Multimodal Pre-Training with Self-Distillation for Product Understanding in E-Commerce

Shilei Liu*
Alibaba Group
Hangzhou, China
liushilei.lsl@alibaba-inc.com

Lin Li*
Alibaba Group
Hangzhou, China
guoba.ll@alibaba-inc.com

Jun Song
Alibaba Group
Hangzhou, China
jsong.sj@alibaba-inc.com

Yonghua Yang
Alibaba Group
Hangzhou, China
huazai.yyh@alibaba-inc.com

Xiaoyi Zeng
Alibaba Group
Hangzhou, China
yuanhan@taobao.com

ABSTRACT

Product understanding refers to a series of product-centric tasks, such as classification, alignment and attribute values prediction, which requires fine-grained fusion of various modalities of products. Excellent product modeling ability will enhance the user experience and benefit search and recommendation systems. In this paper, we propose MBSD, a pre-trained vision-and-language model which can integrate the heterogeneous information of product in a single stream BERT-style architecture. Compared with current approaches, MBSD uses a lightweight convolutional neural network instead of a heavy feature extractor for image encoding, which has lower latency. Besides, we cleverly utilize user behavior data to design a two-stage pre-training task to understand products from different perspectives. In addition, there is an underlying imbalanced problem in multimodal pre-training, which will impairs downstream tasks. To this end, we propose a novel self-distillation strategy to transfer the knowledge in dominated modality to weaker modality, so that each modality can be fully tapped during pre-training. Experimental results on several product understanding tasks demonstrate that the performance of MBSD outperforms the competitive baselines.

CCS CONCEPTS

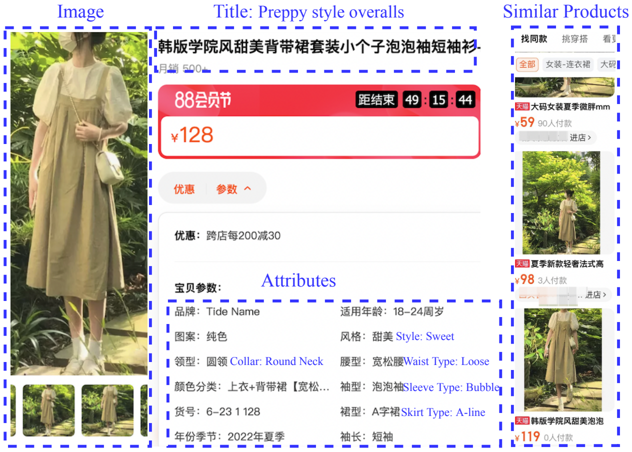
• **Information systems** → *Retrieval tasks and goals*; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

vision-and-language pre-training, product understanding, modality imbalance

ACM Reference Format:
Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. 2023. Multimodal Pre-Training with Self-Distillation for Product Understanding in E-Commerce. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570423

1 INTRODUCTION



韩版学院风甜美背带裙套装小个子泡泡袖短袖衫
月销 500+

优惠：跨店每200减30

宝贝参数：

品牌：Tide Name
图案：纯色
领型：圆领 Collar: Round Neck
颜色分类：上衣+背带裙【宽松...
货号：6-23 1 128
年份季节：2022年夏季
适用年龄：18-24周岁
风格：甜美 Style: Sweet
腰型：宽松腰Waist Type: Loose
袖型：泡泡袖Sleeve Type: Bubble
裙型：A字裙Skirt Type: A-line
袖长：短袖

Figure 1: An e-commerce product in *Taobao*.

Online shopping has been integrated into human life due to the rapid growth of e-commerce in recent years. As shown in Figure 1, in addition to the title and image, some other information such as product attributes and similar style products, is also displayed on the detail page. Search and recommendation engines rely on them to better serve users. However, such information may be incomplete or inaccurate and needs to be supplemented by the e-commerce platforms, which challenges the ability of product modeling. Only a powerful model capable of fusing multimodal information can provide insight into the products.

Transformer [27], first designed for machine translation, then spread to the entire natural language processing community [6, 15] and achieved success in the area of computer vision [8, 18]. Recently, many works try to pre-train the transformer on large-scale text

[*] Contributed equally to this research. Shilei Liu is corresponding author.

WSDM '23, February 27-March 3, 2023, Singapore, Singapore
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570423

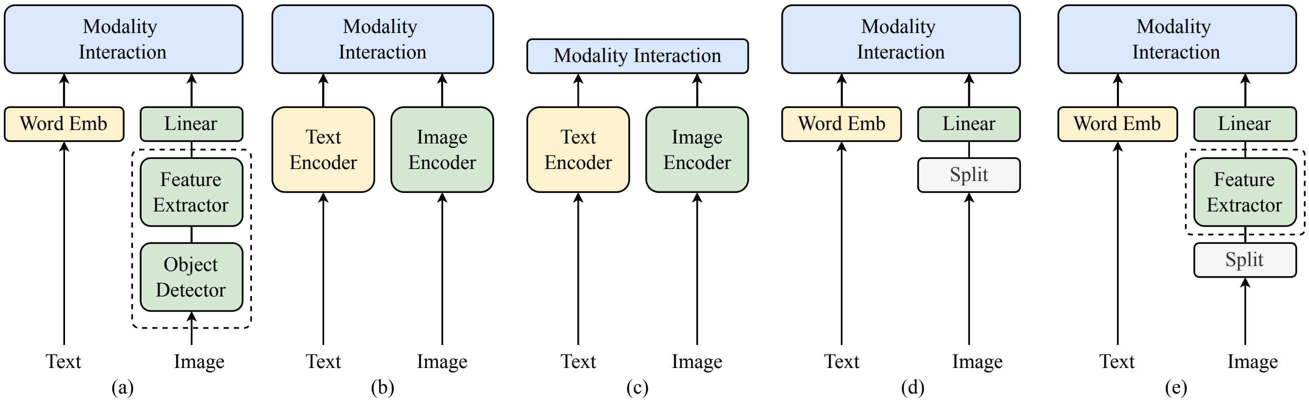


Figure 2: Five categories of vision-and-language models. The parameters in the dotted box are frozen during training. The height of each rectangle indicates its computational cost.

