks typically pre-train by recognizing objects in images, modeling objects, and interacting with other modalities (i.e. text) about object-related information. Those tasks ignore the styles or emotions reflected in the images. Such pre-training fits downstream tasks that require an understanding of the objects in images (e.g. image-text retrieval [3] and visual question answering [2]).

However, images imply both object-related and object-independent (e.g. styles, sentiments, or emotions) information. Object-independent information is more helpful to some applications, including multimodal sentiment analysis [62], multimodal style transfer [18,65], and IgERG [17]. Thus, pre-training of capturing object-independent signals enhances those tasks but faces two challenges: 1. object-independent information is harder to define and describe than objects. Intuitively, recognizing a human face is easier than describing the emotion seen in the facial expressions; 2. if the application involves multiple images (e.g. speakers' images in multi-turn conversations of IgERG)[1], the development of emotions reflected from a sequence of images is hard to model and utilize.

[1] 81.1% samples contain multiple emotions in IgERG.

In this paper, we propose a multimodal pre-training method that enhances the ability to capture emotions from a series of images and learns to incorporate the emotions into text generation. Such pre-training helps image-grounded emotional response generation (IgERG), where models generate responses given textual context and a sequence of speakers' images implying the emotions.

Particularly, to enhance the emotion perception, our model learns to model a series of emotions from a sequence of images via an image emotion sequentially labeling (`IESL`) and an image emotion classification (`IEC`) task. We obtain large-scale coarse data for those tasks via data augmentation, since those tasks have limited supervised data. To enhance the text generation, we apply BART's pre-training and masked language modeling (`MLM`) task to our pre-training. To enhance cross-modal interactions between emotion understanding and text generation, our model learns to incorporate emotions in the generated text by a controllable image-to-text generation (`C-I2T`) task. The text generation is controlled by the emotions detected from images. We construct a transformer-based framework carrying all pre-training tasks. Our contributions are as follows,

- We propose catering to object-independent emotional information in images during pre-training, which is a less explored topic.
- We propose several pre-training tasks in a unified framework that prompts image understanding, text generation, and cross-modal interactions.
- Our model obtains state-of-the-art performance on the IgERG task.
- We construct several multimodal pre-training datasets and will release them.

2 Related Work

Image grounding conversation has two categories. The first one, visual dialog, leads conversations to discuss the objects or events reflected in images [1,8]. Shuster et al. [46] assign specific styles to speakers in visual dialog. Agarwal et al. [1] study the effects of explicitly encoding historical contexts into visual dialog models. The second category captures speakers' emotions from images for an empathetic conversation. Poria et al. [39] and Hazarika et al. [14] discover speakers' emotion in conversations. Huber et al. [17] use a Seq2Seq [52] model to capture scene sentiment and use a facial encoding tool to extract the facial expressions from the images in conversation. Our task falls in this category. The above methods do not involve pre-training.

Emotional conversation systems can be categorized into two directions: expressing chatbot's emotions and catering to speakers' emotions. The first direction aims to enable chatbot to respond conditioned on a given emotion [7,50,68]. Colombo et al. [7] append the given emotion label on the input utterance. Song et al. [50] apply a lexicon-based attention to a Seq2Seq model and encourage Seq2Seq to implicitly express emotion. The second direction detects speakers' emotions and make empathetic responses according to speakers' emotion [12,38,42]. Skowron et al. [47] build a chatbot with the ability to detect user's emotion states. Rashkin et al. [42] propose a pipeline system that predicts emotion words and feed the explicit emotion words into a neural conversation model.

Lin et al. [28] propose an end-to-end neural conversation model. Lin et al. [29] and Zhong et al. [67] apply GPT [41] and BERT [9] to empathetic conversation. Another types of chatbots detect emotions from non-textual modalities [14], including tone or body languages. Some chatbots captures speakers' personalities in conversation [48,54,63].

