M5Product: Self-harmonized Contrastive Learning for E-commercial Multi-modal Pretraining

Xiao Dong†, Xunlin Zhan1,2†, Yangxin Wu1, Yunchao Wei3, Michael C. Kampffmeyer4, Xiaoyong Wei5, Minlong Lu6, Yaowei Wang5, Xiaodan Liang1,2*
1Sun Yat-sen University 2Shenzhen Campus of Sun Yat-sen University 3Beijing Jiaotong University 4UiT The Arctic University of Norway 5PengCheng Laboratory 6Alibaba Group.
{dongx55, zhanxlin, wuyx29}@mail2.sysu.edu.cn, {dx.icandoit, wychao1987, xdlang328}@gmail.com, ymlml@zju.edu.cn, michael.c.kampffmeyer@uit.no, cswei@scu.edu.cn, wangyw@pcl.ac.cn

Abstract

Despite the potential of multi-modal pre-training to learn highly discriminative feature representations from complementary data modalities, current progress is being slowed by the lack of large-scale modality-diverse datasets. By leveraging the natural suitability of E-commerce, where different modalities capture complementary semantic information, we contribute a large-scale multi-modal pre-training dataset M5Product. The dataset comprises 5 modalities (image, text, table, video, and audio), covers over 6,000 categories and 5,000 attributes, and is 500⇥larger than the largest publicly available dataset with a similar number of modalities. Furthermore, M5Product contains incomplete modality pairs and noise while also having a long-tailed distribution, resembling most real-world problems. We further propose Self-harmonized ContrAstive LEarning (SCALE), a novel pretraining framework that integrates the different modalities into a unified model through an adaptive feature fusion mechanism, where the importance of each modality is learned directly from the modality embeddings and impacts the inter-modality contrastive learning and masked tasks within a multi-modal transformer model. We evaluate the current multi-modal pre-training state-of-the-art approaches and benchmark their ability to learn from unlabeled data when faced with the large number of modalities in the M5Product dataset. We conduct extensive experiments on four downstream tasks and demonstrate the superiority of our SCALE model, providing insights into the importance of dataset scale and diversity. Dataset and codes are available at 1.

1. Introduction

Self-supervised learning has been driving the rapid development of fields such as computer vision and natural language processing, as well as research on multi-modal representation learning. In particular, it has been shown both from a theoretical [18] and a practical [16, 58] perspective that large scale datasets with diverse modalities can effectively enhance the discrimination of generated features and thus improve the performance in vision-language tasks. However, current advances are severely limited by the lack of such large-scale diverse-modality datasets, with the largest public multi-modal datasets only containing text and image modalities and no category information [41].

Given the prevalence of online shopping in daily life, with its natural occurrence of multi-modal information.

1 https://xiaodongsuper.github.io/M5Product_dataset/
† Equal contribution. * Corresponding Author.
and diverse categories, multi-modal pre-training on E-commerce products has received increasing attention and led the developments of next-generation technology for several downstream tasks (e.g., multi-modal retrieval, multi-modal classification, and clustering). However, even among the present product datasets (e.g., RPC checkout [48], Dress Retrieval [9] and Product1M [53]), the number of categories is insufficient to robustly verify the performance of downstream tasks.

More importantly, the current research community mostly focuses on two modalities (text and image) in both general multi-modal and E-commerce datasets, while ignoring the importance of additional complementary information from structural data as well as video and audio modalities. Tabular data for instance can provide detailed information about properties and characteristics, such as brand, materials, attributes, and scenarios, while audio and video can convey different perspectives, scales, affordances, selling points, characteristics, and use scenarios that are not obvious from images or text alone. The focus on these two modalities is partly due to the lack of datasets with diverse modalities as well as an under-exploration of approaches to balance the modality importance in these settings. In particular, two key challenges are: 1) Modality Interaction: How to learn common representations from unimodal, bimodal, trimodal, and even multi-modal relationships between different modalities using an elegant approach that scales to a large number of modalities; 2) Modality Noise: How to reduce the influence of modality noise (missing and incomplete modalities) during the training process.

