

How to Attend to Different Modalities Equally: Unsupervised Learning for Multi-Modal Product Embedding

Anonymous EMNLP submission

Abstract

Product embedding, serving as the cornerstone for a wide range of applications in e-commerce, has been heavily studied in the past few years. The product embedding learned from multiple modalities shows significant improvement than that from a single modality, since complementary information is given with different modalities. However, some modalities are normally informatively dominant than the others. How to teach a model to learn embedding from different modalities without putting too much attention on the dominant one is an unsolved problem. We present Vision-Text BERT (VT-BERT), an unsupervised learning method that is designed to better attend to vision and text modalities equally. We extend BERT by (1) learning embedding from text and image without knowing the regions of interest; (2) training a global representation to predict masked title tokens and to construct masked image patches without the individual representations. We evaluate our VT-BERT on two tasks: the search for extremely similar products and the prediction of product categories, showing substantial gains compared to strong baseline models¹.

1 Introduction

Online shopping sites tend to employ visual similarity to improve recall and relevance of search and recommendation results (Jing et al., 2015; Shankar et al., 2017; Yang et al., 2017). The visual similarity is quite useful in some domains, like fashion and furniture. However, the product titles also play a key role when the outlook of the products is similar, like the phones of different generations. These two modal information can be easily obtained from the seller. One might improve the recommendation system with both product image and title.

Figure 1 shows some product examples from our own collected dataset, vision-text online prod-







Meta Category	Dolls & Bears	Dolls & Bears	Cameras & Photo	Business & Industrial	Clothing	Home & Garden
Leaf Category	Care Bears	Care Bears	Telescopes	Multimeters	T-shirts	Generators
Title	Care Bears Vintage Bottomline Bear 13" Plush Kamekiden Creations 1983 Excellent Con	Care Bears Vintage Bottomline Bear 13" Plush Kamekiden Creations 1983 Excellent Con	Bushnell Deep Space 78-9512 80mm Reflector Telescope	Fuke Electrical Tester T+PRO	GUCCI T-shirt Sherry Line Washed Black Logo XS Men Cotton Kn3015	Champion Power Equipment 6500 Watt 986cc 4-Stroke Gas Portable Generator
Image						

Figure 1: Examples from the index set of VTOP. Four types of information are given for each product, i.e. meta category, leaf category, title and image. Compared to the index set, the meta categories and the leaf categories of products in the query set are not given.

uct dataset (VTOP). Compared to the image, one can usually obtain more information from the title. Sometimes the category name even exists in the title. If we want to train a model to extract embedding from this cross-modal dataset, one problem is to prevent the model from attending too much to the title. Otherwise the image is useless.

In this paper, we propose a new model, Vision-Text BERT (VT-BERT), to learn the fine-grained embedding from both product image and title in an unsupervised way. Inspired by Transformer (Vaswani et al., 2017), our model use the Transformer encoder as the basic structure. Both image and title are first encoded into a sequence of vectors. These vectors are then concatenated and fed into the transformer encoder. Different with other similar models Li et al. (2019); Su et al. (2019); Qi et al. (2020); Lu et al. (2019); Tan and Bansal (2019) that feeds the regions of interest of a image as input, we input the whole image without object detection. Besides, we design two new objectives to lead the model to attend to image and title as equally as possible. One is to predict the missing title tokens with their corresponding position embeddings and a global representation. The another is to construct the masked patches of a image with their corresponding position embeddings and a global representation. With these two new objectives, VT-BERT could learn fine-grained embedding from

¹The code and the pre-trained model will be released.

image and text.

We evaluate our VT-BERT on two tasks: the search for extremely similar product and the prediction of the product category. Since there lacks of relevant annotated data, we collect our own data and train VT-BERT on it. The result shows that VT-BERT can retrieve almost the “same” products, and has a higher accuracy for the prediction task.

To the best of our knowledge, this is the first work using multi-modal information to retrieve the “same” products. Our contributions are summarized as follows:

- We collect a large-scale multi-modal product dataset, and annotate it for the search of extremely similar products.
- We learn the cross-modal embedding without knowing the regions of interest. The regions of interest need to be detected by a pre-trained model, which requires the evaluation task to be in the same domain of the pre-trained dataset. Without knowing the regions of interest gives our model more possibility on various tasks.
- We design two new pre-training objectives, forcing the model to extract information from different modalities as equally as possible, making full usage of all modalities.

2 Related Work

Image and text are two most common modalities. Bridging them with an efficient way has a long history. Various tasks, such as image captioning (You et al., 2016), textual grounding (Kazemzadeh et al., 2014; Plummer et al., 2015), visual question answering (Antol et al., 2015; Goyal et al., 2017) and visual reasoning (Suhr et al., 2018; Zellers et al., 2019), have been proposed. To solve these tasks, various models have been developed. These models can be split into two types: a single-stream model (Li et al., 2019; Su et al., 2019; Qi et al., 2020) that fuses vision-and-text information at the very beginning and a two-stream model (Lu et al., 2019; Tan and Bansal, 2019) that consists of an image encoder, a text encoder and a multi-modal fusion module at the end.

Our VT-BERT is a single-stream model with only one encoder for both vision and text. There are three main differences between VT-BERT and other cross-modal models:

Subset	# of Images	Titles	Meta	Leaf
Index	1,101,396	Yes	15	1,275
Query	5,000	Yes	-	-

Table 1: VTOP statics.