Pre-trained language models significantly enhance natural language understanding (NLU) tasks [9,60], since language models contains commonsense knowledge [61] or language understanding abilities [33,37]. BERT [9] achieves state-of-the-art results on a wide range of NLU tasks. For natural language generation (NLG), GPT [41] trains to generate texts auto-regressively. BART [19] propose several text permutation techniques, such as text infilling and sentence shuffling. Multimodal (image-text) pre-training [23] mainly has four kinds of tasks as follows. 1. *object classification*: Su et al. [51] extract object representation with Faster-RCNN [43] model and apply object classification to pre-training. Li et al. [20] consider linguistic clues in object classification. 2. *object region modeling*: Li et al. [25] reconstruct the masked image regions by referring to the remaining part. Chen et al. [5] jointly train the masked region classification and the masked region modeling. 3. *image conditioned text generation*: Li et al. [26] pre-train with image captioning that typically describes the semantic information of image with a textual title. 4. *image-text matching*, encourages models to align texts and images in the semantic level [20]. Li et al. [22] learn to predict whether the given image-text pair are semantically aligned. Huang et al. [16] consider the alignment between multilingual texts and images. Those methods typically focus on pre-training to understand image objects, but our pre-training caters to object-independent information (e.g. emotions) in images.

3 Approaches

3.1 Overview

Our downstream task is the image-grounded emotional response generation (IgERG) that generates a response $\hat{S}_n$ given a series of context sentences $S = \{S_1, S_2, ..., S_{n-1}\}$ and the speakers' images corresponding to each sentence $I = \{I_1, I_2, ..., I_n\}$, where I_i, with speakers' facial expressions and gestures, reflects the emotional state of the speaker in the i-th turn. We construct a framework for the pre-training tasks and the fine-tuning task. As shown in Fig. 2, the framework mainly consists of four components:

- **Image encoder** E_{img} represents an image I_i with a vector $\boldsymbol{v}_i$.
- **Generator** is a transformer model [56] to generate text, where the pink and grey colored blocks in Fig. 2 indicate its encoder and decoder. Its input can be a sequence of sentences S, a sequence of images I, or the mixture of S and I; its output is the generated texts. If the input comes across an image I_i, the generator employs the image encoder to transfer the image into a vector $\boldsymbol{v}_i$ as its input. The generator's encoder output is defined as $\boldsymbol{e}_i$ at the i-th step.
- **Image emotion classifier** $D_{\text{img\_emo}}$ predicts emotion labels based on $\boldsymbol{e}_i$.

- **Text emotion evaluator** $D_{\text{txt\_emo}}$ evaluates texts generated by the generator.

Our tasks involve four types of datasets: 1.image-only datasets: each sample has one image or a sequence of images $I = \{I_1, I_2, ..., I_m\}$, $(m \geq 1)$; 2. text-only datasets: each sample is a raw sentence; 3. image-text datasets: each sample consist of several sentences $S = \{S_1, S_2, ..., S_n\}$ and each sentence has an aligned image $I = \{I_1, I_2, ..., I_n\}$; 4. textual emotion datasets: each sample has a sentence and its emotion label. We propose three kinds of pre-training tasks:

- **Pre-train for image emotion discovery** learns to detect image emotions on all datasets involving images, including image-only and image-text datasets.
- **Pre-train for text generation** learns to generate text on all datasets with texts, including text-only and image-text datasets.
- **Pre-train for cross-modal interaction** learns to generate text controlled by images on image-text datasets.

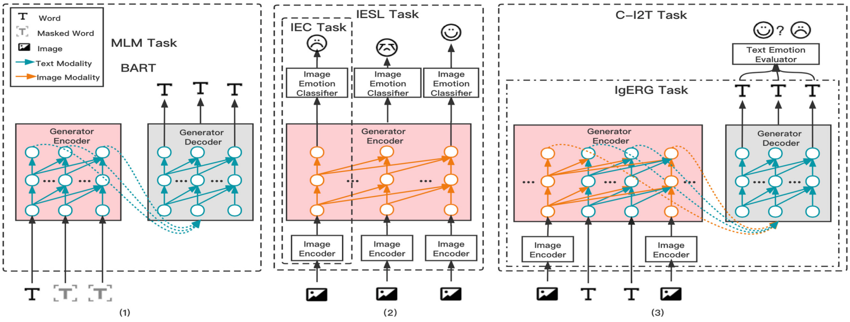

Fig. 2. Our model architecture with pre-training tasks (MLM, BART, IESL, IEC, and C-I2T) and fine-tuning tasks (IgERG). The generator, with an encoder (in pink) and a decoder (in grey), is shared among all subfigures. The image encoder and image emotion classifier are also shared in all time steps. Subfigure 1 shows the pre-training on text generation (`MLM` and `BART`) mentioned in Sect. 3.2; Subfigure 2 shows the pre-training of emotion discovery (`IEC` and `IESL`) mentioned in Sect. 3.2 and Subfigure 3 shows the pre-training of `C-I2T` and the fine-tuning of `IgERG` mentioned in Sect. 3.2 & 3.3. The orange and green arrows indicate information flows in image and text modality. (Color figure online)