and image pairs, which lead to impressive performance gains in a wide variety of vision-and-language downstream tasks such as cross-modal retrieval [28], visual question answering [2] and image captioning [1]. However, these studies are centered on the general domain, which are sub-optimal for product understanding tasks. On the one hand, some popular works [16, 19, 24, 26] rely on RoI-based (i.e., Region of Interest) object detectors, which do not work well in the e-commerce domain since relatively-rare RoIs can be detected from fashion images and the extracted objects have high overlapping with each other [10, 17]; On the other hand, some of the latest models [23, 34, 40] have carefully designed model structures and losses for specific public benchmarks such as image-text retrieval, but these methods are not suitable for the product understanding, because the information about the item itself is indivisible. So these designs are rarely bring performance gains. Although some vision-and-language models in the e-commerce field have recently emerged [10, 41], they use heavy external feature extractors to encode images and cannot achieve a trade-off between performance and speed.

Besides, there are rich user behavior data in e-commerce field, such as searches, clicks, and purchases, which are not available in general fields but contain useful information. If they are integrated into the pre-training of e-commerce multimodal models, it will help to improve the performance in the downstream tasks.

In addition, there is an imbalance phenomenon in multimodal learning tasks [20], which refers to the potential of multimodal information is not fully exploited even when the multimodal model outperforms its uni-modal counterpart [21]. This violates the intention of improving model performance through fusing features from multiple modalities, and will damage downstream tasks if not handled properly.

In this paper, to address the above problems, we propose a **M**ultimodal **B**ERT with **S**elf-**D**istillation pre-training for product understanding named **MBSD**. We choose BERT as the backbone and use a lightweight convolution network to split the image into a sequence of patches to fit the BERT. Two mask-based losses are employed to learn joint representations of vision and language on large-scale products. Besides, we perform two strategies to construct an additional dataset from click-through data and continue pre-train MBSD on it. In addition, we propose a self-distillation training method to transfer the knowledge in dominated modality to weaker modality, which can benefits the downstream tasks.

We evaluate MBSD with several product understanding tasks on four real-world datasets, and the results show that our MBSD significantly outperforms other pioneer approaches.

In summary, our contributions are as follows:

- Propose a simple end-to-end single-stream architecture which can fuse the features from textual and visual information of products. Compared to some complex designs, this approach leads to significant runtime and parameter efficiency.
- Take advantage of user behavior data in e-commerce web search to continue pre-train the model to further enhance the performance.
- Propose a self-distillation pre-training strategy to exploit the knowledge contained in weaker modality and alleviate modality imbalance phenomenon.
- Achieve state-of-the-art performance on various product understanding tasks.

Part of the code and a small amount of pre-training and fine-tuning data are available at https://github.com/NPEL-ll/mbsd.

2 BACKGROUND

2.1 Product Understanding

Product understanding refers to a series of downstream tasks based on products, such as classification [39], clustering [7], attribute values prediction [3, 35] and similar product recognition. The above tasks play a crucial role in the e-commerce systems, which help customers to make purchasing decisions and facilitate retailers on many applications, such as product search, product recommendation and product representation.

It should be noted that different from the mainstream multimodal tasks such as cross-model retrieval (search image based on text and vice versa) and image captioning (generate text based on image), all modalities are regarded as inseparable in product understanding,

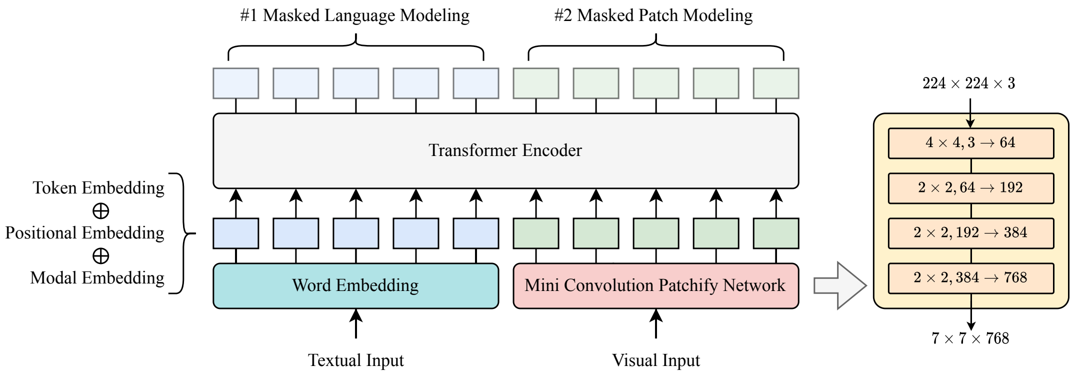

#1 Masked Language Modeling
#2 Masked Patch Modeling
Transformer Encoder
Token Embedding
⊕
Positional Embedding
⊕
Modal Embedding
Word Embedding
Mini Convolution Patchify Network
Textual Input
Visual Input
$224 \times 224 \times 3$
$4 \times 4, 3 \rightarrow 64$
$2 \times 2, 64 \rightarrow 192$
$2 \times 2, 192 \rightarrow 384$
$2 \times 2, 384 \rightarrow 768$
$7 \times 7 \times 768$

Figure 3: Model structure and preliminary pre-training paradigm. Layer normalization and ReLU are inserted between every two adjacent convolutional layers.

and models should pay more attention to the integration rather than alignment between heterogeneous modalities.