- Inspired by Dosovitskiy et al. (2020), VT-BERT use the whole image as input without the detection of the regions of interest, while the other methods (Li et al., 2019; Su et al., 2019; Qi et al., 2020; Lu et al., 2019; Tan and Bansal, 2019) employ a pre-trained model to detect the regions of interest and apply these regions as input. The detection of regions of interest is beneficial for some specific tasks, like image captioning, visual question answering and so on. However, if the dataset of the downstream task is out-of-domain, the performance of the object detection in this dataset is poor, which restricts the implementation of the model for poorly annotated tasks.
- We introduce new objectives to force the model to attend to different modalities equally, and to learn a better embedding. Text is normally informatively dominant for most cross-modal datasets. If the objective is not well-defined, the model tend to attend more to the text and ignore the vision.
- We collect a new dataset to evaluate the pre-trained model. Li et al. (2019); Su et al. (2019); Qi et al. (2020); Lu et al. (2019); Tan and Bansal (2019) evaluate the pre-trained models by finetuning them on downstream tasks. There is a mismatch between the pre-training and finetuning, making it difficult to evaluate the pre-training method. We evaluate the pre-trained VT-BERT directly with a super fine-grained product recommendation task.

3 Vision-Text Online Product Data Collection

How to evaluate the pre-trained VT-BERT is a key point in our paper. Most cross-modal pre-trained models (Lu et al., 2019; Li et al., 2019; Su et al., 2019; Tan and Bansal, 2019; Desai and Johnson, 2020) use the Conceptual Captions dataset (Sharma et al., 2018) or the Microsoft COCO Captions dataset (Chen et al., 2015) as pre-training

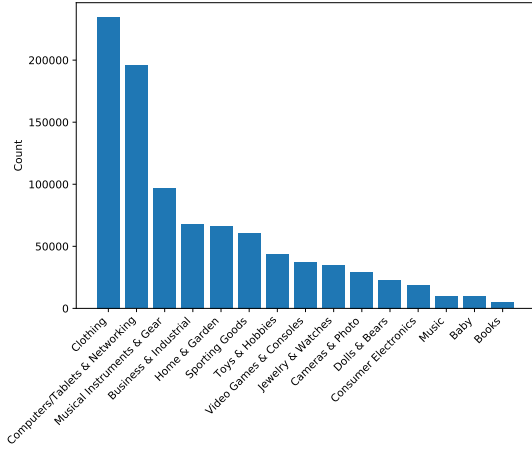


Figure 2: Index set distribution over meta category.

datasets. These models are then finetuned on retrieval tasks or question-answering tasks for evaluation. There is a mismatch between the pre-training and finetuning methods, which makes it difficult to demonstrate the effectiveness of a pre-training method. Even though some works, like Qi et al. (2020), evaluate the embedding from a pre-trained model on zero-shot retrieval tasks, there still lacks of dataset to evaluate the fine-grained embedding.

In this paper, we collect a large-scale vision-text online product dataset from an online shopping website. VTOP is a large-scale, diverse, well-distributed and cleanly annotated dataset for training, given a large number of likely fine-grained categories, many with subtle differences not easily distinguishable. VTOP has two sets: query set and index set that is the search space for queries. Some examples from the index set are shown in Figure 1. VTOP uses the hierarchical taxonomy structure which is a sub-tree from a real and large e-commerce inventory. It includes 15 meta cate-

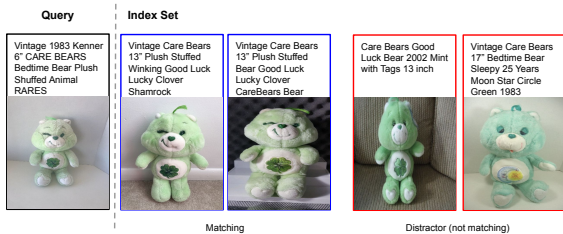


Figure 3: A search example of VTOP: we show a query on the left with its two true matches and two distractors on the right. Distractors are “hard” examples because they all come from the same leaf category as the query, i.e. “Care Bears”, yet only the true matches share the same product model.

gories and 1275 leaf categories. 5,000 query products alongside 1.1 million index products are used for a search benchmark. Its statics are shown in Table 1. The sampled 15 meta-category distribution of the index set is shown in Figure 2, a long-tailed distribution.

Construction of Query Set Diversity was prioritized when collecting the queries. Firstly, we collect a 25 million dataset which covers 27 meta and 3,500+ leaf categories. We then manually select over 200 leaf categories from the 25 million dataset to construct our 5,000 queries.

Construction of Index Set The index set is the search space in our search task. It implicitly consists of two subsets, groundtruth matches and distractors. A groundtruth match is an exact product match to any of the queries, while a distractor is not a match to any queries. A clean and large-scale index set makes a multi-modal search dataset valuable in terms of accurate evaluations of a pre-trained cross-modal model. The index set contains 1 million products that covers 1,275 leaf categories. More details on the construction of the index set are shown in A.

With all of these procedures, we finally build a large dataset for super fine-grained recognition. A search example is shown in Figure 3.

4 Methodology

Figure 4 illustrates the overall architecture of our VT-BERT model. Similar to BERT (Devlin et al., 2018), we use the transformer encoder as our basic structure. The transformer encoder shares the same setting as BERT_{BASE}, i.e. with 12 transformer encoder layers where the hidden size is 768 and the number of attention heads are 12. The image and title are encoded into different embeddings with different embedding layers. Then theses embeddings are fed into a bidirectional self-attention transformer encoder to model the cross-modal relationship between vision and text.

4.1 Embedding Modeling

We use three types of embeddings to encode image and text, i.e. token embeddings, segment embeddings and position embeddings. Compared to the embeddings in BERT, the most different part in VT-BERT locates on the embeddings for image.