3.2 Pre-training

Pre-training for Image Emotion Discovery. We propose two pre-training tasks: image emotion sequentially labeling (`IESL`) and image emotion classification (`IEC`) (Fig. 2.2), which enables our model to discover emotions from an image or a sequence of images. As a result, the image encoder and the generator's encoder represent each image I_i with a vector $\boldsymbol{e}_i$ that implies emotions.

IESL models a series of emotions reflected in a sequence of images $I = \{I_1, I_2, ..., I_m\}$, where the image sequence is consecutive screenshots from a video. Its output is a sequence of emotion labels. IESL involves the image encoder (E_{img}), generator's encoder (GE), and the image emotion classifier ($D_{\text{img\_emo}}$). The image encoder E_{img} first maps each image I_i into a vector $\boldsymbol{v}_i$. The generator's encoder GE takes the output vectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_m\}$ as the input and treats it as the sequence of "word embeddings" (As a transformer model, the generator's encoder inputs are word embeddings). Then, the generator's encoder obtains the output vectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_m\}$. To train the emotion labeling task, we feed each encoder output $\boldsymbol{e}_i$ to the image emotion classifier $D_{\text{img\_emo}}$ to predict the emotion at i-th step. Equation 1 and 2 show those operations, where E_{I_i} is the emotion label of I_i, K is emotion category number, $\mathbb{I}$ is an indicator function. We introduce the structure of the image encoder and emotion classifier in Sect. 3.4.

$$\boldsymbol{v}_i = E_{\text{img}}(I_i) \ \text{ for } \ i \in [1, m], \qquad \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_m\} = GE(\{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_m\}), \tag{1}$$

$$L_{\text{IESL}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \mathbb{I}(E_{I_i} = k) \log P(D_{\text{img\_emo}}(\boldsymbol{e}_i) = k). \tag{2}$$

IEC learns to detect the emotion reflected in a single image, where the input is an image and the output is an emotion label. This task shares the same components as IESL. As for the training procedure, this task can be treated as a special case of IESL, where the image sequence length is one. IEC's advantage is this task can work on the datasets with only one image in each sample.

We apply data augmentation to build large-scale coarse supervised datasets since supervised data in the above tasks are limited. We first collect several videos from TV series and movies, where the characters have rich facial expressions or gestures during the conversations. Then, we collect unlabeled raw images by taking screenshots from the TV series videos. Finally, we generate pseudo labels for the unlabeled images with Face++ API[2], which categorizes images into seven classes of emotion (e.g. fear). Therefore, each sample in those datasets has a sequence of images and labels, which is used for both IEC and IESL tasks.

In total, we obtain 5.9M training samples from six datasets for the above tasks. Except for one dataset that is an existing supervised image emotion dataset, we apply data augmentations on the other five since they do not have emotion labels. Among the five datasets, four come from the TV series and one is an image-text dataset [32].

Pre-training for Text Generation. To enhance text generation ability, our model applies the pre-training of BART [19] and pre-trains on the text data from the image-text datasets via masked language modeling (MLM) task (Fig. 2.1). BART pre-trains a model combining bidirectional and auto-regressive transformers. As rerunning BART's pre-training requires a great number of GPUs[3], we

[2] console.faceplusplus.com.cn.

[3] It roughly requires 64 GPUs for more than 2 weeks.

use the pre-trained BART released by FairSeq[4]. BART provides initial parameters for the generator. The motivation behind using BART is that BART has the same architecture as the generator and BART has achieved state-of-the-art performance on text generation (e.g. dialogue and summarization).

MLM learns to generate randomly masked tokens in a raw sentence [31]. MLM only employs our generator, where the encoder receives a masked sentence and the decoder reconstructs the original sentence. We train MLM on 6M sentences from the image-text dataset, and each sentence is a sample in the pre-training.

