V2P: Vision-to-Prompt based Multi-Modal Product Summary Generation

Xuemeng Song
Shandong University
Qingdao, China
sxmustc@gmail.com

Liqiang Jing
Shandong University
Qingdao, China
jingliqiang6@gmail.com

Dengtian Lin
Shandong University
Qingdao, China
lindengtian@mail.sdu.edu.cn

Zhongzhou Zhao
Alibaba Group
Hangzhou, China
zhongzhou.zhaozz@alibaba-inc.com

Haiqing Chen
Alibaba Group
Hangzhou, China
haiqing.chenhq@alibaba-inc.com

Liqiang Nie
Shandong University
Qingdao, China
nieliqiang@gmail.com

ABSTRACT

Multi-modal product summary generation is a new yet challenging task, which aims to generate a concise and readable summary for a product given its multi-modal content, *e.g.*, its long text description and image. Although existing methods have achieved great success, they still suffer from three key limitations: 1) *overlook the benefit of pre-training*, 2) *lack the representation-level supervision*, and 3) *ignore the diversity of the seller-generated data*. To address these limitations, in this work, we propose a Vision-to-Prompt based multi-modal product summary generation framework, dubbed as V2P, where a Generative Pre-trained Language Model (GPLM) is adopted as the backbone. We design V2P with two key components: *vision-based prominent attribute prediction*, and *attribute prompt-guided summary generation*. The first works on obtaining the vital semantic attributes of the product from its image by the Swin Transformer, while the second aims to generate the summary based on the product's long text description and the attribute prompts yielded by the first component with a GPLM. Towards comprehensive supervision, apart from the conventional output-level supervision, we introduce the representation-level regularization. Meanwhile, we design the data augmentation-based robustness regularization to handle the diverse inputs and improve the model robustness. Extensive experiments on a large-scale Chinese dataset verify the superiority of our model over cutting-edge methods.

CCS CONCEPTS

• **Information systems** → **Summarization**;

KEYWORDS

Multi-modal Summarization; Product Summary Generation; Pre-trained Language Model

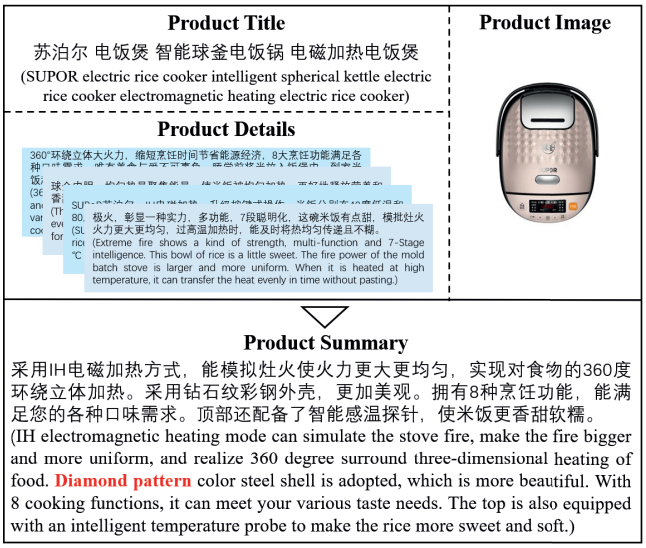

Figure 1: Illustration of the task of multi-modal product summary generation. The English texts are translated from the Chinese texts.

ACM Reference Format:
Xuemeng Song, Liqiang Jing, Dengtian Lin, Zhongzhou Zhao, Haiqing Chen, and Liqiang Nie. 2022. V2P: Vision-to-Prompt based Multi-Modal Product Summary Generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3477495.3532076

1 INTRODUCTION

In E-commerce advertising, each product is usually displayed in a web page with rich data, such as the product's title, detailed description, and image. Although these data provide detailed product information, going through the entire web page increases the cognitive load and time cost for users to judge whether the product meets their purchase needs. Therefore, it is highly desired that the seller can provide a concise product summary to highlight the product characteristics and advantages, so as to save the consumers' time, optimize their consumption experience, and hence boost the product sales. It is, however, intractable to manually

*Liqiang Nie (nieliqiang@gmail.com) is the corresponding author.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532076

