Visual Emotion Representation Learning via Emotion-Aware Pre-training

Yue Zhang[1] , **Wanying Ding**[2] , **Ran Xu**[3] and **Xiaohua Hu**[1]
[1]Drexel University, College of Computing & Informatics, Philadelphia, PA, USA
[2]JPMorgan Chase & Co., Palo Alto, CA, USA
[3]Salesforce Research, Palo Alto, CA, USA
yz559@drexel.edu, wanying.alice@gmail.com, xurantju@gmail.com, xh29@drexel.edu

Abstract

Despite recent progress in deep learning, visual emotion recognition remains a challenging problem due to the ambiguity of emotion perception, diverse concepts related to visual emotion, and lack of large-scale annotated datasets. In this paper, we present a large-scale multimodal pre-training method to learn visual emotion representation by aligning emotion, object, and attribute triplet with a contrastive loss. We conduct our pre-training on a large web dataset with noisy tags and fine-tune on smaller visual emotion classification datasets with class label supervision. Our method achieves state-of-the-art performance for visual emotion classification.

1 Introduction

Understanding the emotion from visual content is an important and attractive interdisciplinary research area that is influenced by research in computer vision, psychology, and natural language processing. Recent advances in deep learning [He *et al.*, 2016] and representation learning [Radford *et al.*, 2021] have driven significant progress in various computer vision and multimodal tasks, yet visual emotion categorization remains a very challenging task. Intuitively, humans can perceive the emotion from an image "in-the-wide" due to different reasons, for example, the object and scene, action and activity, photo style, text in the image, concepts or story inferred from the image, to name a few (see Fig.1 and Fig. 4 for illustration).

Most visual emotion categorization works adopt an image classification framework to learn a mapping from pixels to emotion categories with neural networks in an end-to-end fashion, however, there are two major challenges in this direction. First, due to the intrinsic ambiguity of emotion perception and the mechanics of data acquisition (more details in Sec. 2), severe dataset bias has been identified by [Panda *et al.*, 2018]. Particularly, the issues of *positive bias* where images with certain emotion tag are dominated by a single concept, scene, or object; and *negative bias* where diverse visual concepts "in-the-wild" related to a certain emotion are not collected in the datasets, limiting

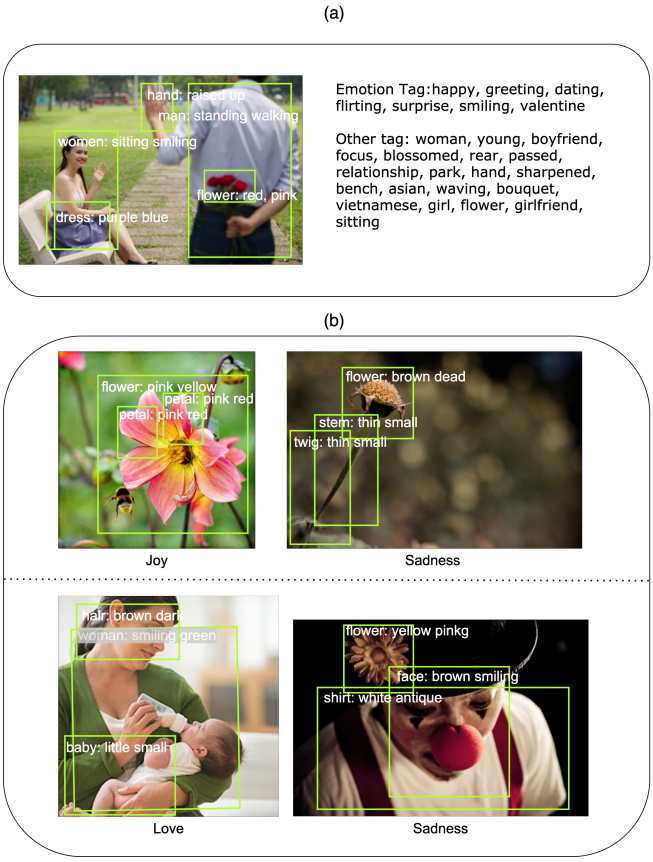

Figure 1: Block (a) illustrates a sample from the pre-training dataset Emotion Stock, where both visual features such as objects of "person", "dress", "flower" and their corresponding attributes such as "red", "smiling", as well as a set of emotion or non-emotion tags from the dataset form a visual-language pair. Block (b) illustrates samples from emotion classification datasets Emotion6 and UnBiasedEmotion with emotion labels. See Sec. 1 for more analysis.