Pre-training for Cross-Modal Interaction. As shown in Fig. 2.3, we propose a pre-training task, controllable image-to-text generation (C-I2T), to enhance the interaction between images and texts. C-I2T learns to incorporate the emotions detected from the images into text generation and encourages emotions in texts and images to be consistent. The training samples consist of a series of images $I = \{I_a, I_{a+1}, ..., I_b\}$ and a series of sentences $S = \{S_a, S_{a+1}, ..., S_b\}$. The pre-training requires the model to generate a sentence $\hat{S}_b$ given the corresponding images $\{I_a, I_{a+1}, ..., I_b\}$ and the previous sentences $\{S_a, S_{a+1}, ..., S_{b-1}\}$.

We concatenate the sentences and images as a long sequence $\{I_a, S_a, I_{a+1}, S_{a+1}, ..., S_{b-1}, I_b\}$, where the sentence and image occurs alternately. Each utterance and its corresponding image gather together and the image is ahead of the utterance since we treat the image as the condition of its utterance. We use the image encoder to transfer each image I_i into a vector $\boldsymbol{v}_i$, and generator's word embedding layer transfers a sentence with l words into a sequence of l word vectors $\{\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, ..., \boldsymbol{w}_{il}\}$. Then, we obtain a sequence of vectors $\{\boldsymbol{v}_a, \boldsymbol{w}_{a1}, \boldsymbol{w}_{a2}, ..., \boldsymbol{v}_i, \boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, ..., \boldsymbol{w}_{(b-1)l}, \boldsymbol{v}_b\}$ and feed the sequence into the generator to generate $\hat{S}_b$. Feeding a combined image-text sequence into the generator bridges the cross-modal attention to enhance image-text interaction, since every two inputs have a connection via self-attention in the generator (i.e. transformer) including the two inputs in different modalities. The text emotion evaluator $D_{\text{txt\_emo}}$ (See details in Sect. 3.4) predicts emotion of the generated text $\hat{S}_b$.

C-I2T's loss has two terms as Eq. 3: a negative likelihood loss $\mathcal{L}_{NLL}$ that encourages generating the ground truth and a cross-entropy loss $\mathcal{L}_{emo}$ that encourages the generated text $\hat{S}_b$ to have the same emotion as I_b, where E_{I_b} is I_b's emotion, $D_{\text{txt\_emo}}$ is the text emotion evaluator, and α balances the two losses.

$$\mathcal{L}_{\texttt{C-I2T}} = \mathcal{L}_{NLL} + \alpha\mathcal{L}_{emo}, \quad \mathcal{L}_{emo} = \sum_{k=1}^{K} \mathbb{I}(E_{I_b} = k) \log P(D_{\text{txt\_emo}}(\hat{S}_b) = k). \tag{3}$$

C-I2T uses an image-text dataset collected from the TV series. Each original sample is an image-text sequence consisting of a series of images $I = \{I_1, I_2, ..., I_n\}$ and sentences $S = \{S_1, S_2, ..., S_n\}$. The images I are from the screenshots of a video clip and the sentences S are the corresponding subtitles. All the sub-sequences in the image-text sequence act as training samples for

[4] dl.fbaipublicfiles.com/fairseq/models/.

C-I2T: each original sample (image-text sequence) is decomposed into multiple C-I2T's samples: $\{I_a, I_{a+1}, ..., I_b\}$, $\{S_a, S_{a+1}, ..., S_b\}$ (for $\forall a, b \in (0, n]$, $a+1 < b$). As mentioned in Sect. 3.2, we generate pseudo emotion E_{I_i} for I_i with Face++.

Procedure of Pre-training. We conduct the above pre-training tasks in a unified framework. Our pre-training consists of two phases as mentioned in Fig. 2. In the first phase, we load the pre-trained BART model to initialize the parameters of our generator and obtain the text emotion evaluator $D_{\text{txt\_emo}}$ by fine-tuning BERT on textual emotion datasets. In the second phase, based on the BART and $D_{\text{txt\_emo}}$, we jointly pre-train the model on IEC, IESL, C-I2T, and MLM tasks. To train the second phase, we mix up all the datasets and shuffle all batches of samples (Samples in a batch come from one dataset). During the second phase of pre-training, we pre-train IEC and IESL on the samples from datasets involving images; we pre-train MLM on samples involving texts and we pre-train C-I2T on samples from image-text datasets. After all the pre-training, the pre-trained model provides the initial parameters for the downstream IgERG task.