2.2 Vision-and-Language Models

We propose a taxonomy of vision-and-language models from the perspective of modality embedding and interaction.

The RoI-based models such as LXMERT [26], ViLBERT [19], VL-BERT [24], UNITER [5] and M5Product [7] belong to Figure 2(a). They use the object detector such as Mask R-CNN [12] to extract several regions of interest as object-level visual token. The embedding of the visual token is obtained by the feature extractor. This kind of methods has severe performance overhead and cannot achieve end-to-end training.

FLAVA [23], METER [4] and CAPTURE [38] and belong to Figure 2(b). They use separate but equally expensive embedders for each modality and another heavy encoder for modality fusion. This class of methods achieves better performance in cross-modal retrieval tasks, but has a larger number of parameters and slower speed.

CLIP [22] belongs to Figure 2(c), it performs shallow interaction rather than deep fusion between the pooled text vector and image vector. CLIP is good at the zero-shot classification of images.

There are also some works such as ViLT [14] use shallow text and image embedders and deep fusion layer like Figure 2(d). This design inherently leads to significant speed and parameter efficiency, but the performance is not the best at present.

2.3 E-Commerce Domain Pre-Training

FashionBERT [10] was the first published work in the e-commerce fashion domain, which cut each image into 16 patches with the same pixels, and use ResNet-50 to encode each patch independently. The embedding of each patch constitutes the BERT input on the image side. M6 [17] uses the same patchify method with FashionBERT. Kaleido-BERT [41] is an improvement of FashionBERT which utilize kaleidoscope patchify strategy to split the image into 55 patches with different pixels.

We argue that the above approaches (can be summarized as Figure 2(e)) restrict the power of the pre-trained representation learning and damage inference speed. First of all, ResNet is trained on general domain image gallery that is not inconsistent with the distribution of e-commerce images. This inconsistency may cause errors in the extracted features, increasing the model burden during training. Besides, for a single image, FashionBERT and Kaleido-BERT will call ResNet 16 times and 55 times respectively, which cannot achieve accuracy/FLOPs trade-offs, and brings challenges for large-scale deployment. As a consequence, an end-to-end solution for downstream tasks is urgently required. In addition, the rich user behavior data in e-commerce platforms are also not utilized in the above model.

2.4 Balanced Multimodal Learning

Multimodal learning aims to achieve performance improvement by integrating different modal features. Many studies have found that multimodal learning cannot bring sufficient performance gains due to modal discrepancy and modal imbalances [9, 21, 25, 30].

Wang et al. [30] found that different modalities converge at different rates, and proposed Gradient-Blending for the late-fusion multimodal network, which computes an optimal blending of modalities based on their overfitting behaviors. Du et al. [9] argue that the imbalance of modalities lead to the insufficient feature learning of the weaker modality, and trained the multimodal model by distilling the trained uni-modal features to the multimodal networks. Further, Peng et al. [21] dynamically adjust the gradient to alleviate the insufficient learning of weak modalities by monitoring the contribution of different modalities to the learning objective.

For late-fusion multimodal models, gradient-based methods can lead to improved performance. However, it's not suitable for early-fusion models with fully integrated features, such as single-stream BERT-style models. In this work, to learn the weaker modality sufficiently, we design a self-distillation training strategy with an adaptive loss weight, which can automatically teach weak modality from well-trained strong modality.

3 METHODOLOGY

3.1 Overview

The goal of this work is to learn a pre-trained vision-and-language model which can be applied to a range of product understanding tasks with a simple and elegant architecture. As shown in Figure

3, the inputs of MBSD are the linguistic and visual information of items, followed by a standard BERT structure. A mini convolutional neural network will split the image to a sequence of patches. The pre-training consists of two stages. In the first stage, we preform masked language modeling and masked patch modeling to learn the primary semantic representation of products. In the second stage, user behavior data are introduced to help to learn from a holistic perspective. Besides, we propose a self-distillation training strategy to alleviate the modal imbalance problem during the pre-training, which benefits downstream tasks.

3.2 Text Representation

The title of an product will be wrapped with two special symbols [CLS] and [SEP] to form the textual information T. T is then tokenized into a sequence of discrete tokens and get the distributed representation by a lookup table $\mathbf{M} \in \mathbb{R}^{|V| \times d}$, where $|V|$ represents the vocabulary size and d is the hidden size. Finally, the sum of the word embeddings, position embeddings and modality embeddings is regarded as the linguistic inputs of BERT.

3.3 Image Representation

BERT is first designed for natural language processing, so some additional operations are required to convert image into a sequence of d-dimensional visual features. Some works in the e-commerce field uses the pre-extracted image features as the input of the image side [10, 17, 41], but this will inevitably bring error transmission. We try to extract the primary features of the image end-to-end.

Considering that the heavy backbone network will bring large FLOPS, we use a lightweight network to rasterise 2D patches. The architecture of our patchify network is shown in the right part of Figure 3, which contains several cascaded non-overlapping convolution stems to quickly downsample a 224×224 input image I to 7×7, and reshape to 49 patches.

It should be noted that we do not use linear projection in ViT [8] to partition image, which will lead to the training instability [4, 32]. The introduction of convolution does not significantly increase the runtime. The length of the image sequence produced by our method is equivalent to linear projection with a patch size $p = 32$.

3.4 Preliminary Pre-Training Task

In the first stage, we only use mask-based pre-training tasks to learn atomic (terms or image patches) semantic of the product.