the accuracy, generalization, and robustness of visual emotion categorization. Second, different from image classification or object detection where a particular visual region with a specific category is learned, the reasoning of emotion from an image usually involves the compositionality of multiple elements such as object-in-context [Kosti *et al.*, 2017;

Kosti *et al.*, 2020], human-object interaction, object attribute, intention, and functionality, etc. For example, as illustrated in Fig. 1, the combination of "smiling women" and "little baby", under the bright and warm lighting, evoke the human perception of *joy*. The difficulty of learning the mapping between low-level features and high-level emotion is also known as "affective gap" [Machajdik and Hanbury, 2010], while the compositionality of visual concepts is largely unexplored.

Recently, a line of research [Panda *et al.*, 2018; Wei *et al.*, 2020] works leverage large-scale web images to train visual emotion representations to help mitigate dataset bias, and partially address the sample efficiency problem. The web data contains both image and weakly-labeled tags from either image metadata or search query. This paradigm improves system generalizability. However, the visual emotion representation is trained on visual and texture features separately before a linear transformation to align features from both modalities. In addition, visual signals such as objects and attributes are not explicitly modeled to infer emotion. We believe large-scale datasets provide new opportunities to study the compositionality of objects and attributes for emotion understanding, as illustrated in Fig. 1, dense and accurate object categories and bounding boxes could be obtained quite accurately thanks to recent advances. Moreover, the same object with different attributes, such as the "pink flower" versus the "dead flower" could lead to dramatically different emotion perceptions; and similarly, for the same attribute of "smile" and even similar body pose, different objects and context also evoke different emotions.

In this paper, we propose to learn visual emotion representations via multimodal pre-training on a large-scale **emotion related** dataset - Stock Emotion [Wei *et al.*, 2020]. First, we utilize scene graph detection [Han *et al.*, 2021] trained on large-scale open domain datasets to extract a diverse set of objects and attributes from an image, together with noisy tags from the image, we use a transformer-based fusion model for the pre-training task. With knowledge from open-domain datasets, we hope the model could learn both low-level features and middle-level visual concepts and narrow down the 'affective gap'. Second, we attempt to learn the alignments between emotion and underline visual concepts such as objects and attributes apart from image-text alignments; in particular, we design emotion-object alignment and emotion-attribute alignment objectives for pre-training. Existing works [Li *et al.*, 2020; Zhang *et al.*, 2021] inspire our cross-modal fusion module design, while our research focus and major contributions are on the emotion-aware module and analysis of visual emotion understanding.

Section 3 describes details of the model and Section 4 presents the experiments and analysis. From our experiments, we find

1. Pre-training helps the model to generalize better.
2. Emotion tag is critical to learning visual emotion representation.
3. Emotion related dataset is critical compared with pre-training on general image-text datasets.

2 Related Works

In this section, we focus on the literature review of visual emotion categorization and vision-language pre-training (VLP).

Visual Emotion Categorization Early works [Machajdik and Hanbury, 2010; Wang *et al.*, 2013] leverage low-level image features such as color, edge, texture, etc, and train a linear emotion classifier such as SVM for categorization tasks. More recently, deep learning based feature extractor, especially CNN, has been used to learn visual representation [You *et al.*, 2015; Peng *et al.*, 2015], and the accuracy has been significantly improved from hand-crafted low-level features.