3.3 Fine-Tuning on the IgERG Task

IgERG is our downstream task. As shown in Fig. 2.3, most layers in our pre-trained model provide the initial parameters for the downstream task, except the classifier $D_{\text{img\_emo}}$ mentioned in Sect. 3.2 and the evaluator $D_{\text{txt\_emo}}$ mentioned in Sect. 3.2. Fine-tuning on the downstream task follows the pre-training operations of C-I2T except for the use of L_{emo} loss. We concatenate the images and context sentences into a sequence $\{I_1, S_1, I_2, S_2, ..., S_{n-1}, I_n\}$ as the input of our model. The image encoder and generator's word embedding layer transfers images and words into vectors. The generator's encoder receives those vectors and the generator's decoder outputs the response $\hat{S}_n$.

3.4 Structure of Model Components

Our model (Fig. 2) consists of an image encoder, an image emotion classifier, a text emotion evaluator, and a generator. The image encoder represents an image I_i with a fixed dimensional vector $\boldsymbol{v}_i$ and feeds the vector to the generator. The image encoder is ResNet-50 [15] without softmax layer. We replace the top fully-connected layer with another fully-connected layer, in which the output dimension equals the dimension of the generator's word embedding. Thus, the image encoder's output $\{\boldsymbol{v}_0, \boldsymbol{v}_1, ..., \boldsymbol{v}_m\}$ is fed into the generator by acting as its input embeddings. The image encoder is trained by emotion discovery (IEC and IESL) task (Sect. 3.2), and it is frozen in other pre-training tasks and fine-tuning.

The image emotion classifier $D_{\text{img\_emo}}$ predicts the emotion of an image. Its input is an image vector (i-th image's encoder output $\boldsymbol{e}_i$). The image emotion classifier consists of a fully-connected layer and a softmax layer. It transfers $\boldsymbol{e}_i$ into a probability distribution. IEC and IESL train and use this classifier.

The text emotion evaluator $D_{\text{txt\_emo}}$ is a BERT-based classifier that predicts the emotion of the generated text in `C-I2T` task. It is pre-trained by BERT[5] and fine-tuned on text emotion datasets. Only `C-I2T` involves this $D_{\text{txt\_emo}}$.

The generator is the core part that accomplishes the pre-training and fine-tuning tasks. It is trained in all pre-training tasks. The generator is the "transformer base" model in [56] with encoder (pink blocks in Fig 2) and a decoder (grey blocks in Fig 2). Both encoder and decoder have 6 stacked layers.

4 Experiments

4.1 Experimental Settings

Datasets and Hyper-parameters. Our pre-training tasks involve four types of datasets: 1. for image-only datasets, we use a supervised image emotion dataset, RAF-DB [24], with 15k samples. We collect 4.9M samples from four TV series (How I Met Your Mother, This Is Us, The Big Bang Theory, and Person of Interest). Each of the samples is a sequence of screenshots, and we assign emotional labels on those samples via Face++ API. 2. The text-only datasets are BART's datasets consisting of four datasets [13,35,55,69]. 3. For image-text datasets, we use OpenViD [32] dataset with 1M samples. It comes from the movies with subtitles, where the image comes from the sequence of screenshots and texts are the corresponding subtitles. Notice that, we split original OpenViD samples to obtain 21M samples for `C-I2T` as mentioned in Sect. 3.2. 4. We use three textual emotion datasets [4,27,39] with 0.92M samples in total, and they share the same emotion space with Face++ API. Our downstream task uses OpenViD [32] dataset with 0.9M/50k/50k samples for training/validation/testing. In all baselines and our model, following [32], we set the dimension of word embeddings and hidden layers to 512 and the dimension of image vector $\boldsymbol{v}$ to 1000. We set the dropout rate to 0.3 and the learning rate to 3e−5 during pre-training and fine-tuning. The factor α in Eq. 3 is 1. $D_{\text{txt\_emo}}$ is fine-tuned with a learning rate of 1e−5. We released our code[6].