3.4.1 Masked Language Modeling. For the textual input, the data preprocessor chooses 15% of the tokens for prediction following BERT. When masking indices set $\mathcal{M}_T$ are determined, we decompose masking words into 80% [MASK] token, 10% random words and 10% unchanged. This task requires model to predict the tokens in the masked positions based on the surrounding tokens and image.

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|\mathcal{M}_T|} \sum_{i \in \mathcal{M}_T} \log \left(T_i | T_{\{i \notin \mathcal{M}_T\}}, I \right) \tag{1}$$

3.4.2 Masked Patch Modeling. BERT is pre-trained on text corpus and does not have the ability to perceive images. Similar to MLM, we add an masked patch modeling task to enhance the visual modeling ability. Specifically, 50% of patch features output by patchify network are replaced with a trainable *mask* vector, and the model needs to predict raw pixels of the masked patch by regression:

$$\mathcal{L}_{\text{MPM}} = \frac{1}{|\mathcal{M}_I|} \sum_{i \in \mathcal{M}_I} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_1 \tag{2}$$

where $\mathbf{y}_i \in \mathbb{R}^{p \cdot p \cdot 3}$ are the raw RGB values of i-th patch, and $\hat{\mathbf{y}}_i$ are the prediction RGB values obtained by an affine transformation (from d-dimension to $3p^2$-dimension) of the i-th patch representation generated by MBSD. The information density of image is much lower than that of text, and more patches need to be masked for a better performance [11, 33].

Mask-based pre-training tasks allow pre-trained models to understand product terms or image patches well, but lack understanding of product overall information. Taking advantage of the rich user behavior data in the e-commerce, we design a pre-training task to learn products holistically.

3.5 Data Construction for Modality Balance Pre-Training

User behaviors lead to a large amount of data related to products such as search queries and product comments, and these product-related data usually have clear semantics, and many researchers have used them to construct training data in e-commerce tasks [36].

In search logs, the search query can be regarded as a partly equivalent description of the product clicked or purchased under the query. Therefore, we take judging whether the query and the product are matched as a pre-training task, which allows the model to understand the product as a whole.

The instance we constructed is in the form of (*matched query*, *product*, *true*) and (*unmatched query*, *product*, *false*). We only regard the product purchased with queries by multiple people as a positive instance since the click-behavior data is easily affected by many factors such as price, attractive images, product ranking positions, and tends to be noisy.

For the construction of negative instances, the product non-clicked under the search query cannot be used as negative instances, as the lack of behavior between the query and the show product does not mean that they do not match. Besides, the queries that are randomly selected for the product are too easy to distinguish and useless for model learning. Inspired by Xiao et al. [31], we construct hard negative instances in the following two ways.

Random select queries from the same category. Queries from the same category are all demand for the same kind of products, but different in purpose, description and focus. It is very suitable to sample hard negative instances from these similar but different queries. Specifically, we collect search queries corresponding to the purchased products of a specific category as the negative instance candidates, and randomly select the queries that do not have click behaviors with the product as hard negative instance.

Query rewriting. Query rewriting is a similar approach to produce hard negative instances. First, we get the type of each term in matched query through an e-commerce NER system, and generate negative instances by randomly replacing a term in the matched query which has the same type. For example, given a matched query *long-sleeved dress*, a rewritten unmatched query can be *short-sleeved dress* or *sleeveless dress*, with the sleeve type in the query replaced.

Some additional checks will ensure that there is no click behavior between this rewritten query and product.

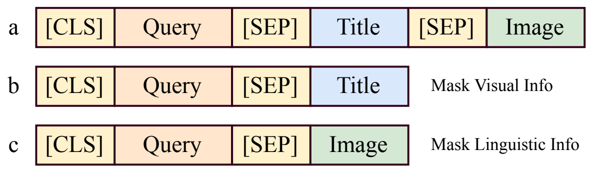


Figure 4: Three input forms for each query-item sample.

3.6 Modality Balance Pre-Training

As illustrated in Figure 4 (a), given i-th training sample, the query will be concatenated with the multimodal product information, and feed into MBSD. A linear layer and a sigmoid function will be applied to the representation of `[CLS]` token (denoted $\mathbf{h}_i \in \mathbb{R}^d$) to obtain the probability distribution $p_i \in (0,1)$ of whether the query matches the item. Binary cross-entropy is used to optimize the model parameters:

$$\mathcal{L}_{\text{QI}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i) \tag{3}$$

where N represents the number of samples of a mini batch, and $y_i \in \{0,1\}$ is the label.

However, both query and title are linguistic information, which have less discrepancy, so the visual modal will be gradually suppressed in the training process, and the valuable information in the image cannot be fully exploited. Some works [9, 21, 30] have attempted to solve the modal imbalance problem for late fusion architecture (different modalities are encoded using separate encoders, and finally there is only a simple layer of interaction, usually concatenation), but this is still a thorny problem for early fusion (i.e., single-stream) architecture. In this section, we propose **SDMI**, a **S**elf-**D**istillation pre-training paradigm to alleviate the **M**odality **I**mbalance problem and improve the model's ability to understand weak modality.