With the success of CNN based deep features, there are two research directions that attempt to address the "affective gap" we discussed in Section 1. The first line of research targets at understanding the relationships between objects and scenes that evoke emotion. [Kosti *et al.*, 2020] create a dataset Emotic to focus on images containing people in context, and CNNs for both human body and whole image are used for better understanding of the visual emotion evoked by people and context. [Peng *et al.*, 2016] and [Yang *et al.*, 2018b] utilize salient objects or regions from an image and attempt to find "affective regions" that align with emotion concepts better. The second line of research attempt to address dataset bias and sample efficiency problem with larger and more diverse datasets. [You *et al.*, 2016] collect a dataset of over 3 million weakly labeled images by querying 8 emotion tags as keywords from Flickr and Instagram, then curate 23,000 images with human annotation via Amazon Mechanic Turk (AMT). [Panda *et al.*, 2018] collect WEBEmo with 268,000 images from a stock image dataset consists 6 high-level categories and 25 fine-grained categories. The search queries are diversified to reduce bias and a curriculum learning method is used to learn image representation with noise labels. [Wei *et al.*, 2020] collect an even larger dataset with 1.17 million images, and each image is accompanied by a set of tags from image metadata. Emotion related tags are extracted from the entire tag set, alignment between image feature and tag embedding, as well as emotion tag prediction from image feature are jointly learned.

Vision-Language Pre-training VLP leverages image-text pairs to conduct multimodal representation learning. Training on large-scale unlabeled datasets usually produces effective multimodal representations for downstream tasks such as image caption, VQA, cross-domain retrieval, etc. A number of works [Lu *et al.*, 2019; Su *et al.*, 2020; Chen *et al.*, 2020; Gan *et al.*, 2020] use object detectors to detect ROI of the image as visual tokens, then use multi-layer transformers [Vaswani *et al.*, 2017] across both modalities. [Li *et al.*, 2020; Zhang *et al.*, 2021] propose to use object tags detected in images as anchors to help bridge the semantic gap between vision and language. Recently, CLIP [Radford *et al.*, 2021] and ALIGN [Jia *et al.*, 2021] scale the number of image-text pairs to 400 million and 1.2 billion; without recomputing region-based image features, very strong multimodal representation can be learned. [Li *et al.*, 2021] propose to use an image-text contrastive loss to align image and text features from the unimodal encoders before fusing them to a multimodal encoder.