To address the problem of insufficient modality diversity and limited scale, while at the same time providing a challenging real-world scenario, we present a very large-scale E-commerce multi-modal product dataset M5Product, which is one of the largest and most diverse multi-modal product dataset to date. Our M5Product dataset contains more than 6 million multi-modal samples from 6,232 categories and has more complex and diverse modalities than existing datasets. This allows M5Product to be used for more comprehensive evaluation of the practical application and generalization abilities of multi-modal pretraining models and can improve the modality fusion performance, facilitating new directions in multimodal research. Figure 1 shows the five modalities (image, caption, video, audio, and specification (table)) of our dataset.

To further address the modality fusion limitations of existing methods as well as handle modality noise, we propose a generic framework that takes five-modality data as inputs, as shown in Figure 2. The framework consists of a simple and efficient multi-modal five stream pre-training model named Self-harmonized ContrAstive LEarning (SCALE) and is evaluated on several downstream tasks and compared with several recent state-of-the-art vision-language models [7, 27, 30, 38, 42, 45, 56]. SCALE increases modality alignment effectiveness by implementing a self-harmonized strategy that adapts the alignment weights between different modalities in the contrastive learning modules and masked tasks to adaptively integrate complementary modality information. In summary, our contributions are as follows:

- We provide the largest five-modality E-commerce dataset M5Product. Through its large scale, diversity, complex real scenarios and number of modalities, M5Product provides a comprehensive environment for evaluating the generalization performance of multi-modal pre-training models.
- Our Self-harmonized Contrastive Learning (SCALE) framework learns adaptive modality interactions, resulting in more effective modality fusion. We compare SCALE to a comprehensive set of baseline methods and demonstrate its superior performance on the M5Product dataset.
- Interesting Observations: 1) In large-scale and complex scenarios, the complementary gain of different modalities increases. Learning modality alignment weights allows our SCALE framework to effectively coordinate complementary information to achieve better performance. 2) For multi-modal pre-training models in the E-commerce domain, dataset scale and diversity are relatively important for the downstream tasks. Given the large-scale and diverse products, our SCALE framework generalizes better than other baselines to downstream tasks.

2. Related Work

Multi-modal pre-training datasets. Most multi-modal pre-training datasets are collected from social websites (e.g., Twitter and Facebook) and are limited to just two modalities collected for specified tasks. These datasets can be divided into four categories according to their modality composition, i.e., audio/text, video/text, image/text, and others. Among these, LJ Speech [19] and SQuAD [25] are classical audio/text datasets and used for voice synthesis and audio Q&A, while most video/text datasets [2, 20, 24, 32, 46, 47, 51, 57] are used for video Q&A. However, these datasets commonly only contain a limited number of samples, limiting their applicability to multi-modal pretraining. Image/text datasets [1, 4, 8, 17, 22, 23, 29, 34, 41, 43, 48, 53], on the other hand, tend to be larger and have been widely used for pretraining multi-modal models. Among these, the CC 3M [41] with more than three million image-text pairs is the most widely used pre-training dataset, and has recently been expanded to CC 12M [5], the largest text-image cross-modal dataset currently. Apart from these, commonly used Image/text datasets for multi-modal retrieval tasks are MS COCO [29], Flickr30K [53], INRIA-Websearch [22] and
Table 1. Comparisons with other widely used multi-modal datasets. “-” means not mentioned. Our M5Product is one of the largest multi-modal datasets compared with existing datasets. Six modalities are separately denoted as: Image (I), Text (T), Video (V), Audio (A), Table (Tab) and 3D Image (3D).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Samples</th>
<th>Categories</th>
<th>Instantes</th>
<th>Modalities</th>
<th>Modal type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD [25]</td>
<td>37,111</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>A/T</td>
</tr>
<tr>
<td>HowTo100M [32]</td>
<td>1,220,000</td>
<td>12</td>
<td>-</td>
<td>2</td>
<td>V/T</td>
</tr>
<tr>
<td>CC 3M [41]</td>
<td>3,300,000</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>I/T</td>
</tr>
<tr>
<td>CC 12M [5]</td>
<td>12,423,374</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>I/T</td>
</tr>
<tr>
<td>CMU-MOSEI [54]</td>
<td>23,500</td>
<td>2</td>
<td>-</td>
<td>3</td>
<td>T/V/A</td>
</tr>
<tr>
<td>XMedia [36]</td>
<td>12,000</td>
<td>20</td>
<td>-</td>
<td>5</td>
<td>I/T/V/A/3D</td>
</tr>
<tr>
<td>RPC checkout [48]</td>
<td>30,000</td>
<td>200</td>
<td>367,955</td>
<td>2</td>
<td>I/T</td>
</tr>
<tr>
<td>Dress Retrieval [9]</td>
<td>20,200</td>
<td>50</td>
<td>~20,200</td>
<td>2</td>
<td>I/T</td>
</tr>
<tr>
<td>Product1M [55]</td>
<td>1,182,083</td>
<td>458</td>
<td>92,200</td>
<td>2</td>
<td>I/T</td>
</tr>
<tr>
<td>MEP-3M [6]</td>
<td>3,012,959</td>
<td>599</td>
<td>-</td>
<td>2</td>
<td>I/T</td>
</tr>
<tr>
<td>M5Product</td>
<td>6,313,067</td>
<td>6,232</td>
<td>-</td>
<td>5</td>
<td>I/T/V/A/Tab</td>
</tr>
</tbody>
</table>