Text Embedding We tokenize the product title into N sub-word tokens t_0, \dots, t_{N-1} with WordPiece embeddings (Wu et al., 2016) which has a

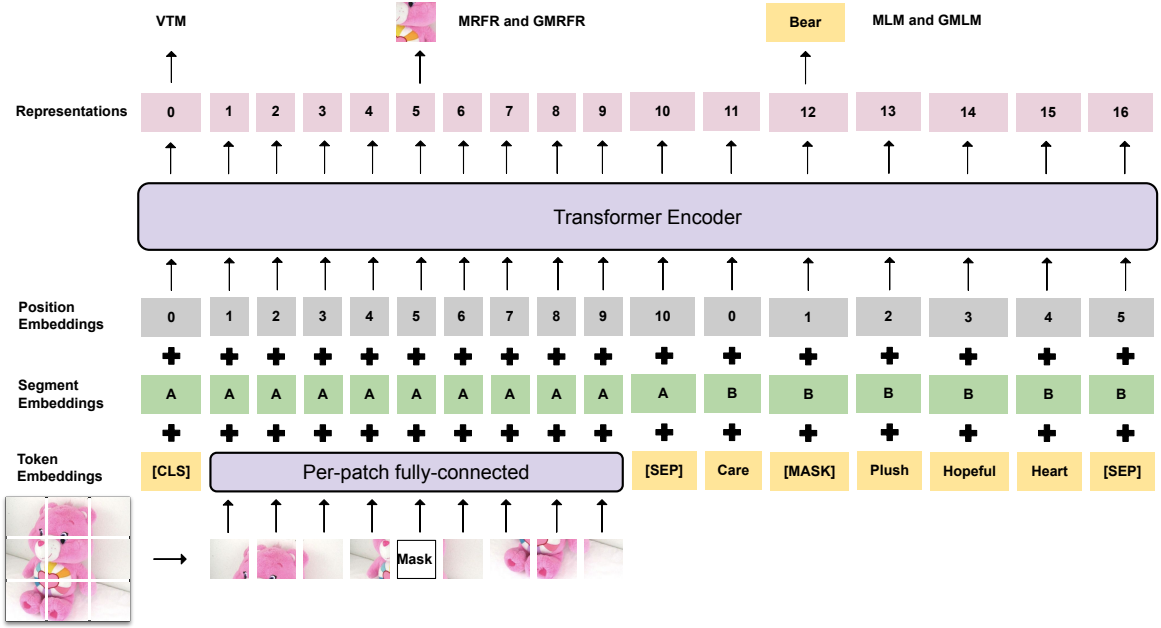


Figure 4: Architecture of VT-BERT model. The full image is split into multiple patches before input to a fully-connected layer. VT-BERT is pre-trained with five objectives: vision-text matching (VTM), masked region feature regression (MRFR), masked region feature regression based on the global information (GMRFR), masked language modeling (MLM) and masked language modeling based on the global information (GMLM). The sampling of patches or tokens to be masked is random.

30,000 token vocabulary. This 30,000 English uncased vocabulary is borrowed directly from BERT. Special token [SEP] is added to the end of the title (t_{N-1} stands for the token [SEP]). The final embedding for each sub-word token is generated by combining its token embedding, segment embedding and position embedding.

Vision Embedding The whole image is split into multiple patches, and then fed into a fully-connected layer to generate a sequence of vectors (v_0, \dots, v_{M-2}). The token embedding of the special token [SEP] is then added to the end of the image vectors as v_{M-1} . The vision embedding is generated by combining the segment embedding and the position embedding.

Different with other cross-modal models (Lu et al., 2019; Li et al., 2019; Su et al., 2019; Tan and Bansal, 2019) where regions of interest detected by a pre-trained model are fed into the fully-connected layer, we ignore object detection and input the whole image to our VT-BERT. There are two reasons for this: (1) Most online product images only contain one object. I.e. we only have one region of interest, it is the whole image. (2) Finding regions of interest requires an in-domain pre-trained object-detection model, which weak-

ens the wide application of a pre-training method. We want to build a model that can use as many vision-and-text data as possible, whereas the pre-trained object-detection model requires expensive annotated data.

Global Embedding A virtual token [CLS] is added at the beginning of the input sequence to the transformer encoder. Similar to the text embedding, the final global embedding is generated with combining token embedding, segment embedding and position embedding. The global embedding shares the same segment embedding with the vision embedding.

The final global, image and text embedding are concatenated together before fed into the transformer encoder.

4.2 Pre-training VT-BERT

Our VT-BERT is pre-trained with five unsupervised tasks, trying to learn the relationship between vision and text, and make the model attend to two modalities as equally as possible.

Vision Text Matching (VTM) Some important cross-modal downstream tasks, e.g. category prediction and visual question answering, require the understanding of the relationship between image

and text. Similar to the next sentence prediction task of BERT, we pre-train for a binarized matching prediction task. During our pre-training, the product image and title are not always matched. 50% of the time, the product title is the actual title of the image. 50% of the time, the product title is a random title from an arbitrary product. The global representation, i.e. the output representation of [CLS] token, is fed into a fully-connected layer to obtain the vision-text similarity score. A binary classification loss is used for optimization.

Masked Language Modeling (MLM) MLM is the task of predicting missing tokens in a sequence from their placeholders. Same as BERT’s implementation, in 15% of the title tokens, 80% are replaced by [MASK] token, 10% are replaced with a random token (according to the unigram distribution), and 10% are kept unchanged. We build a fully-connected layer and a token embedding layer on top of the output representations of these tokens to predict them. A cross entropy loss is used for optimization.

Masked Region Feature Regression (MRFR) Inspired by Chen et al. (2019), MRFR task is to construct the masked patches of a image from their placeholders. In 15% of the image patches, 80% are masked and 15% are kept unchanged. A fully-connected layer is built on top of these masked output representations to project it back to the original patch dimension. A L2 loss is applied to regress the ground truth feature. This task is formulated as:

$$\mathcal{L}_{\text{MRFR}} = -\frac{1}{I} \sum_{i=1}^I \|h(v_m^i) - r(v_m^i)\|_2^2 \quad (1)$$

where I is the total number of masked patches, $h(v_m^i)$ is the constructed embedding of the masked patches and $r(v_m^i)$ are the original patches.

Masked Language Modeling Based on the Global Information (GMLM) Compared to the next sentence prediction in BERT’s implementation, VTM is a rather easy task. Empirically, we can obtain about 98% accuracy for VIM within a few of iterations. This easy task makes it difficult to summarize the cross-modal information in a global output representation, i.e. the output representation of [CLS].