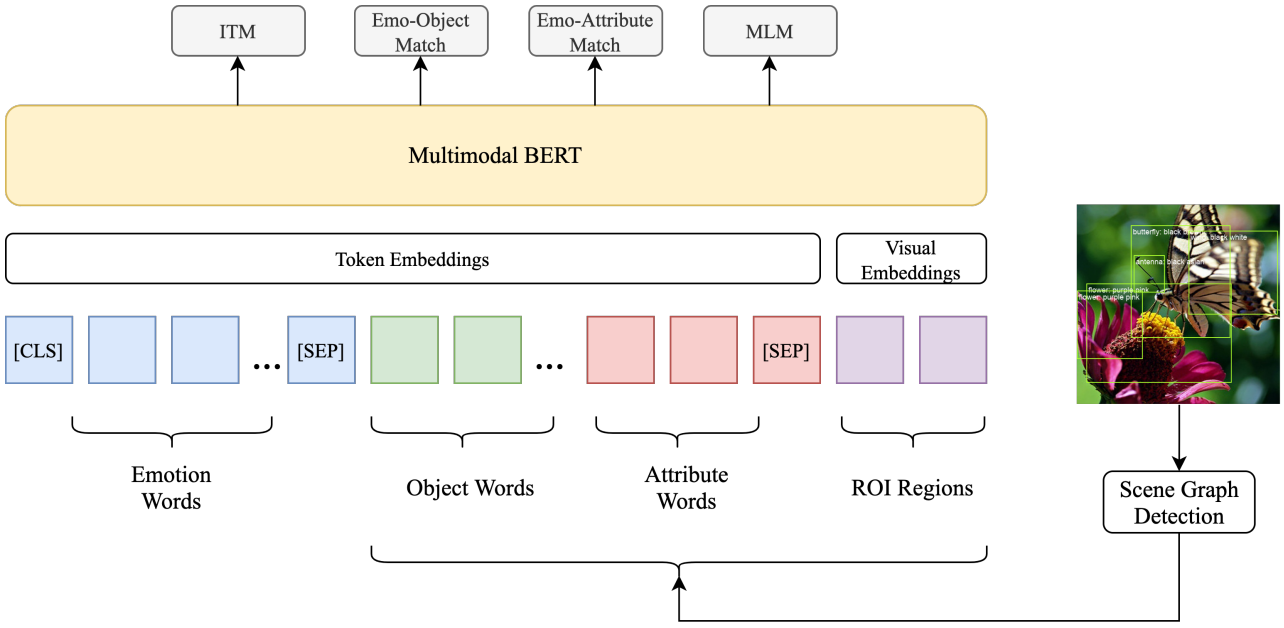

Figure 2: Architecture of pre-train model. Region based visual features, object words, and the corresponding attribute words are extracted with a scene graph detection model. A masked language modeling (MLM) loss is applied on all texture tokens, an image-text matching (ITM) loss is applied on emotion words and visual and texture tokens from an image, and another two emotion-object and emotion-attribute match losses are applied to object and attribute separately

.

3 Methods

In this section, we introduce our emotion aware pre-training architecture in Section 3.1, then describe the adaptation to the visual emotion categorization task in Section 3.2.

3.1 Pre-training

We utilize Stock Emotion [Wei *et al.*, 2020] for pre-training because it is the largest dataset with image-text pair where the text contains emotion related words due to curation methodology. As introduced in 1, our core motivation is to embed the emotion reasoning into the pre-training objective by aligning the compositional visual features and concepts with emotion words. Specifically, we propose an emotion-object matching loss and an emotion-attribute matching loss in addition to traditional masked language modeling loss and image-text matching loss.

Model Architecture As illustrated in Fig. 2, for an image from the dataset, we use a scene graph detector [Han *et al.*, 2021] to predict object category $\boldsymbol{o}$, bounding box $\boldsymbol{b}$ and corresponding attributes $\boldsymbol{a}$ from the object region, region of interest (ROI) feature $\boldsymbol{v}$ in each object bounding box is also extracted. Specifically, the visual feature is extracted with a Mask-RCNN [He *et al.*, 2017] detector. The four corners as well as the width and height of the object bounding box are normalized w.r.t. image size to produce a 6-dimensional spatial feature. Similar to [Su *et al.*, 2020; Li *et al.*, 2020], we concatenate both content and spatial feature and use a linear layer to project visual features to the same dimension of word embedding.

During dataset curation, a list of tags provided by the image uploaders is also collected to pair with each image. Among the entire tag list, a subset of 690 emotion related tags are manually selected by [Wei *et al.*, 2020] (see detailed process in Sec. 4.1). We regard emotion related tags $\boldsymbol{e}$ as our text inputs. A tuple of ($\boldsymbol{e}$, $\boldsymbol{o}$, $\boldsymbol{a}$, $\boldsymbol{v}$) is sent to the embedding layer and produce text and visual embeddings, which are further used as inputs to a BERT model.

Emotion-Aware Matching We design a contrastive loss for both image-text matching and matching of emotion to objects and attributes. For image-text matching, we consider emotion words as language modality and both visual embedding and predicted objects and attributes as vision modality, similar to OSCAR [Li *et al.*, 2020]. This concept fits our downstream task better in the sense that emotion words are usually high-level concepts whereas object-attribute pairs are usually ground with images better. And with this setting, we are able to leverage objects and attributes as anchors to ground the learned emotion words in images.

To ground the emotion concept to objects or attributes, we design an intuitive matching mechanics for contrastive loss. We keep both emotion tags and object tags aligned and pollute attribute tags to learn the alignment between emotion and attributes, and similarly pollute object tags to learn object-emotion alignment. Visual features are kept all the time to help the grounding.