Table 2. The characteristics of different modalities for E-products.

<table>
<thead>
<tr>
<th>Modality</th>
<th>APP</th>
<th>USA</th>
<th>SPEC</th>
<th>SELL</th>
<th>PROD</th>
<th>MATE</th>
<th>CATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Table</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

NUS-WIDE [8] with standard annotations. Other datasets include CMU-MOSEI [54] and XMedia [36], where CMU-MOSEI mainly focuses on the emotional analysis and XMedia is utilized for cross-modal retrieval.

Aside from the abovementioned datasets, there exist several E-commerce datasets. The Dress Retrieval [9], RPC checkout [48] and Product1M [55] are typical E-commerce multi-modal datasets. The Dress Retrieval dataset contains 20,200 samples from 50 clothing categories, RPC checkout offers 30,000 samples of small retail goods on simple backgrounds and Product1M provides 1.18 million samples from 458 cosmetics classes. Compared with these three datasets, our M5Product is not only larger in terms of categories and data scale, but also contains a more diverse set of modalities. A detailed comparison with other multi-modal pre-training datasets is provided in Table 1.

Multi-modal pre-training for E-commerce products.
Several vision-language pre-training models have been explored for visual-text multi-modal learning in recent years. They can coarsely be divided into two categories: 1) Single-stream models whose transformer layer operates collectively on the concatenation of the visual and text inputs, e.g. VL-BERT [42], Image-BERT [37], ViLBERT [30], VisualBERT [27] and UNITER [7]. 2) Dual-stream models whose image and text inputs are not concatenated, such as ViLBERT [30], LXMERT [45], CLIP [38] and DALL-E [39].

Within E-commerce, fashion-based tasks have been addressed in among others FashionBERT [13], MAAF [11], Kaleido-BERT [59], M6 [28] and CAPTURE [55]. All existing studies in the E-commerce scenarios focus solely on the image and text modalities and none of the approaches can utilize more modalities. Besides, all existing methods default to assigning the same contribution to different modalities when modeling multi-modal interactions. More specifically, transformer-based approaches combine high-level features that are extracted from the different inputs via concatenation, where the uni-modal transformers are trained via masked task constraints or via constructing inter-modality losses between different modalities. This restricts the models from effectively prioritizing modalities and tends to limit performance improvements as the number of modalities increases.

Our proposed benchmark fills this gap by exploiting all the diverse modalities of the M5Product dataset and provides a strong baseline for multi-modal pre-training research in the field of E-commerce and beyond.

3. M5Product Dataset

Data Collections. The dataset is crawled from a popular E-commerce website 2, and the front page of each E-commerce product is analyzed to collect the multi-modal information consisting of product images, captions, videos, and specifications (table information) 3. Duplicate data is removed and audio information is extracted from videos via the moviepy 4 tool and saved in mp3 format. For product specifications, we extract 5,679 product properties and 24,398,673 values to construct a table database coarsely labeled by e-commerce merchants. After processing, the dataset contains 6,313,067 samples. Note, being a real-world dataset, our M5Product is, unlike traditional multi-modal datasets, not a complete paired dataset and contains samples with only a subset of modalities as well as long-tailed distributed (Figure 3). We summarize the product characteristics that are relayed by the different modalities in our dataset in Table 2, where APP, USA, SPEC, SELL, PROD, MATE and CATE denote Appearance, Usage, Specification, Selling Point, Production, Material and Category Descriptions, respectively.