To summarize the global information in the global output representation, we design a new objective, predicting the missing tokens with the global output representation and the correspond-

ing position embeddings rather than their individual representations. Supposed $h(t_0), \dots, h(t_{N-1})$ is the output representations of the title tokens and t_n is the masked token, we want to predict the original token with its own position embedding p_n and the global output representation $h([\text{CLS}])$.

$$y_n = f(h([\text{CLS}]), p_n) \quad (2)$$

Inspired by the span boundary objective in SpanBERT (Joshi et al., 2020), we implement the representation function $f(\cdot)$ as a 2-layer feed-forward network with GeLU activations (Hendrycks and Gimpel, 2016) and layer normalization (Ba et al., 2016):

$$\begin{aligned} h_0 &= [h([\text{CLS}]); p_n] \\ h_1 &= \text{LayerNorm}(\text{GeLU}(W_0 h_0)) \\ y_n &= \text{LayerNorm}(\text{GeLU}(W_1 h_1)) \end{aligned} \quad (3)$$

We use the representation vector y_n to predict the masked token t_n . For the example in Figure 4, it is $\mathcal{L}_{\text{GMLM}}(\text{Bear}) = -\log(\text{Bear}|h([\text{CLS}]), p_1)$. With this objective, we force the model to summarize the information of the masked tokens in the global output representation.

Masked Region Feature Regression Based on the Global Information (GMRFR) Similar to GMLM on the text side, we design GMRFR on the image side to construct the masked patches with the global output representation and their individual position embeddings, forcing the model to summarize the information of the masked patches in the global output representation:

$$\begin{aligned} h_0 &= [h([\text{CLS}]); p_m] \\ h_1 &= \text{LayerNorm}(\text{GeLU}(W_0 h_0)) \\ z_m &= \text{LayerNorm}(\text{GeLU}(W_1 h_1)) \\ \mathcal{L}_{\text{GMRFR}} &= -\frac{1}{I} \sum_{i=1}^I \|z_m^i - r(v_m^i)\|_2^2 \end{aligned} \quad (4)$$

where p_m is the position embedding of the masked patch.

Combining these five objectives, we can get the final loss of VT-BERT for optimization:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \lambda_{\text{VTM}} \mathcal{L}_{\text{VTM}} + \lambda_{\text{MLM}} \mathcal{L}_{\text{MLM}} \\ &\quad + \lambda_{\text{GMLM}} \mathcal{L}_{\text{GMLM}} + \lambda_{\text{MRFR}} \mathcal{L}_{\text{MRFR}} \\ &\quad + \lambda_{\text{GMRFR}} \mathcal{L}_{\text{GMRFR}} \end{aligned} \quad (5)$$

Empirically, we can obtain good performance without tuning the interpolation weight λ . We only need to set $\lambda_{\text{VTM}} = 1$, $\lambda_{\text{MLM}} = \lambda_{\text{GMLM}} = 0.1$ and $\lambda_{\text{MRFR}} = \lambda_{\text{GMRFR}} = 0.01$, making sure all of these losses are at the same scale, i.e. equally important.

5 Experimental Setup

5.1 Tasks

There are two downstream tasks to evaluate our pre-trained VT-BERT: same product recommendation and leaf category prediction.

Same Product Recommendation As shown in Figure 3, this task requires fine-grained embedding from a pre-trained model to retrieve the same product according to the query. Compared to the similar product recommendation, the same product recommendation has higher requirement for the cross-modal embedding. It requires the embedding to be good enough for subtle difference among products.

Our major criteria for this task are Macro-Average Recall@k (MAR@k) and Macro-average Precision@k (MAP@k) (Yang, 1999). The higher the better. We choose $k = 10$.

We randomly split the index set of VTOP by a ratio of 8/2 into train/valid set, and also split the query set by a ratio of 2/3 into development/test set. VT-BERT is trained on the index set with our five objectives and never sees the data from the query set during training. For evaluation, we don't mask any patch or token, using the embedding from a pre-trained VT-BERT to compute the similarity scores between a query and all products in the index set.

There are three encoded embeddings we can obtain from a pre-trained VT-BERT, i.e. the global embedding, the text embedding and the vision embedding. The global embedding e_g is the output of the $h([CLS])$ through a fully-connected layer in VIM task, i.e. the pooled output of the BERT implementation. The text embedding e_t is the average representation of the title tokens:

$$e_t = \frac{1}{N} \sum_{n=0}^{N-1} h(t_n) \quad (6)$$

where N is the number of the title tokens. The image embedding e_i is the average representation of the image patches:

$$e_i = \frac{1}{M} \sum_{m=0}^{M-1} h(v_m) \quad (7)$$

where M is the number of the patches, including the special token [SEP] at the end of image patches.

We employ cosine distance to measure the similarity between a query and a product in the index set (normalize the embedding before computing the cosine distance):

$$s_g(i, j) = (e_g^q)^T e_g^i \quad (8)$$

where e_g^q is the global embedding for the q -th query and e_g^i is the global embedding for the i -th index item. The similarity score for the text embedding $s_t(i, j)$ or the vision embedding $s_v(i, j)$ have the similar definition. The final similarity score for the q -th query and the i -th index item is:

$$s(i, j) = \max(s_g(i, j), s_t(i, j), s_v(i, j)) \quad (9)$$

Leaf Category Prediction The same product recommendation task can be viewed as a super fine-grained recognition task. Besides this, leaf category prediction (can be viewed as a similar product recommendation task) is another import task in e-commerce. Since the query set doesn't contain leaf categories, we don't use it for this task. We randomly split the index set by a ratio of 8/1/1 into train/valid/test set.

Different with the same product recommendation task, the leaf category prediction requires training VT-BERT in a supervised way. We first pre-train VT-BERT with extra 10 million dataset on five objectives, then finetune it on the train set with classification loss by building a new classifier layer on top of the global representation. Our major criteria for this task are top1 and top5 accuracy.