Concretely, for image-text pairs represented by ($\boldsymbol{e}$,$\boldsymbol{o}$,$\boldsymbol{a}$,$\boldsymbol{v}$), 50% of the sequence remain matched, 20% of $\boldsymbol{e}$ are randomly swapped with emotion words from another pair to construct negative samples for image-text matching, 15% of ob-

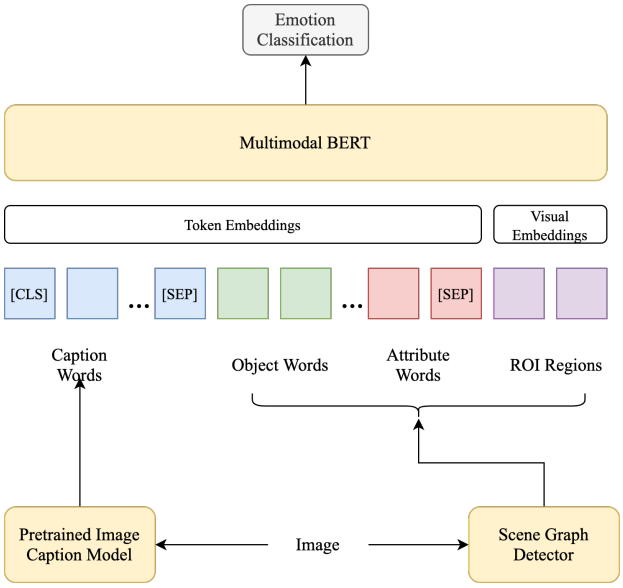

Figure 3: Architecture to adapt the pre-trained BERT model for emotion classification tasks. Images are sent to a pre-trained image caption model to generate captions as image tags, and the same scene graph detector as pre-training phase to generate object/attribute labels and visual features.

ject words and 15% of attribute words are randomly swapped with corresponding words from another pair to construct negative samples for emotion-object-attribute reasoning. The emotion-aware contrastive matching loss is defined as:

$$\mathcal{L}_{\mathcal{EAC}} = -E_{(e,o,a,v)\sim D} log p(y|f(e,o,a,v)) \quad (1)$$

We use [CLS] token to represent the representation of the multimodal sequence, and $f(e,o,a,v)$ represents a fully connected classification layer with softmax to predict a matching label p from [CLS]. As discussed above, p $\in$ (0,1,2,3) for matched sequence, unmatched image-text pair, unmatched emotion-object, and unmatched emotion-attribute.

Masked Language Modeling We use both visual features and text to predict masked tokens from the text. In particular, for this target we regard original emotion tags ***e*** and tags ***o*** and ***a*** inferred from the vision model as text because they share the same linguistic semantic space in spite of the different resources. We randomly masked out 15% of word tokens from the text and replace the original token with a special token [MASK], then a cross-entropy loss is applied to learn to infer the original token, concretely MLM is minimized as:

$$\mathcal{L}_{\mathcal{MLM}} = -E_{(v,T)\sim D} log p(\hat{y}|\hat{T}, v) \quad (2)$$

where T represents (***e***,***o***,***a***) tuple, $\hat{y}$ represents masked token and $\hat{T}$ represents remaining unmasked tokens from one sample.

To sum, the full pre-training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathcal{EAC}} + \mathcal{L}_{\mathcal{MLM}} \quad (3)$$

Implementation Details Our pre-training model consists of a pre-trained BERT base model (bert-base-uncased) with 123.7M parameters. The scene graph detection model is pre-trained on OpenImagesV5, Visual Genome, Objects365V1, and COCO datasets, with ResNeXt152-C4 backbone with an extra branch to predict attributes. We select up to 20 objects and 4 attributes per object for pre-training, based on detection confidence scores. We pre-train our model with 8 NVIDIA V100 GPUs, and the batch size of 384 image-text pairs. The learning rate is set as $2e^{-5}$ and we train for 30 epochs with AdamW. We set a maximum number of token sequence (including both visual features and words) length to 100.

3.2 Adapting to Visual Emotion Categorization

We frame the emotion classification task as a multi-label sequence classification task, as illustrated in Fig. 3, the input sequence consists of image tags, predicted objects and attributes words, and visual features from the scene graph detector. We regard the [CLS] token embedding as sequence embedding and use a linear layer (FC) and softmax layer to predict the emotion class for multi-class classification problems and use an FC layer without softmax to regress emotion classes for multi-label classification problems. Detailed settings for adaptation datasets are described in Sec. 4.1.