Quantitative analysis. 1) Diversity: The dataset consists of more than 6,000 classes covering various and massive amounts of E-commerce products such as clothes, cosmetics, and instruments. Figure 1 illustrates the diversity of the modalities and categories and we further provide a description of the data format and the collection process in Section E of the supplementary materials. Finally, a quantitative analysis of the category and modality distributions can be found in Section F. Note that about 5% of products are unimodal samples e.g. only either contain images, captions,

2 We are authorised by the company to access and obtain the data. We are further authorised to share the dataset and the detailed license is given in Section A of the supplementary material.

3 In this work we focus on core data modalities (image, text, video, audio, and table data) only and do not consider extracted feature representations such as OCR and Motion embeddings that are extracted from core modalities as separate modalities.

4 https://pypi.org/project/moviepy/
Finally, we provide further insights into the composition of the dataset (missing modalities, unimodal data analysis, and how annotations are obtained in Section D). We further provide a smaller split, referred to as subset, which is used to show the difference in performance for a smaller dataset.

Moreover, our dataset contains a large amount of instances, i.e., more than six million samples from the 6,232 coarse categories. These abundant data will benefit several downstream tasks such as self-learning, weakly-supervised learning, multi-modal retrieval, cross-modal generation and fine-grained recognition.

Additional analysis. In the supplementary materials, we provide dataset collection details in Section B and detail how the dataset is split into training and test in Section D and how annotations are obtained in Section C. We further provide a smaller split, referred to as subset, which is used to show the difference in performance for a smaller dataset. Finally, we provide further insights into the composition of the dataset (missing modalities, unimodal data analysis, and data format) in supplementary Section F.

4. Our Methodology

As shown in Figure 2, our SCALE framework consists of a self-harmonized contrastive learning module and a self-supervised multi-modal transformer. In this section, we first provide the architectural design of SCALE in Section 4.1 before describing the five masked tasks that enable the self-supervised learning of SCALE in Section 4.2. Finally, we present the detailed learning process of SCALE and detail how multi-modal alignment is achieved in Section 4.3.

4.1. Architectural Design of SCALE

As depicted in Figure 2, SCALE is a typical single-stream transformer architecture. In the bottom part, the Image/Text/Table/Video/Audio embedding layers and transformers aim to extract modality features and generate token features. Specifically, the text and table encoders are standard transformers to encode the caption and table information of products, respectively. The image encoder instead takes proposals extracted by bottom-up-attention [3] as inputs, while ordinal frames sampled from the video are fed into the video encoder. For the audio encoder, SCALE extracts MFCC [33] features from audio. After being processed by the separate modality encoders, the token features of different modalities are concatenated and fed into a Joint Co-Transformer (JCT) module to capture the token relationships between different modalities.

Missing Modalities. Zero imputation of missing modalities is leveraged to utilize all available data when training...
SCALE. We provide experimental evidence that SCALE benefits from the incomplete samples in Section I of the supplementary material.

4.2. SCALE by Masked Multi-Modal Tasks

Similar to previous works, we utilize several pretext tasks (PRE) to facilitate self-supervised learning of SCALE in the Joint Co-Transformer module. For modality-wise feature learning from the image and text modalities, we adopt the Masked Region Prediction task (MRP) and the Masked Language Modeling task (MLM), respectively, after the JCT. Utilizing the characteristics of the table, video, and audio modalities, we further propose a Mask Entity Modeling task (MEM), Mask Frame Prediction task (MFP), and Mask Audio Modeling task (MAM) following a similar strategy of predicting masked tokens. In all masked tasks, the ground-truth labels are the features of masked areas. For all masking tasks, 15% of the inputs are masked out and the remaining inputs are used to reconstruct the masked information. Please note that unlike in the MLM task, where 15% of individual words are masked, 15% of the entities (properties, brand names, etc.) are entirely masked out for the MEM task. This drives our model to learn better table representations to recover the masked inputs, which is illustrated in Section 5.3. The loss function of the $i$th modality is defined as:

$$L_M(\theta) = -\mathbb{E}_{t_{msk} \sim t} \log p_{\theta}(t_{msk} \mid t_{msk}, M_{-i}),$$

where $t_{msk}$ denotes the unmasked tokens surrounding the masked token $t_{msk}$, $\theta$ represents the network parameters, and $M_i$ and $M_{-i}$ are the $i$th modality and the remaining modalities, respectively.