Specifically, for the same query-item pair, we remove either title or image information, and send them to MBSD again like Figure 4 (b) and (c). Similar to $\mathbf{h}_i$, we use $\mathbf{h}_i^t$ (or $\mathbf{h}_i^v$) denote the item representation with only textual (or visual) information and p_i^t (or p_i^v) denote the corresponding probability distribution of whether the query matches the item. We add two additional partial modal losses like Eq. 3:

$$\mathcal{L}_{\text{QT}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i^t) + (1-y_i)\log(1-p_i^t) \tag{4}$$

$$\mathcal{L}_{\text{QV}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i^v) + (1-y_i)\log(1-p_i^v) \tag{5}$$

Besides the hard targets, we also add some soft targets to inject the knowledge from the dominant modality to weak modality as presented in Figure 5. We first define the stronger modality s and weaker modality w and their corresponding probability distribution as Eq. 6.

$$p_i^s = \begin{cases} \min\left(p_i^t, p_i^v\right), & y_i = 0 \\ \max\left(p_i^t, p_i^v\right), & y_i = 1 \end{cases} \qquad p_i^w = \begin{cases} p_i^t, & p_i^s = p_i^v \\ p_i^v, & p_i^s = p_i^t \end{cases} \tag{6}$$

We want the prediction distribution of weak modal to be closer to that of the dominant modal. We minimize the Kullback—Leibler divergence between p_i^w and p_i^s:

$$\mathcal{L}_{\text{KL}}(i) = D_{\text{KL}}\left(p_i^w \| p_i^s\right) \tag{7}$$

We also want the representations of weak modal to be closer to that stronger modal. We minimize the euclidean distance between sample overall representations of these two partial-modal inputs:

$$\mathcal{L}_{\text{RS}}(i) = \frac{1}{d}\left\|\mathbf{h}_i^s - \mathbf{h}_i^w\right\|^2 \tag{8}$$

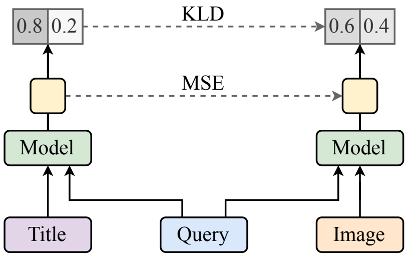


Figure 5: A demonstration of two self-distillation losses.

Applying a smaller weight to the knowledge distillation loss usually achieves better results [13]. In order not to introduce additional hyperparameters, we propose an adaptive loss weight:

$$w_i = \begin{cases} p_i^w - p_i^s, & y_i = 0 \quad \textit{and} \quad p_i^s < 0.5 \\ p_i^s - p_i^w, & y_i = 1 \quad \textit{and} \quad p_i^s > 0.5 \\ 0, & \textit{else} \end{cases} \tag{9}$$

which means that optimization is performed only when the prediction of the strong modality is correct, and the greater the gap with the weak modality, the greater the weight of the loss. These two knowledge distillation losses in a mini-batch can be expressed as:

$$\mathcal{L}_{\text{KL}} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \mathcal{L}_{\text{KL}}(i) \qquad \mathcal{L}_{\text{RS}} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \mathcal{L}_{\text{RS}}(i) \tag{10}$$

The final loss of SDMI is the sum of the above losses:

$$\mathcal{L} = \mathcal{L}_{\text{QI}} + \mathcal{L}_{\text{QT}} + \mathcal{L}_{\text{QV}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{RS}} \tag{11}$$

4 EXPERIMENTS

4.1 Datasets

All data used in this study is collected from *Taobao.com*, the largest e-commerce platform in China. For the preliminary pre-training, we random sample about 10M items, and for modality balance pre-training, we construct about 10M (*query*, *item*, *binary label*) triplets from search logs.

Table 1: Experimental results on two product attribute prediction dataset. † is the text model, ‡ is the visual model, and the others are vision-and-language models. Bold indicates the best result, the same below.

Model	IAVP-G					IAVP-F				
	Accuracy	P	R	F1	ROC-AUC	Accuracy	P	R	F1	ROC-AUC
BERT†	72.43	75.93	64.84	69.95	80.10	84.78	87.39	92.05	89.66	88.90
ViT‡	59.20	62.82	43.08	51.11	63.37	70.29	76.24	85.05	80.40	62.92
M6	70.26	67.92	75.66	71.59	76.70	84.34	85.48	94.14	89.60	88.36
FashionBERT	72.48	75.74	65.32	70.14	80.23	84.87	87.05	**92.68**	89.78	89.09
ViLT	69.11	73.34	59.08	65.45	76.27	83.20	85.51	92.17	88.72	86.78
OFA	71.20	73.31	65.78	69.34	78.50	83.38	86.46	91.08	88.71	86.86
MBSD	**74.14**	**77.92**	**66.66**	**71.85**	**82.13**	**85.57**	**87.94**	92.57	**90.19**	**89.86**

4.2 Baselines and Evaluation Metrics

The text description of products in our experiments is Chinese, so we will compare MBSD with the following pre-trained multimodal models with official Chinese versions:

- **FashionBert** [10]: The well-known vision-and-language model for cross-modal retrieval in fashion industry of e-commerce. In FashionBert, an image is divided into 16 patches, and ResNet-50 is employed to extract features of each patch. The visual features will be concatenated with word embeddings as the inputs of BERT. There are three pre-training tasks of FashionBert: masked language modeling, masked visual feature reconstruction and image-text matching.
- **M6** [17]: A transformer decoder-based model that perform prefix language modeling tasks such as image captioning and image-based text denoising for pre-training. The image input mode of M6 is the same as that of FashionBert. M6 excels at various text generation tasks such as product title generation. There are several versions of M6, and we use M6-base-chinese with 327M parameters in our experiments.
- **OFA** [29]: A unified architecture multimodal model based on encoder-decoder transformer, which can achieve eight vision-and-language tasks such as visual grounding, image captioning and visual question answering through the prompt mechanism. For visual information, OFA uses a learnable ResNet-101 to convolve the image into 196 patches as the transformer inputs. We use OFA-base-cn with 182M parameters in our experiments.

Besides, we implemented ViLT [14], the popular general domain multimodal on the same e-commerce data as MBSD. Compared with MBSD and FashinBERT, official ViLT uses ViT [8] instead of BERT as the backbone.

In addition, we also compare MBSD with some single-modal modal. We choose BERT as the text model and ViT as visual model.