Since we do not assume the emotion classification task and datasets will include image tags, we propose two variants to address this problem. First, we simply do not include any image tags during emotion classification, the task is still sequence classification but there exists a gap between pre-training and adaptation. The second variant utilizes a pre-trained image caption model to generate captions for each image as image tags. This method keeps the gap between pre-training and adaptation small but potentially brings additional noise to the inputs. We will discuss the results of each variant in the ablation study section 4.3.

4 Experiments

In this section, we first describe the datasets for pre-training and adaptation in Section 4.1, then introduce baseline methods and experimental results in Section 4.2, we finally conduct ablation studies and summarize in Section 4.3.

4.1 Dataset

We pre-train our model utilizing one large dataset and test our pre-training model performance on four widely-used emotion classification datasets.

Stock Emotion [Wei *et al.*, 2020] is collected in several steps. The authors first search Adobe Stock with various emotion keywords and concepts, and rank the words associated with the images. After removing low-frequency keywords, 690 emotion words are manually selected from 2000 candidates. A total of 1.17 million images are collected with these 690 emotion words as queries. We use a 33k images subset of this dataset for the downstream tasks so remove this subset and use the remaining 1.13 million images for pre-training.

DeepEmotion [You *et al.*, 2016] contains 23,815 images labeled with eight emotion categories. This dataset is queried online by eight emotion keys from Flickr and Instagram, and

Figure 4: Sample images from the dataset for pre-training (StockImage) on the first row, and five datasets (EMOTIC, Emotion6, SE20K8, UnBiasedEmo, DeepEmotion) for emotion classification task.

each weakly-labeled image is categorized by 5 Amazon Mechanical Turk (AMT) workers. We follow the split setting in [Panda *et al.*, 2018] and [Wei *et al.*, 2020], randomly select 80% data for training and the rest 20% data for testing.

Emotion6 [Peng *et al.*, 2015] is collected by searching images related to six emotion keywords and their synonyms. The erroneous images are removed and each emotion category contains 330 images, thus 1980 images in total. The author lately created another dataset EmotionROI [Peng *et al.*, 2016] based on this dataset to provide areas evoking emotions in each image. In our experiment, we only train and test on Emotion6 dataset without utilizing the saliency map in EmotionROI. We follow the training setup in [Yang *et al.*, 2018a] with 80% data for training and 20% for testing.

SE30K8 is a subset of Stock Emotion, with 33k images and categorized into eight emotions. [Yang *et al.*, 2018a] utilize Amazon Mechanical Turk (AMT) and each image is labeled by five workers. In total 85% of images have been annotated by at least three annotators. Therefore, we use 27k images for our downstream task following [Wei *et al.*, 2020], with 22k training images and 5k testing images.

UnBiasedEmotion [Panda *et al.*, 2018] contains 3045 images with six emotion categories collected from Google. We follow the setting from [Wei *et al.*, 2020] and [Panda *et al.*, 2018] with 80% images for training and 20% for testing.

EMOTIC [Kosti *et al.*, 2020] contains 23,571 images and 34,320 annotated people with body and face bounding boxes. Each people is labeled with 26 categories in a multi-label setting. We follow the original split provided by the authors with 70% images for training, 10% for validation, and 20% for testing. We modify our single label prediction paradigm to multi-label prediction following [Kosti *et al.*, 2020], but only use the whole image and emotion classes (denoted as EMOTIC-I) instead of the additional body bounding box and VAD annotations in our experiments.

As illustrated in Fig. 4, the dataset distribution is diverse and across multiple domains such as advertisement images, natural images, posters, and cartons, etc.

4.2 Baseline Methods and Results

In this section, we compare our methods with a few baseline variants and other state-of-the-art algorithms. [Wei *et al.*, 2020] reports state-of-the-art performance on DeepEmotion [You *et al.*, 2016], SE30K8, [Wei *et al.*, 2020] and UnBiasedEmotion [Panda *et al.*, 2018], which outperforms other non pre-training methods by a large margin. [Yang *et al.*, 2018a] reports state-of-the-art performance on Emotion6 [Peng *et al.*, 2015] with additional annotations of saliency maps. [Kosti *et al.*, 2020] reports state-of-the-art performance on EMOTIC dataset.