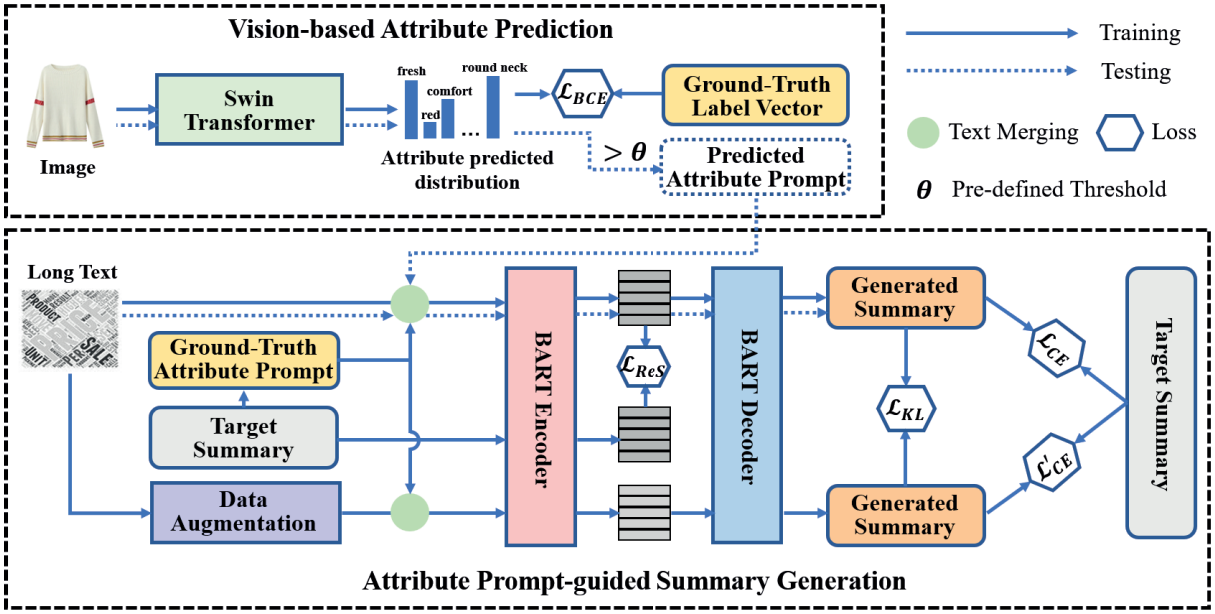

Figure 2: The proposed V2P scheme, which consists of with two key components: *Vision-oriented Attribute Prediction*, and *Attribute Prompt-guided Summary Generation*.

write summaries for numerous products, leaving alone that writing concise summaries for products needs certain expertise. In light of this, automatic product summary generation gains increasing research attention.

Early studies [2, 35] mainly fulfill the product summarization purely relying on the product's textual modality, such as the title and attribute values. Despite their promising performance, these efforts overlook the visual modality (*i.e.*, the image) of each product, which indeed conveys important cues regarding the product's features. Therefore, a few pioneer studies [14, 38] resorted to the Multi-modal Product Summary Generation (MPSG). As illustrated in Figure 1, MPSG aims to generate a textual summary for a product based on its multi-modal contents, such as the long text description (including the brief title and other detailed descriptions) and image. Although the pioneer studies have achieved significant progress, they mainly suffer from three key limitations.

- **L1: Overlook the benefit of pre-training.** Previous methods follow the conventional train-from-scratch paradigm, and fail to take advantage of the pre-training technique, especially the powerful text generation capability of Generative Pre-trained Language Models (GPLMs), which have shown compelling success in various natural language generation tasks [39, 41].
- **L2: Lack the representation-level supervision.** Existing methods mainly focus on the overall output-level supervision, *i.e.*, using the cross-entropy loss to regulate the generated summary to be similar to the ground-truth one, which entangles the optimization of the whole model. Beyond that, we argue that the representation-level supervision, which can propel the model to learn the meaningful representation of the input towards the product summary generation, also merits our special attention.
- **L3: Ignore the diversity of the seller-generated data.** In practice, the long text descriptions of products are often seller-generated, and it is thus likely that for the same product, different sellers could provide different long textual descriptions. This requires the model to be able to correctly generate the product summary with even disturbed input textual description. Existing methods overlook this issue, and hence cannot guarantee the robustness of the model.

To address these limitations, as shown in Figure 2, we propose an effective and robust Vision-to-Prompt based multi-modal product summary generation scheme, V2P for short, where the generative pre-trained language model BART is adopted as the backbone. In particular, we design V2P with two key components: *vision-based prominent attribute prediction*, and *attribute prompt-guided summary generation*. The first component, implemented with the Swin Transformer [18], works on predicting the to-be-summarized semantic attributes of each product from its image, while the second one with BART as backbone, concentrates on generating the product summary given its long text description and attribute prompts. The underlying philosophy is three-fold. 1) As the saying goes, a picture is worth a thousand words. As shown in Figure 1, given the image of the product, we can infer the product attributes, like "Diamond pattern", which benefits the product summary generation. 2) We believe that the text phrases of the semantic attributes derived from the product's image should be more helpful towards the summary generation, as compared with the heterogeneous visual features extracted by Convolutional Neural Networks (CNNs). And 3) by converting the product's image into the semantic attribute prompts, we can project the image of the product into the same text embedding space of the original pre-trained language model, which facilitates the maintenance of the original text generation capability of the