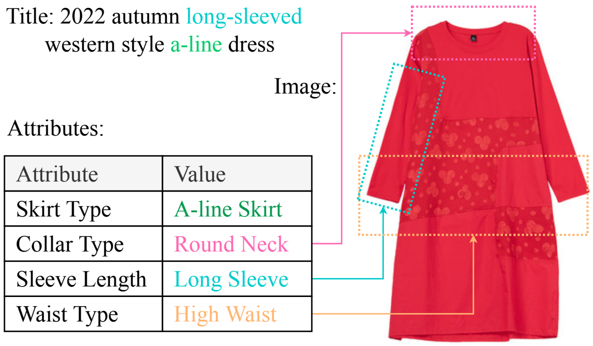
For classification task, accuracy is used to evaluate models' performance. And for alignment task, we also use accuracy, precision, recall, F1, and ROC-AUC.

4.3 Implementation Details

In this paper, we take the official bert-base-chinese as the backbone which has 768 hidden units and 12 self-attention heads in each layer. The maximum text sequence length is set to 144 and images are resise to 224 × 224. Lamb [37] optimizer is applied with the initial learning rating of 2e-4, $\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay of 0.01, learning rate warmed up at the first 1/10 steps, and then linear decay. For preliminary pre-training, we train 10 epoch with a batch size of 2048, and for modality balance pre-training, we train 3 epoch with a batch size of 1024. The experiments are implemented with PyTorch and conducted on 16 * NVIDIA Tesla V100-32G GPUs with mixed precision training. For downstream tasks, we tuning hyperparameters for each model and report the best results.

Table 2: Statistics of two item attribute prediction dataset.

Dataset	#Item	#Values	#Train	#Test
IAVP-G	61,581	945	48,625	13,376
IAVP-F	53,114	238	44,364	9,380



Title: 2022 autumn long-sleeved western style a-line dress

Image:

Attributes:

Attribute	Value
Skirt Type	A-line Skirt
Collar Type	Round Neck
Sleeve Length	Long Sleeve
Waist Type	High Waist

Figure 6: An example of predicting the product attributes from the textual description with the aid of the visual information.

4.4 Task #1: Product Attribute Prediction

Although product attribute provide details of the item and help users to make purchasing decisions, many retailers do not complete the attributes (or fill with the wrong values), and this task needs to be done by the e-commerce platform. As shown in Figure 6, it requires a better fusion of title and image to extract.

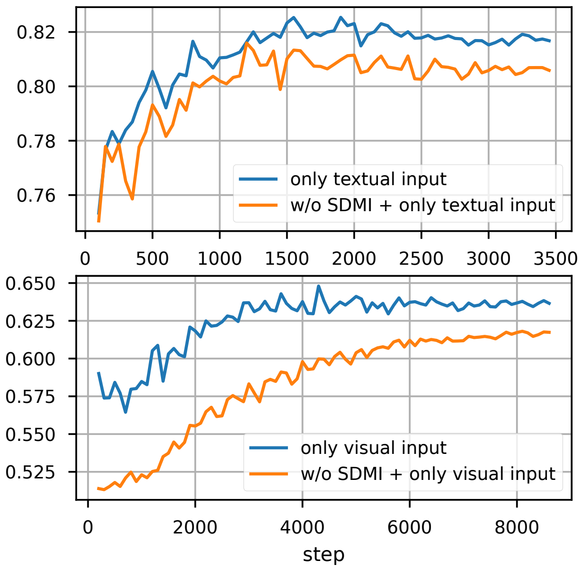


Figure 7: Performance (AUC) of the uni-modal modals with our proposed SDMI on the validation set of the IAVP-G.

To simplify the problem, we turn this task into a matching task, that is, given an attribute value (abbreviated as tag) and item pair, to determine whether the two are aligned. We simply concatenated the tag with item information like X={[CLS], *Tag*, [SEP], *Title*, [SEP], *Image*}, and feed to MBSD. A linear layer is will be applied to the representation of [CLS] token to calculate the match score. For ViT, we randomly initialize a lookup table of linguistic input, and text and images share the backbone like ViLT. We conduct experiments on two human-annotated datasets in the fashion domain, and the statistics are illustrated in Table 2. Table 3 shows some cases of product attribute. It can be seen from the literal sense that IAVP-F requires the model to pay more attention to the understanding of some details in the picture (such as sleeve length), while IAVP-G is more inclined to the overall understanding of the picture.

The experiment results are shown in Table 1, which indicate that our MBSD has excellent ability to understand the global and local multimodal information of items. We can observe that the language and visual modalities do not have equal contribution for the learning objective due to the natural imbalance existed in the two datasets. In addition, some multimodal models do not pay attention to the fusion between modalities, so the performance is even weaker than that of BERT which only accepts text input.

4.5 Task #2: Product Alignment

Product alignment, also known as similar product recognition, is widely used to deduplicate web search results and recommend similar products to customers.

We collected a manually labeled data set, including 46,290 training samples and 11,174 test samples, totaling 71,170 products. The form of each sample is (*source item*, *target item*, *binary label*).

We use a late-fusion structure for this task: the source and target item are mapped into latent vectors $\mathbf{h}_s$ and $\mathbf{h}_t \in \mathbb{R}^d$ by a shared MBSD separately, and a linear layer will perform on $\mathbf{h} = [\mathbf{h}_s; \mathbf{h}_t] \in \mathbb{R}^{2d}$ to calculate whether the two items match or not.

Table 3: Some cases of tags in IAVP-G and IAVP-F. The original tags are in Chinese, and we translated them into English.