5.2 Implementation

We implement VT-BERT, baseline models and pre-training method in fairseq (Ott et al., 2019). For the pre-training on the index set, the learning rate is warmed up over the first 4,000 steps to a peak value of $1e-4$, and then linearly decayed. We set β hyperparameters ($\beta_1 = 0.9$, $\beta_2 = 0.98$) and a decoupled weight decay (Loshchilov and Hutter, 2017) of $1e-4$. We deviate from the optimization by running for 100k steps and using an ϵ of $1e-8$ for AdamW (Kingma and Ba, 2014).

The hyperparameters' setting (including the number of layers, dropout, the number of attention heads, activation function and hidden size) of the transformer encoder stays the same as the original BERT_{BASE}. We use a batch size of 512, a max title length of 36 and an image size of 224 with 16

Model	Dev. Query Set		Test Query Set	
	MAR@10	MAP@10	MAR@10	MAP@10
ResNet50 on image	0.3851	0.3067	0.3843	0.3024
BERT on title	0.4261	0.3725	0.4627	0.4113
CLIP on image and title	0.5905	0.5249	0.5798	0.5150
VT-BERT with cls on image and title	0.5479	0.4946	0.5414	0.4873
VT-BERT on image and title	0.6753	0.6056	0.6696	0.6005

Table 2: The Performance of different models on the same product recommendation task. The higher the better. The fourth experiment, “VT-BERT with cls on image and title”, demonstrates that we train VT-BERT in a supervised way with a classification loss to predict the leaf category.

Model	Dev. Query Set		Test Query Set	
	MAR@10	MAP@10	MAR@10	MAP@10
VTM	0.0106	0.0069	0.0117	0.0078
VTM + MLM + MRFR	0.6775	0.6069	0.6653	0.5935
VTM + GMLM + GMRFR	0.6617	0.5929	0.6565	0.5884
MLM + GMLM + MRFR + GMRFR	0.6763	0.6065	0.6690	0.5959
VTM + MLM + GMLM + MRFR + GMRFR	0.6724	0.6043	0.6676	0.5985

Table 3: The Performance of VT-BERT with different pre-training objectives on the same product recommendation task. Except for the pre-training with only VTM, the other experiments provide similar performance.

as the patch size. Random resizing and horizontal flip techniques are used for image augmentation.

The pre-training was done on 8 Volta V100 GPUs and takes 2 days to complete. The pre-training on 10 million dataset is updated 164k times with a batch size of 2304 on 36 Volta V100 GPUs, completed in 4 days. The other settings stay the same as the above. The training settings for baseline models and for the finetuning of VT-BERT are shown in Appendix B. The best checkpoint is determined by the minimal $\mathcal{L}_{\text{total}}$ on the valid set.

6 Results and Analysis

6.1 Same Product Recommendation

The same product recommendation is a much more difficult task than the similar product recommendation. It requires the model to learn the fine-grained feature of a product.

Performance of Different Models As shown in Table 2, the cross-modal models, VT-BERT and CLIP, outperform the models trained on a single modality. Because Complementary information is offered from different modalities, making them more informative. Another observation is that the models trained with unsupervised learning achieve better results than the one with supervised learning. When training to predict the leaf category, in a supervised way, the model tends to learn the com-

mon pattern within each category. During search, the supervised-learning model might consider the products with the same leaf category as the similar products, resulting worse score for the same product recommendation. The best result is obtained by our VT-BERT trained with five objectives. It is powerful enough to retrieve the products with similar images and similar titles, resulting in the same products. Some retrieval examples are shown in Figure 5.

Ablation Study of the Pre-training Objectives Is it necessary for us to pre-train VT-BERT with all of the five objectives? Or can one obtain a rather good performance with a few of them? Table 3 shows the performance of models with different objectives. The results from the last four models trained with different combination of objectives are very similar, almost at the same scale. However, if we look at the performance of different embeddings (as shown in Table 4), the VT-BERT trained with all of the five objectives get the best scores for all three embeddings. This phenomenon gives more potentials to our VT-BERT for different downstream tasks. If the dataset of a downstream task is text dominant (like VTOP), one can only use the text embedding for a good performance. Besides, the global embedding from VT-BERT trained with five objectives is the best, way better than the models trained with other objectives. The reason

Model	Test Query Set					
	MAR@10			MAP@10		
	Global	Vision	Title	Global	Vision	Title
VTM	0.0025	0.0015	0.0117	0.0020	0.0012	0.0078
VTM + MLM + MRFR	0.2305	0.1320	0.6653	0.1841	0.1072	0.5935
VTM + GMLM + GMRFR	0.2225	0.1396	0.6565	0.1729	0.1128	0.5884
MLM + GMLM + MRFR + GMRFR	0.2054	0.0553	0.6690	0.1703	0.0417	0.5959
VTM + MLM + GMLM + MRFR + GMRFR	0.2843	0.1427	0.6676	0.2287	0.1184	0.5985

Table 4: The Performance of different embeddings (the global embedding e_g , the vision embedding e_v and the text embedding e_t) from VT-BERTs trained with various combination of objectives. The VT-BERT trained with all five objectives almost outperform the other models for all of these three embeddings, especially for the global embedding.

Model	Valid Set		Test Set	
	Top 1	Top 5	Top 1	Top 5
ResNet50 on image	74.72	90.68	74.92	90.84
BERT on title	88.79	98.07	88.92	98.05
VT-BERT with cls on title and image	89.49	98.12	89.82	98.11
VT-BERT on image and title	90.79	98.51	91.09	98.52

Table 5: The Accuracy of different models on validation set and test set. The third experiment, “VT-BERT with cls on image and title”, demonstrates that we train VT-BERT in a supervised way on the train set with a classification loss to predict the leaf category. We first pre-train the last model, “VT-BERT on image and title”, on 10 million dataset, and then finetune it on the train set from the index set.

for this is that GMLM and GMRFR force the VT-BERT to summarize cross-modal information in the global output representation.