Base-VG and Base-OID To compare pre-training with emotion awareness design with pre-training on general purpose image-text pair datasets, we directly use pre-trained models from OSCAR+ [Li *et al.*, 2020] for emotion dataset fine-tuning. Base-VG is trained with 5.5M image-text pairs from multiple large-scale datasets, and image tag inputs are trained with Visual Genome [Krishna *et al.*, 2017] dataset. Base-OID is pre-trained on the same dataset as Base-VG but the image tags are predicted from a model trained with Open Image V5 [Krylov *et al.*, 2021].

EAP-AT-T and EAP-ET-T We design a variant of emotion-aware objective that instead of randomly selecting emotion-object and emotion-attribute pairs separately as negative samples for contrastive loss, we select both object and attribute words and swap the entire object-attribute sequence with another sequence from the dataset. The motivation is to help ground emotion with image region easier from separately sampled counterparts. Specifically, we remain 50% matched sequence, and swap 25% of emotion tag sequence and swap another 25% of object-attribute sequence to construct the loss. Both EAP-AT-T and EAP-ET-T use this sampling strategy. "AT" here indicates another design choice to use both emotion related tags and other image tags together, while "ET" indicates we only use emotion related tags as inputs, as described in our method section.

EAP-ET-S1 and EAP-AT-S2 We design two variants of emotion-aware sample selection strategies where object and attribute are separately selected to construct contrastive objectives. EAP-ET-S1 is described in our method section, where either object word sequence or attribute word sequence is randomly selected for pollution. EAP-ET-S2 use both emotion tags and other image tags to construct matched sequence and pollute both tags as sequence to construct unmatched image-text sequences, but we omit attribute word sequence

Dataset	DeepEmotion	Emotion6	SE30K8	UnBiasedEmotion	EMOTIC-I
Metric	Accuracy	Accuracy	Accuracy	Accuracy	mAP
Previous SOTA	65.81 [1]	58.25* [2]	**69.78** [1]	81.45 [1]	**27.38*** [3]
Base-VG	64.18	53.28	63.06	59.93	20.03
Base-OID	66.29	54.55	63.68	58.46	20.09
EAP-AT-T	67.93	54.29	66.08	72.74	24.41
EAP-ET-T	68.73	**62.37**	**68.70**	**81.77**	**25.02**
EAP-ET-S1	**68.75**	60.61	66.08	74.71	24.19
EAP-ET-S2	68.67	59.60	68.14	76.25	24.18

Table 1: Emotion classification prediction accuracy on state-of-the-art methods and our baselines. Base-VG and Base-OID are pre-trained with general image-text pairs while our methods are pre-trained with domain specific (emotion related) image-text pairs. Previous SOTA [1] from [Wei *et al.*, 2020], [2] from [Yang *et al.*, 2018a] and [3] from [Kosti *et al.*, 2020]. [2] uses additional saliency maps and [3] uses both image and human body image, as well as additional VAD annotations to train the models.

when object word sequence is selected to construct emotion-object aware negative sequence, and omit object word sequence when attribute sequence is selected for negative sequence. The motivation of this design is to learn MLM loss with more diverse tokens but still only use emotion tags for emotion-object and emotion-attribute alignments.

We summarize our results in Table 1, and the results are evaluated with the same fine-tune architecture design, which uses captions generated from pre-trained model as image tags. First, compared with state-of-the-art methods, our methods outperform EmotionNet [Wei *et al.*, 2020] on DeepEmotion and UnBiasedEmotion datasets, while underperforming EmotionNet on SE30K8; and our methods outperform [Yang *et al.*, 2018a] on Emotion6 dataset even without saliency maps. EMOTIC-I dataset features human emotion, we suspect fewer objects and corresponding attributes from the dataset limit the advantages of our approach, but our simpler adaptation method is still competitive against the traditional method with the additional human body and VAD annotations. The results demonstrate the generalization ability of our pre-training methods, and also the advantages of deep emotion aware vision-language modeling.