Dataset	IAVP-G	IAVP-F
Tags	Style: Animation, Style: Gothic, Style: Ethnic, Style: Elegant, Style: Bohemian.	Sleeve Type: Flared Sleeves, Collar Type: Round Neck, Skirt Type: Pleated, Skirt Type: Suit Sleeve: Sleeveless

The results of product alignment are shown in Table 4, and our MBSD surpasses baselines on accuracy, recall and F1.

Table 4: Results on product alignment dataset.

Model	Acc	P	R	F1	AUC
BERT	83.69	82.71	81.61	82.16	90.60
ViT	84.40	82.62	**83.69**	83.15	91.11
ViLT	84.19	83.15	82.31	82.73	91.30
FashionBERT	84.32	83.32	82.43	82.87	90.88
MBSD	**84.73**	**84.37**	82.00	**83.17**	**91.53**

4.6 Task #3: Product Classification

The *category* is a vital attribute for describing a product, and is especially useful in many real-life applications [41]. We consider a classification task that judges the category of a product. This dataset contains 450K products for training and 100K products for testing, and products need to be classified into 543 categories such as table tennis, bumper and shoe brush.

The results are shown in Table 5. Our MBSD is superior in semantic learning compared with other baseline methods.

Table 5: Comparison of product classification task.

Model	BERT	ViT	ViLT	M6	FB	MBSD
Accuracy	91.45	64.32	86.88	92.03	91.69	**92.20**

4.7 Ablation Study and Discussions

We conduct extensive experiments on IAVP-G dataset to verify the effectiveness of the proposed MBSD, and results are presented in Table 6. For each experiment, We vary the value of batch size and learning rate to form 16 sets of hyperparameters and reported the mean and variance of ROC-AUC.

4.7.1 Effectiveness of Components. We first removed in-domain pre-training, and use a vanilla BERT with a randomly initialized image patchify network to fine-tune in downstream task directly. Result is illustrated in row 2 and the performance slumps, which means further pre-training in the field of e-commerce is necessary.

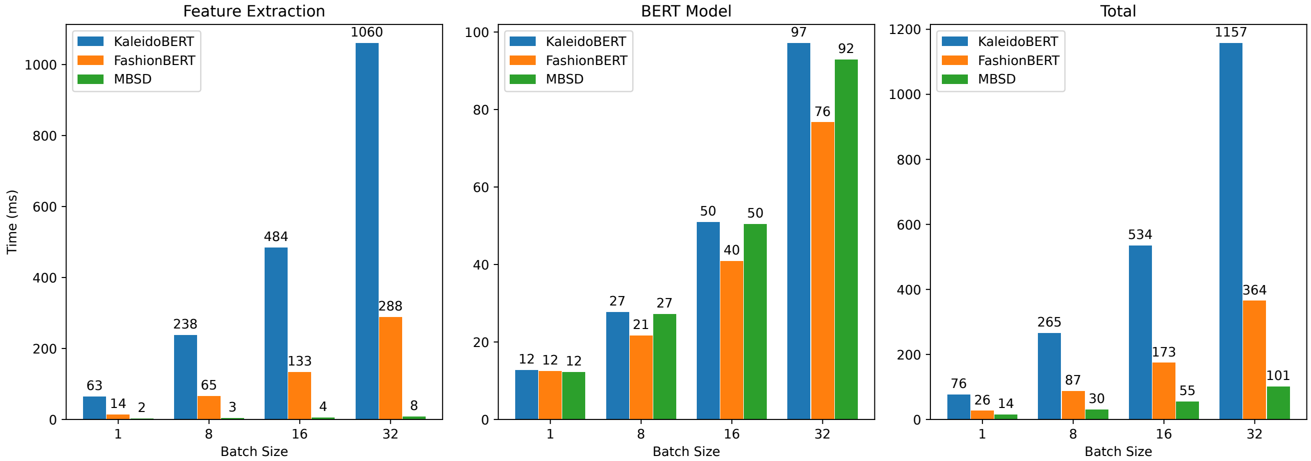


Figure 8: Comparison of testing time (ms for each step) at different batch size.

Table 6: Performance of models in the ablation study on IAVP-G dataset, where "w/o" is short for *without*.

ID	Model	AUC
1	MBSD	81.74 ± 0.30
2	w/o pre-training	78.70 ± 0.70
3	w/o preliminary pre-training	80.48 ± 0.41
4	w/o continue pre-training	81.31 ± 0.33
5	w/o continue pre-training and mini CNN	81.10 ± 0.58
6	w/o $\mathcal{L}_{\text{QT}}$, $\mathcal{L}_{\text{QV}}$, $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{RS}}$	81.35 ± 0.35
7	w/o $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{RS}}$	81.58 ± 0.32
8	w/o adaptive weight of $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{RS}}$	81.61 ± 0.31

As shown in row 3-4, both preliminary pre-training and modality balance pre-training contribute to the performance improvement, and removing either one will result in a performance degraded.

4.7.2 Effectiveness of Mini CNN. According to Kim et al. [14], the vision-and-language pre-training of BERT will collapse when equipped with linear projection for image inputs due to its post-layernorm mechanism. We replaced our patchify network with a linear layer, and results in row 4-5 demonstrate that convlutional patchify network has higher performance and smaller variance.

4.7.3 Influences of SDMI. To further analyze SDMI, we removed $\mathcal{L}_{\text{QT}}$, $\mathcal{L}_{\text{QV}}$, $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{RS}}$, and only keep the query-item match loss $\mathcal{L}_{\text{QI}}$. Results in row 6 indicates that weak modality may be suppressed during pre-training. Besides, $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{RS}}$ are removed, the results demonstrate that self-distillation helps to inject knowledge from stronger modality to weaker modality, and aid the training of multimodal models with additional uni-modal classifiers is not enough. In addition, adaptive weight of $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{RS}}$ are removed, and the performance decreased slightly, which indicates that specifying a constant weight manually impair model learning. Furthermore, to test the robustness of SDMI, we simulate a modality-missing scenario that remove either vision or language information of an product, and fine-tune our model w/ and w/o SDMI on product attribute values prediction task. As illustrated in Figure 7, our SDMI significantly improves the ability to mine weak modality.