pre-trained language model with no extra parameter introduced. Besides, it is worth mentioning that apart from the conventional output-level supervision over the second component, we also introduce the representation-level supervision by maximizing the mutual information between the latent representations of the input and the target output. Moreover, we resort to data augmentation to make the input text description more diverse and hence improve the model robustness, where we also use Kullback Leibler (KL) Divergence to encourage the generated text for the augmented data to be similar to that for the original one.

Notably, these two components are trained separately to avoid error accumulation. In particular, when training the first component, we pre-define an attribute vocabulary based upon our dataset, and then according to that derive the ground-truth attribute labels for each product from its target summary. Meanwhile, for optimizing the second one, we directly use the ground-truth attribute labels of each product as the input attribute prompts. Once the two components are well-trained, we seamlessly integrate them for testing by replacing the ground-truth attribute labels in the second component with the ones predicted by the first component. Ultimately, to justify the proposed scheme, we conduct extensive experiments on a real-world large-scale Chinese dataset, consisting of around 1.4 million products over three different categories (*i.e.*, Home Appliance, Clothing, and Cases&Bags). The experimental results show that our model significantly outperforms the best baseline by 16.42% in terms of Rouge-2 over the Clothing category.

Our main contributions can be summarized in threefold:

- We design a vision-to-prompt based multi-modal product summary generation scheme, where the heterogeneous multi-modal data are subtly unified in the same embedding space by converting the vision modality into semantic attribute prompts, and the original text generation capability of the generative pre-trained language model can be thus well maintained.
- As far as we know, we are the first to perform both output-level and representation-level supervision towards MPSG simultaneously, and introduce the data augmentation-based robustness regularization to cope with the diverse seller-generated product data.
- To the best of our knowledge, we are the first to adapt the generative pre-trained language model to solve the MPSG task, and achieve superior performance over existing state-of-the-art methods on a large-scale real-world dataset with around 1.4 million Chinese products. As a byproduct, we have released the codes and involved parameters to benefit the research community[1].

2 RELATED WORK

Our work is related to E-commerce product summarization and pre-trained language models.

2.1 E-Commerce Product Summarization

E-commerce product summarization aims to generate a concise, short and readable text that contains the most salient information for a product, to save consumers' time and improve their consumption experience. Early researches mainly focus on generating the product summary based on the product's textual descriptions [2, 8, 15, 35, 40]. For example, Xiao et al. [32] proposed two product summarization approaches to summarize short titles of products via bi-directional long short-term memory encoder-decoder network. In addition, Khatri et al. [11] introduced a novel Document-Context based Seq2Seq model for abstractive and extractive summarizations in the E-commerce settings. In spite of their promising performance, these methods ignore the product's image, which also delivers important cues regarding the product characteristics and benefits the product summarization. To address this issue, some studies [14, 38] resort to the multi-modal product summarization. For example, Zhang et al. [38] proposed a multi-modal generative adversarial network for the short product title generation, which simultaneously explores the visual and textual information of the product. Due to the concern that different products have different selling points, Li et al. [14] presented an aspect-aware multi-modal summarization model that is able to determine the most salient aspects of a product based on its textual description and image.

Although these methods have achieved compelling performance, they suffer from three key limitations: overlook the benefit of pre-training, lack the representation-level supervision, and ignore the diversity of the seller-generated data, which are the major concerns of our work.

2.2 Pre-trained Language Models

Recently, pre-training has become a popular paradigm in Natural Language Processing (NLP) community, which has shown compelling success in many NLP tasks. Early pre-trained models, such as Word2vec [19] and GloVe [20], mainly adopt the shallow architecture and aim to yield good pre-trained word embeddings for various NLP tasks. Since most NLP tasks are beyond word-level, modern pre-training aims to output a general backbone model rather than word embeddings. Due to the superior performance of Transformer [29], a standard encoder-decoder architecture, increasing research efforts [5, 13, 22, 37] have been dedicated to devising Transformer-based pre-trained models. For example, Devlin et al. [5] proposed the deep Bidirectional Encoder Representation from Transformer (BERT), which is optimized by two pre-training tasks: masked language model and next sentence prediction. Despite the great success of BERT in textual representation learning [27], it cannot be directly fine-tuned for language generation tasks. Later, Lewis et al. [13] presented a denoising pre-training sequence-to-sequence model, named BART, which combines bidirectional and auto-regressive transformers, for natural language generation, translation, and comprehension. Meanwhile, Raffel et al. [23] introduced a unified Text-to-Text Transfer Transformer (T5), which converts all NLP tasks into a "text-to-text" format and thus can be used for various downstream NLP tasks, including question answering, document summarization, and sentiment classification.