4.3. Self-harmonized Inter-Modality Contrastive Learning

Self-harmonized Inter-Modality Contrastive Learning (SIMCL) is at the core of our proposed SCALE framework. It aims to facilitate the semantic alignment between different modalities via a self-harmonized strategy for adaptive Inter-Modality Contrastive Learning (IMCL). For a minibatch of modality samples $D \in \mathbb{R}^{B \times M \times F}$, where $B$, $M$ and $F$ denote the batch size, number of modalities, and embedding dimension, respectively, we first construct the contrastive loss between each modality pair:

$$\mathcal{L}_{CL}(d_{i}^{(0)}, d_{i}^{(1)}) = -\log \frac{\exp \left(\frac{\text{sim}(f_{i}^{(0)}, f_{i}^{(1)})}{\tau}\right)}{\sum_{m=0}^{N} \sum_{k \neq i} \exp \left(\frac{\text{sim}(f_{i}^{(m)}, f_{k}^{(i-m)})}{\tau}\right)},$$

where $\text{sim}$ is the cosine similarity, $\tau$ is the temperature parameter and $1_{[k \neq i]}$ is a binary indicator function, and $1=1$ for $k \neq i$ and 0 otherwise.

In most prior work, only two modalities are considered and Eq. 2 can be used. However, when considering trimodal data or data with even more than three modalities, it is not suitable to directly fit the loss function as it does not account for the difference in complementary information that different modalities contribute. To solve this problem, we define a simple but effective self-harmonized method to model the complementary process of the inter-modal relationships. We introduce a modality alignment score matrix, to encode the relationships among the inter-modal losses $\mathcal{L}_{CL}$ and the intra-modal losses $\mathcal{L}_M$. The alignment score matrix $S$ for each data sample is initialized by a zero matrix and updated as free model parameters. To obtain modality importance scores for each modality combination, we apply the softmax function to $S$. Finally, the importance scores are multiplied to generate the modality alignment score $S$ as $S = S \cdot \text{softmax}(S)$. The learning process is shown in Figure 4 and illustrates that SIMCL takes full advantage of the inter-modal relationships. Given the modality alignment score $S$, the triangular part $S_{\triangle}$ is selected to weight the inter-modal losses $\mathcal{L}_{CL}$ and the diagonal part $S_{\delta}$ is utilized to constrain the intra-modal losses $\mathcal{L}_M$, resulting in the weighted loss:

$$\mathcal{L}_{total} = \sum_{S_{i,j}} S_{i,j} \logit_{i,j} + \sum_{S_i} \mathcal{L}_M(S_i \logit_i)$$

where logit is the loss logit.

5. Experiments

Implementation Details. We use BERT [10] to initialize the text transformer of our proposed SCALE framework,
5 region features of selected 10 to 36 bounding boxes with pre-trained on the Visual Genome dataset \cite{krishna2017visual}. The maximum sequence length for the captions and tables are set to 36 and 64, respectively. Using the same setting as in \cite{li2019vlbert}, we utilize Faster R-CNN \cite{ren2015faster} with a backbone ResNet101 \cite{he2016deep} to extract region features of selected 10 to 36 bounding boxes with high-class detection probability for each image. We train SCALE with a total batch size of 64 for 5 epochs using the Adam optimizer \cite{kingma2014adam} with a warm-up learning rate of 1e-4. Additional implementation details of our model are provided in Section G of the supplementary material.

### Baselines
We compare SCALE to the following eight alternative pre-training methods that utilize image and text modalities as well as combinations of both: Bert \cite{devlin2018bert} (Text-based), Image-based, VIbert \cite{li2019vlbert}, CLIP \cite{radford2021learning}, VisualBERT \cite{lei2019visualbert}, UniTER \cite{li2021unicoder}, CAPTURE \cite{yu2021capturing}. Image-based and BERT \cite{devlin2018bert} are 12-layer transformers based on the MLM (Mask Language Modeling) or MRP (Mask Region Prediction) task using image or text modality, providing single-modal baselines for the product retrieval, classification, and clustering tasks. To ensure a fair comparison, the same hidden size of 768 is chosen for all baselines.