Second, comparing models trained with larger datasets without specific consideration of emotion tags and visual-emotion interaction, we observe significant advantages by integrating emotion related texture signals during pre-training. Even with more than five times of image-text pairs, Base-VG and Base-OID still underperform our methods significantly.

Third, comparing our methods, we find using emotion tags only generally performs better than using both emotion tags and other tags as image tag sequence for pre-training, and we suspect the non-emotion tag might introduce some noise for emotion-aware contrastive loss. Comparing EAP-ET-T and EAP-ET-S1, we find group object and attribute tags together as a sequence performs generally better and as discussed in the baseline section, the advantage might come from the easiness of ground emotion with object-attribute pair. In the end, we find EAP-ET-S1 and EAP-ET-S2 comparable, each method performs better on two datasets. More qualitative results are included in supplementary materials.

4.3 Ablation Studies

In this section, we discuss ablation study on fine-tuning tasks. Table. 2 shows results on Emotion6. The first three rows show adaptation variants on EAP-ET-T, one of our strongest pre-trained models. The first row shows results of fine-tune without image caption as inputs, we find it performs much worse than our method with image caption. The 2nd and 3rd rows show accuracy without attribute sequence and without object sequence, and both confirm the necessity of these components during the fine-tuning phase.

Method	Accuracy
EAP-ET-T / caption	57.83
EAP-ET-T / attribute	58.84
EAP-ET-T / object	58.84
EAP-AT-T	54.29
EAP-ET-T	**62.37**
EAP-ET-S1	60.61

Table 2: Ablation study with Emotion6 dataset, first three rows represent adaptation results from EAP-ET-T without caption, attribute or object as inputs.

5 Conclusion and Broad Impact

Conclusion and future works In summary, we introduce an emotion-aware pre-training method to learn representative image embeddings. Our experiments also confirm the effectiveness of the pre-training with emotion-object-attribute alignment target. We believe this work could guide future works on other emotion related pre-training methods, the explainability of the pre-training and adaption algorithm, and more effective ways to collect large datasets for visual emotion understanding.

Broader Impact Our adaptation experiments are conducted in relatively small datasets, as a result, we may face some degree of out-of-distribution problem and model bias induced from the data. Our pretraining models are trained on large-scale weakly supervised data from the Internet, which could encode bias w.r.t. gender, skin color, etc, and it will be an important issue to address in the future. For potential emotion understanding applications in critical areas where human life could be affected, such as education and healthcare, we strongly advise practitioners to only use AI systems to "assist" the decision-making process and ensure fair usage.

References

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[Gan *et al.*, 2020] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020.

[Han *et al.*, 2021] Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Image scene graph generation (sgg) benchmark. *arXiv preprint arXiv:2107.12604*, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*, 2021.

[Kosti *et al.*, 2017] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *CVPR*, 2017.

[Kosti *et al.*, 2020] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. In *TPAMI*, 2020.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, 2017.

[Krylov *et al.*, 2021] Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. Open images v5 text annotation and yet another mask text spotter. In *ACML*. PMLR, 2021.

[Li *et al.*, 2020] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019.

[Machajdik and Hanbury, 2010] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM-MM*, 2010.

[Panda *et al.*, 2018] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K. Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *ECCV*, 2018.

[Peng *et al.*, 2015] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 2015.

[Peng *et al.*, 2016] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. Where do emotions come from? predicting the emotion stimuli map. In *ICIP*, 2016.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Iiya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Su *et al.*, 2020] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *ICLR*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wang *et al.*, 2013] Xiaohui Wang, Jia Jia, Jiaming Yin, and Lianhong Cai. Interpretable aesthetic features for affective image classification. In *ICIP*, 2013.

[Wei *et al.*, 2020] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning visual emotion representations from web data. In *CVPR*, 2020.

[Yang *et al.*, 2018a] Jufeng Yang, Dongyu She, Yu-kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 2018.

[Yang *et al.*, 2018b] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Rosin Paul L., and Liang Wang. Visual sentiment prediction based on automatic discovery of affective regions. In *IEEE Transactions on Multimedia*, 2018.

[You *et al.*, 2015] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015.

[You *et al.*, 2016] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, 2016.

[Zhang *et al.*, 2021] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021.