4.7.4 Latency Comparison. We evaluate the inference latency of some vision-and-language models, and results are reported in Figure 8. The latency is averaged over 128 steps on a NVIDIA V100 16GB GPU. The above methods use the BERT-style backbone, but the processing on the image side is quite different. Although KaleidoBERT [41] does not provide Chinese version, we can still compare with it in terms of latency. M6 has a large number of parameters and OFA uses a non-BERT architecture, so they not participate in the latency comparison. It can be seen that our model has an extremely fast speed in image encoding. The visual token sequence length of FashionBERT, KaleidoBERT and our MBSD are 16, 49 and 55 respectively, so FashionBERT has lower latency in executing BERT. Right part of Figure 8 shows that our parameter-efficient model at least three times faster than those with heavy image encoder.

5 CONCLUSION

In this paper, we proposed MBSD, an end-to-end pre-trained vision-and-language model for product understanding. In this study, we devise an multimodal transformer to fuse the feature for the textual and visual information of products, and replace the heavy image feature extractor with a simple and lightweight convlutional network. We first perform masked language modeling and masked patch modeling to learn the preliminary representation of products. Besides, we construct an additional dataset from the large scale web search logs to further pre-train our model from a holistic perspective. In addition, to alleviate the modality imbalance issue during pre-training, we proposed a novel self-distillation training strategy named SDMI. SDMI can force the weak modality to imitate the dominant modality, thus improving the model's ability to understand the weak modality. Extensive experiments on four real-world datasets of three product understanding tasks verify the effectiveness of our method. In the future, larger scale data will be use to train MBSD.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proc. of CVPR*.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proc. of ICCV*.

[3] Chaoyu Bai. 2022. E-Commerce Knowledge Extraction via Multi-modal Machine Reading Comprehension. In *DASFAA'22 (Lecture Notes in Computer Science)*. Springer. https://doi.org/10.1007/978-3-031-00129-1_21

[4] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. In *Proc. of ICCV*.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Proc. of ECCV*.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of AACL*.

[7] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Xiaoyong Wei, Minlong Lu, and Xiaodan Liang. 2021. M5Product: A Multi-modal Pretraining Benchmark for E-commercial Product Downstream Tasks. (2021).

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc. of ICLR*.

[9] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. 2021. Improving Multi-Modal Learning with Uni-Modal Teachers. In *arXiv: 2106.11059*.

[10] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval. In *Proc. of SIGIR'20*.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. In *arXiv: 2111.06377*.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *Proc. of ICCV*.

[13] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv: 2203.05557.

[14] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML'21*.

[15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proc. of ICLR'20*.

[16] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Proc. of ECCV*.

[17] Junyang Lin, Rui Men, An Yang, Chang Zhou, Yichang Zhang, Peng Wang, Jingren Zhou, Jie Tang, and Hongxia Yang. 2021. M6: Multi-Modality-to-Multi-Modality Multitask Mega-transformer for Unified Pretraining. In *Proc. of KDD'21*.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proc. of ICCV*.

[19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Proc. of NeurIPS*.

[20] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention Bottlenecks for Multimodal Fusion. In *Proc. of NeurIPS*.

[21] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *Proc. of CVPR*.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML'21*.

[23] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. FLAVA: A Foundational Language And Vision Alignment Model. In *arXiv: 2112.04482*.

[24] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *Proc. of ICLR*.

[25] Ya Sun, Sijie Mai, and Haifeng Hu. 2021. Learning to Balance the Learning Rates Between Various Modalities via Adaptive Tracking Factor. *IEEE Signal Process. Lett.* (2021).

[26] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proc. of EMNLP*.

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

[28] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *Proc. of CVPR*.

[29] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *ICML'22*.

[30] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What Makes Training Multi-Modal Classification Networks Hard?. In *Proc. of CVPR*.

[31] Rong Xiao, Jianhui Ji, Baoliang Cui, Haihong Tang, Wenwu Ou, Yanghua Xiao, Jiwei Tan, and Xuan Ju. 2019. Weakly Supervised Co-Training of Query Rewriting andSemantic Matching for e-Commerce. In *Proc. of WSDM*.

[32] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross B. Girshick. 2021. Early Convolutions Help Transformers See Better. In *Proc. of NeurIPS*.

[33] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2021. SimMIM: A Simple Framework for Masked Image Modeling. arXiv: 2111.09886.

[34] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. arXiv: 2202.10401.

[35] Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. MAVE: A Product Dataset for Multi-source Attribute Value Extraction. In *WSDM'22*.

[36] Shaowei Yao, Jiwei Tan, Xi Chen, Keping Yang, Rong Xiao, Hongbo Deng, and Xiaojun Wan. 2021. Learning a Product Relevance Model from Click-Through Data in E-Commerce. In *Proc. of WWW*.

[37] Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *ICLR'20*.

[38] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. 2021. Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining. In *ICCV'21*.

[39] Qian Zhao, Jilin Chen, Minmin Chen, Sagar Jain, Alex Beutel, Francois Belletti, and Ed H. Chi. 2018. Categorical-attributes-based item classification for recommender systems. In *RecSys'18*.

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional Prompt Learning for Vision-Language Models. arXiv: 2203.05557.

[41] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-BERT: Vision-Language Pre-Training on Fashion Domain. In *Proc. of CVPR*.