With the advances of generative pre-trained language models, more studies tend to adapt publicly available pre-trained language models that have absorbed rich knowledge from large-scale corpus to solve their specific tasks. For example, Yu et al. [36] proposed

[1] https://xuemengsong.github.io/V2P_Code.rar.

the vision guided generative pre-trained language models based on BART and T5 for summarizing videos into short texts based on their visual modalities and textual transcripts. Inspired by this, we also resorted to the GPLM, to enhance the product summary generation capability of our model. Different from the work [36] that directly incorporates visual features in the middle layer of the pre-trained model, we mapped the visual information to attribute prompts, to eliminate the modality gap between the input of our downstream task and that of the pre-trained language models.

3 PROBLEM FORMULATION

Suppose that we have a set of N products $\mathcal{P} = \{p_1, p_2, \cdots, p_N\}$, where each product $p_i = (X_i, V_i, Y_i)$ involves three elements. $X_i = \{x_1^i, x_2^i, \cdots, x_{M_i}^i\}$ and V_i are the long text description and image of the i-th product, respectively, where x_j^i refers to the j-th token in the long text description. M_i stands for the total number of tokens, which is a variable for different products. $Y_i = \{y_1^i, y_2^i, \cdots, y_{L_i}^i\}$ denotes the ground truth summary of the product p_i, consisting of L_i tokens. Based on these training samples, we aim to learn a multi-modal product summary generation model that is able to generate a brief summary for an arbitrary product given its long text description and image.

Beyond existing studies, we resort to the GPLM BART as the backbone of the product summary generator. As a major novelty, our model is designed with two coherent components: *vision-based prominent attribute prediction*, and *attribute prompt-guided summary generation*, which is able to maintain the original text generation capability of BART by converting the visual modality into semantic attribute prompts. In particular, the former works on extracting the prominent semantic attributes of the product image that are more likely to appear in the target summary, as prompts to guide the following summary generation. The latter is responsible for generating the summary given the product's long text description and attribute prompts with BART. It is worth noting that for avoiding error accumulation, we directly feed the ground-truth attributes as the prompts rather than the ones predicted by the first component to train the second component.

In particular, we first pre-define an attribute vocabulary based on all the product summaries in our dataset, *i.e.*, a set of attributes that tend to appear in products' summaries, denoted as $\mathcal{S} = \{a^1, a^2, ..., a^S\}$, where a^j is the j-th attribute value in the attribute vocabulary. We then can derive a set of ground-truth attributes for each product by checking whether each attribute value a^j appears in the product's target summary. Let $\mathcal{A}_i = \{a_i^{k_1}, a_i^{k_2}, \cdots, a_i^{k_{Q_i}}\}$ denote the set of ground truth attributes of the product p_i, where $a_i^{k_*} \in \mathcal{S}$, and Q_i is the total number of ground-truth attributes. Then the two key components of our model can be formulated as follows, respectively:

$$\begin{cases} \mathcal{G}(V_i|\mathbf{\Theta}_a) \rightarrow \mathcal{A}_i \\ \mathcal{H}(X, \mathcal{A}_i|\mathbf{\Theta}_g, \mathcal{T}) \rightarrow Y, \end{cases} \tag{1}$$

where $\mathcal{G}(\cdot)$ and $\mathcal{H}(\cdot)$ refer to the vision-based prominent attribute prediction model, and attribute prompt-guided summary generation model, respectively, while $\mathbf{\Theta}_a$ and $\mathbf{\Theta}_g$ stand for their corresponding to-be-learned parameters. $\mathcal{T} = \{t_1, t_2, \cdots, t_{|\mathcal{T}|}\}$ denotes the token vocabulary for product summary generation. It is worth noting that the vision-based prominent attribute prediction model $\mathcal{G}$ is trained by the dataset $\mathcal{D}_1 = \{(V_i, \mathcal{A}_i)|i = 1, \cdots, N\}$, while the attribute prompt-guided summary generation model $\mathcal{H}$ is optimized by the dataset $\mathcal{D}_2 = \{(X_i, \mathcal{A}_i, Y_i)|i = 1, \cdots, N\}$. Once the two components get well trained, during the testing phase, we can first utilize the vision-based prominent attribute prediction model G to predict the to-be-summarized attributes of the product from its image, and then feed them as prompts to the model $\mathcal{H}$ to output the final product summary $\hat{Y}$ as follows,

$$\hat{Y} = \mathcal{H}(X, \widehat{\mathcal{A}}_i|\mathbf{\Theta}_g, \mathcal{T}), \tag{2}$$

where $\widehat{\mathcal{A}}_i$ is the set of predicted attribute prompts from the i-th product image by $\mathcal{G}$.