Evaluation Metrics. We evaluate all methods with automatic and human evaluations. Our automatic metrics consists of: (1) Appropriateness, *Bleu-N* [21, 36] and *Nist-N* [10,40,53] measure N-gram match between outputs and ground truthes. *CiDEr* [57] is widely used in image captioning. (2) Informativeness, *Dist-N* [21,49] evaluates the response diversity via unique n-gram proportion in all responses. *Ent-N* [34,53] is the entropy on word count distribution. (3) Emotion Const. *E-Acc* and *E-F1* measure the accuracy and F1 score between the emotion of the last input image and the generated responses. Those metrics show the consistency of images and texts in terms of emotion.

[5] huggingface.co/bert-base-uncased.

[6] Our code is available at: github.com/stupidHIGH/MM-Pre-train.

We conduct human evaluations in three aspects: 1. overall quality *Qual* (fluency, relevance, and grammaticality) of the generated response, 2. informativeness and lexical diversity of results (*Info*). 3. *Emo* (in three factors: emotion consistency between input images and generated texts, emotion being well expressed in text, and whether the personas reflected in outputs and dialogue histories are consistent.) We hire five commercial annotators to annotate five copies of 300 randomly selected test samples with a 5-scale rating.

Baseline Methods. We verify our model by comparing following methods.

- No pre-training. `Trs` is the transformer model [56] without using images. `Trs+FV` and `Trs+CV` denote the transformer using images via Faster-RCNN [43] and ResNet [15], respectively. The above baselines come from Meng et al. [32] and do not involve pre-training. `Trs+CV` serves as the fine-tuning model of all the following methods for a fair comparison.
- Text-only pre-training. `BART` denotes the model pre-trained with BART [19] on text data and fine-tuned with `Trs+CV`'s model, where BART is widely-used pre-train model for text generation tasks (e.g. dialog and summarization).
- Image-Text pre-training. Oscar [24] is the multimodal pre-training model that pre-trains to align the object semantics.[7] `Oscar+BART`'s encoder comes from Oscar and its decoder comes from BART.

Table 1. Overall performance on automatic metrics and human evaluations (Results in rows 1 to 3 match the official github page (github.com/ShannonAI/OpenViDial), where the authors [32] publish the revised results of their paper. Kappa score [11] among annotators is 0.43 (moderate agreement among annotators)).

Pre-train type	Model	Appropriateness						Informativeness					Emotion const		
		Bleu3	Bleu4	Nist3	Nist4	CIDEr	Qual	Dist3	Dist4	Ent3	Ent4	Info	E-F1	E-Acc	Emo
No Pre-train	Trs	1.79	1.00	0.766	0.771	0.069	2.41	0.003	0.004	3.53	3.56	2.05	0.099	0.228	1.87
	Trs+FV	2.07	1.19	0.874	0.880	0.098	2.39	0.027	0.041	4.82	4.99	2.34	0.111	0.217	1.84
	Trs+CV	2.04	1.15	0.875	0.882	0.097	2.58	0.020	0.031	5.01	5.18	1.96	0.118	0.246	2.03
Text-only	BART	2.77	1.70	1.069	1.086	0.140	3.04	0.155	0.224	6.84	7.28	2.64	0.129	0.288	2.32
Image-Text	Oscar	2.18	1.21	0.865	0.870	0.098	2.88	0.021	0.034	5.78	6.07	2.35	0.121	0.248	2.21
	BART+Oscar	1.99	1.07	0.881	0.888	0.104	2.79	0.022	0.037	5.47	5.71	2.31	0.124	0.256	2.09
Image-Emotion-Text	Ours	**3.00**	**1.94**	**1.139**	**1.158**	**0.164**	**3.13**	**0.216**	**0.306**	**7.44**	**7.93**	**2.98**	**0.137**	**0.312**	**2.49**

4.2 Overall Results

Table 1 shows the performance of all methods. Among the baselines without pre-training (first 3 rows), `Trs+FV` and `Trs+CV` perform better than `Trs`, which verifies the image is helpful for our task. `Trs+CV`'s emotion consistency is a little higher than `Trs+FV`, so the following methods employ `Trs+CV` as the fine-tuning

[7] We choose Oscar as our baseline, since Cho et al. [6] and Li et al. [26] reported Oscar outperforms most existing multimodal pre-training models, including UNITER [5], XGPT [58], VL-BART [6] and VL-T5 [6] on VQA, NLVR, and image captioning.

model. Text pre-training BART obtains much higher performance than no pre-training models showing the power of pre-training. Text pre-training enhances *Dist* scores a lot. The reason is BART learned on a corpus with a large variety of sentences (160GB data) during its pre-training.