6.2 Prediction of Product Category

From Table 4, VT-BERT has a well-performed global embedding, which means that it can summarize the cross-modal information in a global representation. Motivated by this, we finetune VT-BERT for the prediction of the leaf category. Table 5 shows that VT-BERT with a pre-training procedure achieves the best results. In addition, there is a huge performance gap between ResNet50 and BERT, while only subtle difference between BERT and VT-BERT. This is caused by the bias of our VTOP dataset, where the title plays the major role and contains more product information than the image. For same cases, the leaf category names could appear in the titles.

7 Conclusion

In this paper, we present VT-BERT, a pre-trained model for joint vision-and-text representation. With five pre-training objectives, especially the two newly designed objectives that lead VT-BERT to extract information from vision and text equally,

and induce VT-BERT to learn fine-grained features from the product. The embeddings from a pre-trained VT-BERT are used for the search of extremely similar products. We also finetune VT-BERT on prediction task and obtain very good accuracy.

In the future, we plan to evaluate our model on public datasets that are rare at this moment. If possible, we also want to make parts of our own collected dataset public, or provide it for research usage.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

569	Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed	Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui	624
570	El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and	Hsieh, and Kai-Wei Chang. 2019. Visualbert: A	625
571	Jingjing Liu. 2019. Uniter: Learning universal	simple and performant baseline for vision and lan-	626
572	image-text representations.	guage. <i>arXiv preprint arXiv:1908.03557</i> .	627
573	Karan Desai and Justin Johnson. 2020. Virtex: Learn-	Ilya Loshchilov and Frank Hutter. 2017. Decou-	628
574	ing visual representations from textual annotations.	pled weight decay regularization. <i>arXiv preprint</i>	629
575	<i>arXiv preprint arXiv:2006.06666</i> .	<i>arXiv:1711.05101</i> .	630
576	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and	Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan	631
577	Kristina Toutanova. 2018. Bert: Pre-training of deep	Lee. 2019. Vilbert: Pretraining task-agnostic visi-	632
578	bidirectional transformers for language understand-	olinguistic representations for vision-and-language	633
579	ing. <i>arXiv preprint arXiv:1810.04805</i> .	tasks. <i>arXiv preprint arXiv:1908.02265</i> .	634
580	Alexey Dosovitskiy, Lucas Beyer, Alexander	Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Sil-	635
581	Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,	vio Savarese. 2016. Deep metric learning via lifted	636
582	Thomas Unterthiner, Mostafa Dehghani, Matthias	structured feature embedding. In <i>Proceedings of</i>	637
583	Minderer, Georg Heigold, Sylvain Gelly, et al. 2020.	<i>the IEEE conference on computer vision and pattern</i>	638
584	An image is worth 16x16 words: Transformers	<i>recognition</i> , pages 4004–4012.	639
585	for image recognition at scale. <i>arXiv preprint</i>		
586	<i>arXiv:2010.11929</i> .	Myle Ott, Sergey Edunov, Alexei Baevski, Angela	640
587	Yash Goyal, Tejas Khot, Douglas Summers-Stay,	Fan, Sam Gross, Nathan Ng, David Grangier, and	641
588	Dhruv Batra, and Devi Parikh. 2017. Making the	Michael Auli. 2019. fairseq: A fast, extensible	642
589	v in vqa matter: Elevating the role of image under-	toolkit for sequence modeling. In <i>Proceedings of</i>	643
590	standing in visual question answering. In <i>Proceed-</i>	<i>NAACL-HLT 2019: Demonstrations</i> .	644
591	<i>ings of the IEEE Conference on Computer Vision</i>		
592	<i>and Pattern Recognition</i> , pages 6904–6913.	Bryan A Plummer, Liwei Wang, Chris M Cervantes,	645
593	Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian	Juan C Caicedo, Julia Hockenmaier, and Svetlana	646
594	Sun. 2016. Deep residual learning for image recog-	Lazebnik. 2015. Flickr30k entities: Collecting	647
595	nition. In <i>Proceedings of the IEEE conference on</i>	region-to-phrase correspondences for richer image-	648
596	<i>computer vision and pattern recognition</i> , pages 770–	to-sentence models. In <i>Proceedings of the IEEE</i>	649
597	778.	<i>international conference on computer vision</i> , pages	650
598	Dan Hendrycks and Kevin Gimpel. 2016. Gaus-	2641–2649.	651
599	sian error linear units (gelus). <i>arXiv preprint</i>	Di Qi, Lin Su, Jia Song, Edward Cui, Taroon	652
600	<i>arXiv:1606.08415</i> .	Bharti, and Arun Sacheti. 2020. Imagebert: Cross-	653
601	Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai,	modal pre-training with large-scale weak-supervised	654
602	Jiajing Xu, Jeff Donahue, and Sarah Tavel. 2015. Vi-	image-text data. <i>arXiv preprint arXiv:2001.07966</i> .	655
603	visual search at pinterest. In <i>Proceedings of the 21th</i>	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya	656
604	<i>ACM SIGKDD International Conference on Knowl-</i>	Ramesh, Gabriel Goh, Sandhini Agarwal, Girish	657
605	<i>edge Discovery and Data Mining</i> , pages 1889–1898.	Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,	658
606	Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld,	et al. 2021. Learning transferable visual models	659
607	Luke Zettlemoyer, and Omer Levy. 2020. Spanbert:	from natural language supervision. <i>arXiv preprint</i>	660
608	Improving pre-training by representing and predict-	<i>arXiv:2103.00020</i> .	661
609	ing spans. <i>Transactions of the Association for Com-</i>	Devashish Shankar, Sujay Narumanchi, HA Ananya,	662
610	<i>putational Linguistics</i> , 8:64–77.	Pramod Kompalli, and Krishnendu Chaudhury.	663
611	Sahar Kazemzadeh, Vicente Ordonez, Mark Matten,	2017. Deep learning based large scale visual rec-	664
612	and Tamara Berg. 2014. Referitgame: Referring	ommendation and search for e-commerce. <i>arXiv</i>	665
613	to objects in photographs of natural scenes. In <i>Pro-</i>	<i>preprint arXiv:1703.02344</i> .	666
614	<i>ceedings of the 2014 conference on empirical meth-</i>	Piyush Sharma, Nan Ding, Sebastian Goodman, and	667
615	<i>ods in natural language processing (EMNLP)</i> , pages	Radu Soricut. 2018. Conceptual captions: A	668
616	787–798.	cleaned, hypernymed, image alt-text dataset for au-	669
617	Jack Kiefer, Jacob Wolfowitz, et al. 1952. Stochastic	tomatic image captioning. In <i>Proceedings of the</i>	670
618	estimation of the maximum of a regression function.	<i>56th Annual Meeting of the Association for Compu-</i>	671
619	<i>The Annals of Mathematical Statistics</i> , 23(3):462–	<i>tational Linguistics (Volume 1: Long Papers)</i> , pages	672
620	466.	2556–2565.	673
621	Diederik P Kingma and Jimmy Ba. 2014. Adam: A	Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu,	674
622	method for stochastic optimization. <i>arXiv preprint</i>	Furu Wei, and Jifeng Dai. 2019. VI-bert: Pre-	675
623	<i>arXiv:1412.6980</i> .	training of generic visual-linguistic representations.	676
		<i>arXiv preprint arXiv:1908.08530</i> .	677

- Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. 2017. Visual search at ebay. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2101–2110.
- Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1):69–90.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

A Construction Details of The Index Set

In order to collect both matches and distractors, we first create a candidate pool by randomly sampling from the 25 million dataset, resulting in about 1 million products that covers 1,275 leaf categories. Next we need to annotate all candidates to be either "exact match" or "not-a-match". To reduce the cost caused by complete human labeling, we used a pre-ranking methodology to reduce the size of shortlists for human review. We also employ categorical filtering, similarity-based thresholding and other methods to further improve the recalls. During the crowd-sourced annotation process, given query-candidate pairs, human annotation review both image and title and give several levels of confidence for "exact match". Each query-candidate pair has been rated by at most ten annotators. We performed post processing to determine the final acceptable labels and discarded any labels with low confidence as either "match" or "not match" to minimize the incorrect annotation risk.

Same Product To define two images as including the "same" product is often subjective and challenging especially when critical aspects are missing from their titles and are invisible from images. In our case, we use images as the primary source for annotations. The criteria for "same" product in two product images are:

- The two products can represent different product conditions. For example, a broken phone and a new phone with the exact same specifications (make, model, color, etc.) are the "same" products.
- They should have the same color/model/style, and other aspects that are visually visible and distinguishable. For example, a golden and a gray phone of otherwise the same make and model are not considered the "same" products.
- They can vary in other aspects that are not distinguished solely based on image, e.g. shoe size, memory size for hardware, etc.

B Baseline Models and Finetuning

B.1 Baseline Models

We mainly compare our VT-BERT with other supervised learning models on the above two downstream tasks.

ResNet50 We train ResNet50 (He et al., 2016) with pure image as input for the leaf category pre-

diction on the index set. For the same product recommendation, we use the pooled embedding from the pre-trained ResNet50.

BERT We train BERT_{BASE} (Devlin et al., 2018) with pure title as input for the leaf category prediction on the index set. For the same product recommendation, we use the pooled output from the pre-trained BERT as the embedding.

VT-BERT with Classification Loss We train our VT-BERT for the leaf category prediction instead of our five objectives on the index set. For the same product recommendation, we use the pooled output, i.e. the global embedding, from the pre-trained VT-BERT as the embedding.

CLIP Similar to VT-BERT, CLIP (Radford et al., 2021) learns image and text embeddings as an unsupervised way. We use ResNet50 as the image encoder and BERT_{BASE} as the text encoder for CLIP. The similarity score is computed with the text and image embeddings in a similar way of Equation 9.

B.2 Training Details for Baseline Models and Finetuning

ResNet50 We train ResNet50 for the leaf category prediction task on the index set, with image as the input. The initial learning rate is set to 0.1, then decayed by a factor of 0.1 if the performance on valid set hasn't improved over 5 epochs. We use SGD (Kiefer et al., 1952) with a momentum of 0.9. The image size is 224, with random resizing and horizontal flip as the data augmentation methods. The training is stopped when the performance on the valid set hasn't improved over 10 epochs. The batch size is 256. We use 8 Volta V100 GPUs to complete the training within 1 day.

BERT We train BERT for the leaf category prediction task on the index set, with title as the input. The initial learning rate is warmed up over the first 1000 steps to a peak value of $1e-4$, and then decayed proportionally to the inverse square root of the step number (Vaswani et al., 2017). We set β hyperparameters ($\beta_1 = 0.9$, $\beta_2 = 0.98$) and decoupled weight decay of $1e-4$. We deviate from the optimization by using an ϵ of $1e-8$ for AdamW. The architecture setting stays the same as the original BERT_{BASE}. The training is stopped when the performance on the valid set hasn't improved over 10 epochs. We use a batch size of 256, training on 8 Volta V100 GPUs and completing in 1 day.

CLIP We train CLIP on the index set, with both

image and title as the input. The training setting for CLIP stays the same as the above setting for BERT.

Finetune VT-BERT for The Leaf Category Prediction For the leaf category prediction task, we first pre-train VT-BERT on 10 million dataset, then finetune the model on the index set. The finetuning setting stays the same as the above setting for BERT.

C Examples for the Same Product Recommendation

Figure 5 shows some results of the same product recommendation. Figure 6 shows the zero-shot performance of our pre-trained VT-BERT on the stanford online product (SOP) dataset (Oh Song et al., 2016).