4 METHOD

In this section, we detail the proposed V2P, which consists of two coherent components: *vision-based prominent attribute prediction*, and *attribute prompt-guided summary generation*. For clarity, we temporally omit the subscript i of the training samples.

4.1 Vision-based Prominent Attribute Prediction

To promote the multi-modal product summary generation performance by GPLM, instead of directly encoding the visual modality with advanced CNNs, we propose to extract the prominent semantic attributes (*i.e.*, the ones tend to appear in the product summary) of each product from its image, and use their semantic embeddings to prompt the product summary generation. In a sense, the prominent attribute prediction can be cast as a task of multi-label image classification. Towards this end, due to the remarkable performance of Transformer in many visual understanding tasks [4, 30, 33], we implement our vision-based prominent attribute prediction component $\mathcal{G}$ with the pre-trained Swin Transformer as follows,

$$\hat{\mathbf{s}} = Swin(V), \tag{3}$$

where $\hat{\mathbf{s}} = [\hat{s}^1, \hat{s}^2, \cdots, \hat{s}^S] \in \mathbb{R}^S$ denotes the predicted attribute probability distribution for the product image V.

Towards the optimization of this component, we assign each product a multi-hot vector as the ground-truth attribute label vector, denoted as $\mathbf{s} = [\mathbf{s}^1, \mathbf{s}^2, \cdots, \mathbf{s}^S] \in \{0, 1\}^S$, where $\mathbf{s}^k = 1$, if the k-th attribute value in the attribute vocabulary is a ground-truth one of the given product, *i.e.*, $a^k \in \mathcal{A}$, and $\mathbf{s}^k = 0$ otherwise. We then use the ground-truth attribute label vector $\mathbf{s}$ to fine-tune the Swin Transformer with the binary cross-entropy loss below,

$$L_{BCE} = \min_{\mathbf{\Theta}_a} -[\mathbf{s} \ln(\sigma(\hat{\mathbf{s}})) + (\mathbf{1} - \mathbf{s}) \ln(\mathbf{1} - \sigma(\hat{\mathbf{s}}))], \tag{4}$$

where $\mathbf{1} \in \mathbb{R}^S$ is the all-one vector, and $\sigma(\cdot)$ refers to the Sigmoid function. Notably, this loss function is defined for a single sample. In practice, we will optimize over mini-batches of samples.

Once this component gets well trained, we employ a threshold θ to determine the final predicted attributes as follows,

$$\widehat{\mathcal{A}} = \{a^k|\hat{s}^k \geq \theta, k = 1, \cdots, S\}, \tag{5}$$

where $\widehat{\mathcal{A}}$ denotes the set of predicted attributes of the given product based on its image, which can be used as prompts for the subsequent product summarization.

4.2 Attribute Prompt-guided Summary Generation

This component aims to generate the product's summary based on the product's long text description and attribute prompts. Towards this end, we devise this component with four key modules: *Text and Attribute Prompt Embedding*, *BART-based Summary Generation*, *Output-level and Representation-level Supervision*, and *Data Augmentation-based Robustness Regularization*.

4.2.1 Text and Attribute Prompt Embedding. Firstly, we introduce the embedding of the input, *i.e.*, the long text description and attribute prompts. As aforementioned, to avoid error accumulation, we employ the ground-truth attributes instead of the predicted ones as prompts for training the second component of our model. Considering the long text description and attribute prompts come from the same token space, we directly treat them as a whole, that it, merge them together. Let $\widehat{X} = [X, a^{k_1}, a^{k_2}, \cdots, a^{k_Q}]$ denote the merged text modality of the product p, which is finally composed of U tokens. We then use the embedding layer of BART to encode the merged text modality of the product. To be specific, we embed each token of $\widehat{X}$ with a linear transformation as follows,

$$\mathbf{e}^j = \mathbf{W}^T \mathbf{g}^j, j = 1, \cdots, U, \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{|\mathcal{T}| \times D}$ is the token embedding matrix to be fine-tuned, which is initialized according to the pre-trained BART. Intuitively, the i-th row of $\mathbf{W}$ refers to the representation of the i-th token in the token vocabulary $\mathcal{T}$.$\mathbf{g}^j \in \mathbb{R}^{|\mathcal{T}|}$ is the one-hot vector indicating the index of the j-th token of $\widehat{X}$ in the token vocabulary, while $\mathbf{e}^j \in \mathbb{R}^D$ denotes the embedding of the j-th token in $\widehat{X}$. D is the dimensionality of the token embeddings.