Image-text pre-training baselines (Oscar and Oscar+BART) outperform no pre-training models in the emotion consistency and get similar performances in other metrics. It shows image pre-training helps models to understand emotions. The performance of the two baselines is not as satisfactory as BART, showing the current image-text pre-training models do not fit for our task. As a model only pre-trained to align the object semantics between image and text, Oscar does not consider the object-independent information (e.g. emotion) during pre-training.

Considering object-independent information (emotion) in image-text pre-training, Ours surpasses all the baselines in all metrics. The improvements in emotion consistency show our pre-training enhances the cross-modal interaction that expresses image emotions in texts. The significant improvements on *Dist* (+39% in *Dist3* and +36% in *Dist4*) indicate suitable image-text pre-training can highly enlarge the informativeness of the generated text. The response generated by our model tends to have higher quality and appropriateness, as our model improves by 8% and 14% on *Bleu3* and *Bleu4*. Our promotion on emotion consistency verifies that considering the emotion reflected from the images during pre-training helps the model on understanding and expressing emotions.

4.3 Ablation Studies on Pre-training Tasks

Table 2. Ablation studies on the effectiveness of different pre-training tasks. "Ours − X" (row 2 to 5) indicates our full mode without pre-training of task X.

	Appropriateness					Informativeness				Emotion Const	
	Bleu3	Bleu4	Nist3	Nist4	CIDEr	Dist3	Dist4	Ent3	Ent4	E-F1	E-Acc
Ours	**3.00**	**1.94**	**1.139**	**1.158**	**0.164**	**0.216**	**0.306**	**7.44**	**7.93**	**0.137**	**0.312**
Ours − IESL	2.82	1.76	1.097	1.116	0.150	0.185	0.268	7.23	7.70	0.136	0.296
Ours − IEC	2.80	1.74	1.065	1.082	0.145	0.163	0.234	6.79	7.19	0.133	0.285
Ours − MLM	2.83	1.77	1.089	1.107	0.147	0.174	0.253	7.01	7.45	0.133	0.291
Ours − C-I2T	2.77	1.73	1.093	1.111	0.146	0.181	0.258	6.99	7.41	0.135	0.302

Table 2 shows the ablation studies on our proposed pre-training tasks. We construct a model variant by removing one specific pre-training task from our full model. As we propose four tasks, we obtain four variants and compare them to our full model Ours. For example, Ours−IEC denotes Ours without IEC.

Ours excels all model variants in row 2 to 5. It demonstrates all the proposed pre-training tasks are necessary and contribute to our full model. Among the rows 2 to 5, removing IESL from Ours (row 2) decrease the performance on all metrics. IESL is crucial since it models the change of emotions in a conversation session, which helps the downstream model capture the tendency of the speaker's emotion and generate suitable responses. Removing IEC from the full model causes the largest performance drop, which shows IEC is much important in pre-training. The reason is that it benefits from high-quality supervised training data with rich emotions (see dataset analysis in Sect. 4.5). Knowing that appropriateness measures the matching between ground truth and generated texts, it is reasonable to see that Ours − C-I2T performs the second-worst on appropriateness because C-I2T is the only task that pre-trains to make generated sentences and the ground truth similar.

4.4 Emotion Expression in Text Generation

To measure the ability to express emotions in the generated texts, we propose *%EW* and *%F1* for evaluation. *%EW* denotes the percentage of emotion words (occurs in an emotion word list[8] of each generated response. Considering the ground truth words, we measure the precision and recall of generated emotion words that match the ground truth as Eq. 4, where E denotes the emotion word list, $\hat{S}$ denotes a generated sentence, and S denotes a ground truth sentence. *%F1* is the harmonic average of the precision and recall.

$$\text{Precision} = \frac{|\hat{S} \cap E \cap S|}{|\hat{S} \cap E|}, \quad \text{Recall} = \frac{|\hat{S} \cap E \cap S|}{|E \cap S|} \tag{4}$$

The experimental results shown in Table 3 reflect the effectiveness of emotion expressions in two aspects: the quantity and quality of generated emotion words. As for the quantity, our model tends to generate far more emotion words in responses than the baselines (shown in the *%EW* of Table 3). This advantage mainly comes from the pre-training of *C-I2T*, since *C-I2T* learns to generate words to express emotions. As for the quality, the results on *%F1* verify the generated emotion words from our model are more likely to match the correct emotions (ground truth emotions). Our model works well owing to the pre-training of understanding emotions (IEC and IESL) and express emotions (C-I2T).