### Evaluation
We consider the following four downstream tasks to evaluate the learned representations: 1) Multi-modal retrieval: This task aims to find the most relevant product (i.e. color, model, shape, and style) are considered same category; 2) Fine-grained multi-modal retrieval: This task aims to find the most relevant modal retrieval: This task aims to find the most relevant modal retrieval: This task aims to find the most relevant

### Table 3
The (pretrain/finetune) performance gains from sequentially adding more modalities using SCALE on the subset (top) and the whole dataset (bottom). The retrieval performances are based on the features extracted from pretrain and finetune stages.

<table>
<thead>
<tr>
<th>Modality Combinations</th>
<th>Accuracy</th>
<th>mAP@1</th>
<th>mAP@5</th>
<th>mAP@10</th>
<th>Prec@1</th>
<th>Prec@5</th>
<th>Prec@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>I+Tab</td>
<td>62.00</td>
<td>44.53</td>
<td>45.97</td>
<td>49.62</td>
<td>48.28</td>
<td>30.33</td>
<td>23.65</td>
</tr>
<tr>
<td>I+V</td>
<td>34.57</td>
<td>20.57</td>
<td>53.26</td>
<td>26.78</td>
<td>24.61</td>
<td>26.64</td>
<td>14.71</td>
</tr>
<tr>
<td>I+W</td>
<td>38.88</td>
<td>15.73</td>
<td>15.64</td>
<td>20.85</td>
<td>20.79</td>
<td>15.73</td>
<td>11.61</td>
</tr>
<tr>
<td>I+T</td>
<td>79.58</td>
<td>67.02</td>
<td>62.20</td>
<td>60.95</td>
<td>68.64</td>
<td>62.21</td>
<td>46.29</td>
</tr>
<tr>
<td>I+T+V</td>
<td>80.34</td>
<td>67.73</td>
<td>63.05</td>
<td>70.29</td>
<td>67.37</td>
<td>68.95</td>
<td>46.45</td>
</tr>
<tr>
<td>I+T+W</td>
<td>73.09</td>
<td>59.81</td>
<td>71.51</td>
<td>64.13</td>
<td>75.21</td>
<td>62.18</td>
<td>47.58</td>
</tr>
<tr>
<td>I+T+Tab</td>
<td>63.09</td>
<td>45.94</td>
<td>47.33</td>
<td>51.32</td>
<td>53.33</td>
<td>49.78</td>
<td>31.69</td>
</tr>
<tr>
<td>I+T+Tab+V</td>
<td>84.31</td>
<td>67.97</td>
<td>68.30</td>
<td>70.34</td>
<td>72.67</td>
<td>69.38</td>
<td>46.85</td>
</tr>
<tr>
<td>I+T+Tab+A</td>
<td>63.54</td>
<td>47.24</td>
<td>48.24</td>
<td>52.07</td>
<td>53.89</td>
<td>50.41</td>
<td>32.19</td>
</tr>
<tr>
<td>I+T+Tab+V+A</td>
<td>80.36</td>
<td>68.80</td>
<td>66.43</td>
<td>70.84</td>
<td>71.12</td>
<td>69.71</td>
<td>47.44</td>
</tr>
<tr>
<td>I+T+Tab+A+V</td>
<td>84.33</td>
<td>70.23</td>
<td>68.97</td>
<td>72.59</td>
<td>73.07</td>
<td>70.94</td>
<td>47.58</td>
</tr>
<tr>
<td>I+T+Tab+A+V+I</td>
<td>59.50</td>
<td>70.62</td>
<td>69.25</td>
<td>73.02</td>
<td>74.08</td>
<td>71.50</td>
<td>58.76</td>
</tr>
<tr>
<td>I+T+Tab+A+V+I+V</td>
<td>59.81</td>
<td>83.68</td>
<td>92.01</td>
<td>70.62</td>
<td>69.25</td>
<td>73.02</td>
<td>58.76</td>
</tr>
</tbody>
</table>

### Table 4
The performance of our model SCALE under different modality combinations on the coarse- and fine-grained multi-modal retrieval and classification tasks. In the following, I, T, Tab, V and A denote image, text, table, video and audio modalities, respectively.