Then to make use of the order information among input tokens, we also introduce the positional encoding [29] of BART to derive the final embedding of the input (*i.e.*, the long text description and attribute prompts) as follows,

$$\mathbf{E} = [\mathbf{e}^1; \mathbf{e}^2; \cdots; \mathbf{e}^U] + \mathbf{E}_{pos}, \tag{7}$$

where $\mathbf{E} \in \mathbb{R}^{U \times D}$ is the final input embedding, and $\mathbf{E}_{pos} \in \mathbb{R}^{U \times D}$ is the positional embedding. $[;]$ denotes the concatenation operation.

4.2.2 BART-based Summary Generation. We then employ the pre-trained BART to fulfil the product summary generation. In particular, we first feed the input embedding into the encoder $\mathcal{E}$ of the pre-trained BART as follows,

$$\mathbf{Z} = \mathcal{E}(\mathbf{E}), \tag{8}$$

where $\mathbf{Z} \in \mathbb{R}^{U \times d}$ is the encoded representation of the product's long text description and attribute prompts. d is the dimensionality of the encoded representation. We then feed the encoded representation $\mathbf{Z}$ to the decoder $\mathcal{D}$ of the pre-trained BART, which works in an auto-regressive manner, namely, producing the next word by considering all the previously decoded outputs as follows:

$$\hat{\mathbf{p}}_\mathbf{j} = \mathcal{D}(\mathbf{Z}, \hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{j-1}), \tag{9}$$

where $\hat{\mathbf{p}}_\mathbf{j} \in \mathbb{R}^{|\mathcal{T}|}$ is the predicted token distribution for the j-th token of the generated summary. According to the largest element of $\hat{\mathbf{p}}_\mathbf{j}$, we can derive the j-th token of the generated summary, denoted as $\hat{y}_j$. The decoding process will terminate when the "End-of-Sequence" token is generated.

4.2.3 Output-level and Representation-level Supervision. To enhance the model generalization, We jointly conduct the output-level and representation-level supervision.

Output-level supervision. To supervise the summary generation, following previous studies [7, 39], we adopt the standard cross-entropy loss to fulfill the output-level supervision as follows,

$$\mathcal{L}_{CE} = -\frac{1}{L} \sum_{j=1}^{L} log(\hat{\mathbf{p}}_j[t*]), \tag{10}$$

where $\hat{\mathbf{p}}_j[t*]$ refers to the element of $\hat{\mathbf{p}}_j$ that corresponds to the j-th token of the ground truth summary Y. L is the total number of tokens in Y. Notably, this loss is defined for a single sample.

Representation-level Supervision. As for the representation-level supervision, we resort to maximizing the mutual information, which has shown great success in representation learning [1, 10]. In particular, we aim to maximize the mutual information between the input data and target summary, to encourage the model to encode the input properly, namely, capture the useful information of the input towards the target summary generation. In particular, we first embed the target summary Y into a continuous vector with the same embedding layer and encoder in Eqn. (7) and Eqn. (8), used for the input. Let $\widetilde{\mathbf{Z}} \in \mathbb{R}^{L \times d}$ denote the encoded representation of the target summary Y. We then regard the encoded representation of the input (*i.e.*, $\mathbf{Z}$) and that of the target summary (*i.e.*, $\widetilde{\mathbf{Z}}$) as an instance of the random variables Z and $\widetilde{Z}$, respectively. Accordingly, we expect that the mutual information between Z and $\widetilde{Z}$ should be as large as possible.

As it is intractable to directly calculate the mutual information between Z and $\widetilde{Z}$, we resort to maximizing the commonly-used lower bound [3, 21], *i.e.*, minimizing the objective function below,

$$\mathcal{L}_{ReS} = \mathbb{E}_{p(z_{1:B})p(\tilde{z}|z_1)}\Big[- log \frac{e^{f(z^y, z_1)}}{\sum_{i=1}^{B} e^{f(z^y, z_i)}}\Big], \tag{11}$$

where $z_{1:B}$ are the batch of B samples from p_Z, while $\tilde{z}$ is a sample from $p_{\widetilde{Z}}$ associated with z_1. We treat $(z_1, \tilde{z})$ as a positive pair, and $(z_i, \tilde{z})$ $(i = 2, \cdots, B)$ as negative. $f(x, y)$ is the real value function that measures the similarity between x and y, defined as follows,

$$f(\tilde{z}, z_i) = \tilde{\mathbf{z}}^T \mathbf{z_i} / \tau, \tag{12}$$

where $\tilde{\mathbf{z}} = norm(avg(\widetilde{\mathbf{Z}}))$ and $\mathbf{z_i} = norm(avg(\mathbf{Z_i}))$ are the normalized representations of samples $\tilde{z}$ and z_i, respectively. $\mathbf{Z_i}$ denotes the encoded input representation of the i-th sample in the batch. $norm(\cdot)$ and $avg(\cdot)$ refer to the l_2 normalization and average pooling operation along the column dimension, respectively. τ is the temperature parameter.