Table 3. The percentage of emotion words in generated responses from different methods, which shows the degree of emotion expressed by the model.

	Trs	Trs+CV	Trs+FV	BART	Oscar	Oscar+BART	Ours
%EW	1.5	3.7	2.7	4.3	4.3	3.5	**5.1**
%F1	3.1	15.3	10.8	60.2	12.1	12.0	**69.4**

[8] saifmohammad.com/WebPages/lexicons.html.

4.5 Studies on the Dataset Selection

Table 4. Comparisons among different types of image datasets on `IEC`. R is to conduct `IEC` on RAF-DB dataset with its original labels. R_FA, T_FA, and P_FA denote `IEC` on RAF-DB, TV, and Pose dataset with pseudo labels by Face++ API.

	Label	Emotion distribution	Image content	Appropriateness					Informativeness				Emotion const	
				Bleu3	Bleu4	Nist3	Nist4	CIDEr	Dist3	Dist4	Ent3	Ent4	E-F1	E-Acc
R	Real	Balanced	Facial Expressions	**2.20**	**1.27**	0.997	**1.009**	**0.107**	0.043	0.073	5.449	5.791	0.134	**0.325**
R_FA	Pseudo	Balanced	Facial Expressions	2.18	1.25	**1.005**	1.007	1.005	**0.052**	**0.087**	**5.707**	**6.105**	**0.136**	0.325
T_FA	Pseudo	Unbalanced	Face + Gestures	2.09	1.17	0.965	0.976	0.099	0.028	0.047	5.357	5.659	0.129	0.293
P_FA	Pseudo	Unbalanced	Few Facial Expressions	2.15	1.18	0.891	0.900	0.092	0.019	0.033	4.981	5.222	0.126	0.276

Most pre-training methods succeed owing to "big data" and our method also requires large-scale image datasets for `IEC`, `IESL`, and `C-I2T`. The qualities and characteristics of those datasets are crucial for our training. However, it's hard to select datasets from various image datasets. Here, we analyze the effectiveness of different types of image emotion datasets, verify the data augmentation, and give suggestions on dataset selection. In Table 4, we choose four datasets with different types and compare the performances of pre-training `IEC` on them.

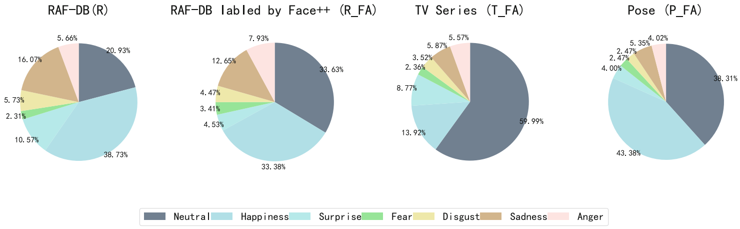

Fig. 3. Emotion distributions of four image datasets in Sect. 4.5.

We choose four datasets considering their image contents and emotion distributions: 1. RAF-DB (`R`) [24] is a supervised image emotion classification dataset, where most images describe facial expressions and the emotion distribution is well balanced. 2. We obtain a new dataset `R_FA` by relabeling the original RAF-DB via Face++. 3. TV dataset (`T_FA`) is a mixture of four TV series datasets (Sect. 4.1) labeled by Face++. The images contain speakers' facial expressions and gestures and its emotion distribution is unbalanced. 4. Pose dataset (`P_FA`) is a mixture of portrait photography datasets [30,64,66], where the face region covers a small part of each image. Its emotion distribution is unbalanced since the only two emotions cover 82% samples. We train all datasets with the same scale (15k samples) for a fair comparison. Figure 3 shows emotion distributions of the above datasets. The distributions of RAF-DB and RAF-DB labeled by