<table>
<thead>
<tr>
<th>Modality Combinations</th>
<th>Accuracy</th>
<th>mAP@1</th>
<th>mAP@5</th>
<th>mAP@10</th>
<th>Prec@1</th>
<th>Prec@5</th>
<th>Prec@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>I+Tab</td>
<td>62.00</td>
<td>44.53</td>
<td>45.97</td>
<td>49.62</td>
<td>48.28</td>
<td>30.33</td>
<td>23.65</td>
</tr>
<tr>
<td>I+V</td>
<td>34.57</td>
<td>20.57</td>
<td>53.26</td>
<td>26.78</td>
<td>24.61</td>
<td>26.64</td>
<td>14.71</td>
</tr>
<tr>
<td>I+W</td>
<td>38.88</td>
<td>15.73</td>
<td>15.64</td>
<td>20.85</td>
<td>20.79</td>
<td>15.73</td>
<td>11.61</td>
</tr>
<tr>
<td>I+T</td>
<td>79.58</td>
<td>67.02</td>
<td>62.20</td>
<td>60.95</td>
<td>68.64</td>
<td>62.21</td>
<td>46.29</td>
</tr>
<tr>
<td>I+T+V</td>
<td>80.34</td>
<td>67.73</td>
<td>63.05</td>
<td>70.29</td>
<td>67.37</td>
<td>68.95</td>
<td>46.45</td>
</tr>
<tr>
<td>I+T+W</td>
<td>73.09</td>
<td>59.81</td>
<td>71.51</td>
<td>64.13</td>
<td>75.21</td>
<td>62.18</td>
<td>47.58</td>
</tr>
<tr>
<td>I+T+Tab</td>
<td>63.09</td>
<td>45.94</td>
<td>47.33</td>
<td>51.32</td>
<td>53.33</td>
<td>49.78</td>
<td>31.69</td>
</tr>
<tr>
<td>I+T+Tab+V</td>
<td>84.31</td>
<td>67.97</td>
<td>68.30</td>
<td>70.34</td>
<td>72.67</td>
<td>69.38</td>
<td>46.85</td>
</tr>
<tr>
<td>I+T+Tab+A</td>
<td>63.54</td>
<td>47.24</td>
<td>48.24</td>
<td>52.07</td>
<td>53.89</td>
<td>50.41</td>
<td>32.19</td>
</tr>
<tr>
<td>I+T+Tab+V+A</td>
<td>80.36</td>
<td>68.80</td>
<td>66.43</td>
<td>70.84</td>
<td>71.12</td>
<td>69.71</td>
<td>47.44</td>
</tr>
<tr>
<td>I+T+Tab+A+V</td>
<td>84.33</td>
<td>70.23</td>
<td>68.97</td>
<td>72.59</td>
<td>73.07</td>
<td>70.94</td>
<td>47.58</td>
</tr>
<tr>
<td>I+T+Tab+A+V+I</td>
<td>59.50</td>
<td>70.62</td>
<td>69.25</td>
<td>73.02</td>
<td>74.08</td>
<td>71.50</td>
<td>58.76</td>
</tr>
<tr>
<td>I+T+Tab+A+V+I+V</td>
<td>59.81</td>
<td>83.68</td>
<td>92.01</td>
<td>70.62</td>
<td>69.25</td>
<td>73.02</td>
<td>58.76</td>
</tr>
</tbody>
</table>

https://github.com/airsplay/py-bottom-up-attention
a match 6; 3) Multi-modal classification: Product category classification given the multi-modal features extracted from the joint co-transformer of SCALE using a linear classifier; and 4) Multi-modal clustering: Product category clustering using k-Means clustering and the same features as in the classification setting. For product retrieval, we adopt the widely used metrics mean Average Precision (mAP) and Precision (Prec) [14, 31, 49] to evaluate the retrieval accuracy on the two retrieval tasks. For product classification and clustering, all methods are evaluated using the Classification Precision (Classification accuracy), Normalized Mutual Information (NMI) [52] and Purity metrics. In all experiments, models are trained on the training split. The pretrained model is then applied to extract the modality features of the gallery and test splits for the product retrieval and clustering tasks. For the classification task, we finetune the pre-trained model on the classification subset containing 1,805 categories/classes and utilize the finetuned model to extract the features of the classification test set.