4.2.4 Data Augmentation-based Robustness Regularization. To deal with the diverse seller-generated long text descriptions of products, we resort to data augmentation, which works on deriving new and realistic-looking training samples from existing ones by altering their inputs, while keeping their labels unchanged. It has

Algorithm 1 The Training Procedure of V2P.

Input: training set $\mathcal{D}_1$ for optimizing the vision-based prominent attribute prediction model $\mathcal{G}$, training set $\mathcal{D}_2$ for optimizing the attribute prompt-guided summary generation model $\mathcal{H}$, the token vocabulary $\mathcal{T}$, hyperparameters $\{\lambda, \beta, \tau\}$.
Output: Parameters Θ_a, Θ_g.

```
Initialize parameters: Θ_a, Θ_g
repeat
    Randomly sample a batch of (V, A)'s from D_1.
    Update Θ_a by optimizing the loss function in Eqn. (4).
until G converges.
repeat
    Randomly sample a batch of (X, A, Y)'s from D_2.
    for each sample (X, A, Y) do
        Generate an augmented sample (X', A, Y).
    end for
    Update Θ_g by optimizing the loss function in Eqn. (15).
until H converges.
```

shown to be effective in promoting the model's robustness in various tasks [16, 28, 34]. In our context, the underlying philosophy for introducing data augmentation is to force the model to learn how to generate the summary correctly even the input text description is somehow disturbed. In particular, we adopt Easy Data Augmentation (EDA) [31] method that has shown remarkable performance in NLP tasks [16, 34]. According to EDA, we randomly choose and perform only one of the following operations: a) randomly replace a few words in the long text description with their synonyms, b) randomly insert a few words into the long text description, c) randomly swap the positions of two words in the long text description, and d) randomly delete a few words from the long text description. Ultimately, for each sample $(X, \mathcal{A}, Y)$ in our training dataset $\mathcal{D}_2$, we can derive a new corresponding training sample $(X', \mathcal{A}, Y)$ by perturbing X.

Following the same processing for the original data, we can obtain the latent input representation of the augmented data as $\mathbf{Z}'$, and the predicted token distribution of the j-th token of the generated summary text as $\hat{\mathbf{p}}'_j$. Then similar to the original dataset, we also adopt the cross-entropy loss to supervise the summary generation for the augmented data as follows,

$$\mathcal{L}'_{CE} = -\frac{1}{L}\sum_{j=1}^{L} log(\hat{\mathbf{p}}'_j[t*]), \tag{13}$$

where $\hat{\mathbf{p}}'_j[t*]$ refers to the element of $\hat{\mathbf{p}}'_j$ that corresponds to the j-th token of the target summary Y.

Moreover, to enhance the model robustness, we expect that the model can generate a similar summary for the augmented data $(X', \mathcal{A})$ as compared with that of the original data (X, A). Towards this end, we employ the KL Divergence to encourage the consistency between the generated summaries of the original and augmented data as follows,

$$\mathcal{L}_{KL} = \sum_{j=1}^{L} D_{KL}(\hat{\mathbf{p}}_j \| \hat{\mathbf{p}}'_j) = \sum_{j=1}^{L}\sum_{k=1}^{L} \hat{p}_{jk} log(\frac{\hat{p}_{jk}}{\hat{p}'_{jk}}). \tag{14}$$

Table 1: Dataset Statistics. The unit of the average input/output length is the Chinese character.

Category	Home Appliances	Clothing	Cases&Bags
#Train Sample	437,646	790,297	97,510
#Valid Sample	10,000	10,000	5,000
#Test Sample	10,000	10,000	5,000
Avg Input Len	335	286	299
Avg Output Len	79	78	79

4.2.5 Training and Testing. Ultimately, the overall objective function for optimizing the attribute prompt-guided summary generation component can be written as follows,

$$\mathcal{L} = \min_{\Theta_g} \mathcal{L}_{CE} + \lambda \mathcal{L}_{ReS} + \beta(\mathcal{L}'_{CE} + \mathcal{L}_{KL}), \tag{15}$$

where λ and β are the non-negative hyper-parameters.The overall procedure of the optimization is briefly summarized in Algorithm 1. As aforementioned, during testing, we replace the ground-truth attributes in this component with the ones (*i.e.* $\widehat{\mathcal{A}}$) predicted by the former component. The predicted attributes would be ordered according to their predicted probabilities (*i.e.* $\hat{s}^k$'s).