<p>Citizen Eco-Drive Skyhawk A.T. JY0000-5326 U600 Stainless Steel Atomic Men's Watch!</p>	<p>ResNet50</p>	<p>5.500 1000</p>
--	-----------------	---

Figure 5: Top5 examples for the same product recommendation. Products in the blue boxes are matching products, while the other are distractors. From these four query results, One can observe: (1) ResNet tends to find the similar image. Especially for the bottom left example, the brown paper box is a key feature; (2) BERT obtains better results when there are common n-grams in the titles (the top right example); (3) VT-BERT offers diverse “same products”. The retrieved products could have very different backgrounds or different titles.


















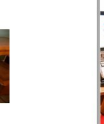






















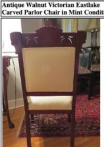
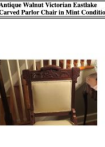
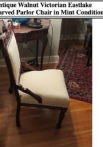


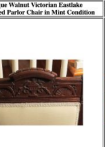
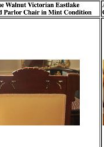
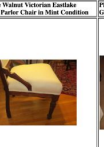


SCHWINN CRUISER BIKE MEN'S "9.5 26" BEACH RIDING CITY COMFORT BIKE PADDLED SADDLE	SCHWINN CRUISER BIKE MEN'S "9.5 26" BEACH RIDING CITY COMFORT BIKE PADDLED SADDLE	SCHWINN CRUISER BIKE MEN'S "9.5 26" BEACH RIDING CITY COMFORT BIKE PADDLED SADDLE	BUFFY CRUISER BIKE 26" MEN'S "9.5 WHITE/LACK TRADITIONAL COMFORT CITY, BEACH BIKE	BUFFY CRUISER BIKE 26" MEN'S "9.5 WHITE/LACK TRADITIONAL COMFORT CITY, BEACH BIKE	SCHWINN Protocol 1.8 Men's Dual Suspension MOUNTAIN BIKE, MEN'S BIKE, RED	SCHWINN Protocol 1.8 Men's Dual Suspension MOUNTAIN BIKE, MEN'S BIKE, RED	SCHWINN Protocol 1.8 Men's Dual Suspension MOUNTAIN BIKE, MEN'S BIKE, RED	BUFFY CRUISER BIKE MEN'S "9.5 26" BLU/FAT TIRE CITY AND BEACH COMFORT BIKE NEW!	BUFFY CRUISER BIKE MEN'S "9.5 26" BLU/FAT TIRE CITY AND BEACH COMFORT BIKE NEW!
									
Korreson Elipse walnut stained adjustable table	Korreson Elipse walnut stained adjustable table	Korreson Elipse walnut stained adjustable table	Korreson task colored corner table	Korreson task colored corner table	Korreson task colored corner table	Vintage Retro Mid-Century Maple Wood Side Table	Vintage Beantail Hammered Coffee Table Wood Cocktail Occasional Open Arm	Vintage wooden "Sailing yacht" table lamp	Vintage wooden "Sailing yacht" table lamp
									
New Portable Small Fan Mini Air Conditioner Personal Compact Cooling Device Blue	New Portable Small Fan Mini Air Conditioner Personal Compact Cooling Device Blue	New Portable Small Fan Mini Air Conditioner Personal Compact Cooling Device Blue	New Portable Small Fan Mini Air Conditioner Personal Compact Cooling Device Blue	Mini Small Fan Cooling Portable Dehumid Dual Blades Air Conditioner USB New	Mini Small Fan Cooling Portable Dehumid Dual Blades Air Conditioner USB New	Mini Small Fan Cooling Portable Dehumid Dual Blades Air Conditioner USB New	Mini Small Fan Cooling Portable Dehumid Dual Blades Air Conditioner USB New	Mini Small Fan Cooling Portable Dehumid Dual Blades Air Conditioner USB New	Mini Small Fan Cooling Portable Dehumid Dual Blades Air Conditioner USB New
									
3 PK Stapler Set-Mini-Stapler/Stapler Remover 4000 No's-15MStaples for Office&HOME	3 PK Stapler Set-Mini-Stapler/Stapler Remover 4000 No's-15MStaples for Office&HOME	3 PK Stapler Set-Mini-Stapler/Stapler Remover 4000 No's-15MStaples for Office&HOME	3 PK Stapler Set-Mini-Stapler/Stapler Remover 4000 No's-15MStaples for Office&HOME	3 PK Stapler Set-Mini-Stapler/Stapler Remover 4000 No's-15MStaples for Office&HOME	BOSTITCH DS-3522 C STAPLER CLINCH BOX CARTON STAPLER CL-100R	BOSTITCH DS-3522 C STAPLER CLINCH BOX CARTON STAPLER CL-100R	Stanley-Smith Flat Clinch Stapler 100 Staples Capacity - Black (90001-A)	Stanley Gun 3 IN 1 W2050 STAPLER AND 2 BOXES OF STAPLING SUPPLIES PERFORMANCE TL	Stanley-Smith Flat Clinch Stapler 100 Staples Capacity - Black (90001-A)
									
Antique Walnut Victorian Enslake Carved Parlor Chair in Mint Condition	Antique Walnut Victorian Enslake Carved Parlor Chair in Mint Condition	Antique Walnut Victorian Enslake Carved Parlor Chair in Mint Condition	Antique Walnut Victorian Enslake Carved Parlor Chair in Mint Condition	Antique Walnut Victorian Enslake Carved Parlor Chair in Mint Condition	Antique Walnut Victorian Enslake Carved Parlor Chair in Mint Condition	Antique Walnut Victorian Enslake Carved Parlor Chair in Mint Condition	Antique Walnut Victorian Enslake Carved Parlor Chair in Mint Condition	Plantation Chair, Can back, Great Condition	Plantation Chair, Can back, Great Condition
									

Figure 6: Zero-shot same product recommendation on SOP test set with the embedding from a pre-trained VT-BERT on VTOP. The products on the first column are the queries. SOP is not a suitable dataset for multi-modal retrieval task, because there are too many products with the same titles, making the search easy. However, looking at the less similar products, we can be still impressed by the performance of the pre-trained VT-BERT.