5.1. Modality Diversity

To examine the performance of our proposed SCALE framework and to verify the benefits of diverse modalities and dataset scale, we train SCALE with an increasing number of modalities and observe the variations in classification and multi-modal retrieval performance both for the whole MSPProduct dataset and the subset. More specifically, fused features are extracted from the joint co-transformer (JCT) of our SCALE after finetuning for the classification task and after pre-training and finetuning for the (coarse) multi-modal retrieval task. Results in Table 3 show that performance increases across all settings as modalities are added, illustrating the benefit of complementary modality information to learning multi-modality feature representations. It can also be observed that modality gains are larger on the whole dataset, supporting Interesting Observation 1.

We further provide results for an extensive set of modality combinations to verify SCALEs effectiveness in leveraging the diverse modalities of our MSPProduct dataset. Table 4 provides results for the coarse- and fine-grained multi-modal retrieval tasks as well as the classification task after finetuning the model. As in the previous experiment, noticeable improvements can be observed as additional modalities are added. In particular, the addition of the text modality leads to high modality gains, verifying the benefits of including more diverse modalities that can capture different views of the same product. Interestingly, performance on the coarse-grained retrieval task is significantly worse than on the fine-grained retrieval task in most cases, indicating the complexity of the MSPProduct dataset and the diversity of the products in each category.

6 A more thorough definition of the term same products and how instance-level labels are obtained is provided in the supplementary.
5.3. Ablation Studies and Visualization

To explore how SIMCL influences IMCL and the Pretext tasks, we conduct several ablation studies. Table 6 illustrates that improvements of approximately 2% are obtained in the classification task and more than 2% for the coarse-grained retrieval task when including both, highlighting the importance of both the Pretext tasks and the effective modality fusion of SIMCL. We further analyze the effect of the MEM pretext task for the table modality and show the benefit of masking out complete entities over masking out individual tokens (MLM) in Table 7. This benefit can be attributed to the fact that MEM ensures that SCALE learns representations that encode the semantic information of complete entities. Finally, we evaluate the performance of modelling the text and the table modalities using individual modality encoders and compare SCALE’s retrieval performance to a baseline where text and table information is concatenated and fed to a single transformer, resembling the process of BERT [10]. By modelling both modalities individually, results in Table 8 illustrate that more information can be preserved and we hypothesize that using a single transformer leads to a loss in table modality information for the benefit of the more expressive text modality.

Figure 6 shows t-SNE visualizations of the extracted features for the JCT module of our SCALE model and the alternative approach ViLBert [30] on the MSPProduct dataset. SCALE not only better distinguishes different classes but also improves class compactness compared to the ViLBert model. Further, the attention attribution for different modalities are shown in Figure 7 and verify that the visual features generated by SCALE are object-oriented and semantically interpretable.

6. Limitations and future work

The experimental evaluation showed that SCALE is able to learn efficient representations from a large number of modalities for retrieval, classification, and clustering. However, more evaluation of the generative capabilities of the models representations is lacking and tasks such as image and caption generation could be promising directions to explore. We further provide some of SCALE’s failure cases in supplementary Section J.

Potential negative societal impact. As a result of the strict ethical considerations used in the data collection process, where among others personally identifiable information has been removed, MSPProduct does not pose any ethical risks.

7. Conclusion and Discussion

To facilitate multi-modal pre-training, we present the MSPProduct dataset, which is the largest available multi-modal E-commerce product dataset, consisting of five core modalities (image, text, table, video, and audio). To further promote multi-modal research in retail and increase seller and buyer engagement, we also propose the novel SCALE multi-modal pre-training framework. By utilizing Self-harmonized Inter-Modality Contrastive Learning (SIMCL), SCALE is able to model and exploit modality relationships effectively and outperforms previous approaches on the MSPProduct multi-modal retrieval, classification, and clustering tasks. We believe that both the dataset and the proposed framework work will inspire research on scaling multi-modal pre-training beyond the commonly used image and text modalities.

8. Acknowledgement

This work was supported in part by Shenzhen Fundamental Research Program (Project No. RCYX20200714114642083, No. JCYJ20190807154211365), National Key RD Program of China (2021ZD0112100), and CAAI-Huawei MindSpore Open Fund. We thank MindSpore for the partial support of this work, which is a new deep learning computing framework, and supported by Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology, Guangzhou 510006, China.

https://www.mindspore.cn/
References