5 EXPERIMENT

In this section, we present the extensive experiments we conducted to answer the following research questions:

- **RQ1.** Does our V2P outperform state-of-the-art methods?
- **RQ2.** What is the contribution of each component of V2P?
- **RQ3.** What is the intuitive performance of our V2P?
- **RQ4.** Is our V2P sensitive to the key hyperparameters?

5.1 Experimental Setting

Dataset. In this work, we adopted the Chinese E-commerce product summarization dataset CEPSUM [14], which consists of around 1.4 million products covering three coarse categories: Home Appliances, Clothing, and Cases&Bags. The detailed statistics of this dataset are given in Table 1. Each product has a long text description, an image, as well as a corresponding summary, written by a qualified writer.

Attribute Vocabulary Construction. To obtain the pre-defined attribute vocabulary, we first employed Jieba[2] to tokenize the dataset and perform the part-of-speech tagging. We then only retained the nouns and adjectives that have at least two Chinese characters to constitute the attribute vocabulary for each category. Notably, we built different attribute vocabularies for products of different categories, due to the fact that products of different categories tend to be summarized by different attributes. For illustration, Figure 3 shows the word clouds over ground-truth attribute values for characterizing products of different categories. As can be seen, home appliances are usually described by "Smart" and "Effective", clothing products by "Fashion" and "Fabric", while cases and bags by "Material" and "Cowhide". To ensure the quality of the attribute vocabulary, we filtered out the nouns/adjectives that have appeared in less than a pre-defined number of products' summaries. Specifically, according to the scale of different product categories, we set the pre-defined number threshold for the Home

[2]https://github.com/fxsjy/jieba.

(a) Home Appliance. (b) Clothing. (c) Cases&Bags.

Figure 3: Word clouds over the extracted attribute values for products of different categories.

Table 2: Performance (%) comparison among different methods on three product categories. The best results are in boldface, while the second best are underlined. * denotes that the p-value of the significant test between our result and the best baseline result is less than 0.01. "Improvement . ↑" refers to the relative improvement by our model over the best baseline result.

Model	Home Appliance			Clothing			Cases&Bags		
	Rouge-1	Rouge-2	Rough-L	Rouge-1	Rouge-2	Rough-L	Rouge-1	Rouge-2	Rough-L
Lead	21.97	9.54	12.79	19.83	8.39	13.56	21.49	9.37	14.19
LexRank	24.06	10.01	18.19	26.87	9.01	17.76	27.09	9.87	18.03
Seq2Seq	21.57	7.18	17.61	23.05	6.84	16.82	23.18	6.94	17.29
MASS	28.19	8.02	18.73	26.73	8.03	17.72	27.19	9.03	18.17
PG	31.31	10.93	21.11	29.11	9.24	19.92	31.11	10.27	21.79
MMPG	32.88	11.88	21.96	30.73	10.29	21.25	32.69	11.78	22.27
VG-BART (Dot-product)	32.71	11.46	22.87	31.36	9.94	21.34	32.73	10.26	22.44
VG-BART (Multi-head)	32.73	11.74	23.61	31.63	10.08	21.53	33.30	11.31	23.13
V2P	**34.47***	**12.63***	**25.09***	**35.05***	**11.98***	**22.62***	**34.65***	**11.89***	**24.53***
Improvement. ↑	4.84%	6.32%	6.27%	10.81%	16.42%	5.06%	4.05%	0.93%	6.05%

Appliance, Clothing, and Cases&Bags categories, as 5,000, 10,000, and 1,000, respectively.

Implementation Details We used the BART provided by Taobao, which is pre-trained by a large-scale Chinese corpus collected from the Weitao[3] platform. For optimization, we adopted the Adam [12] with the fixed learning rate of 5e-5. We employed the grid search strategy to determine the optimal values for the hyperparameters θ, τ, λ, and β. In particular, we searched the threshold θ among values of $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. As for the other three hyperparameters, we first coarsely searched the among values of $\{0.001, 0.01, 0.1, 1, 10, 100\}$, and then finely tuned them at the step of 0.05. Ultimately, the optimal values of θ, τ, λ, and β are 0.2, 1, 0.05, and 0.1, respectively. The batch size is set to 16. The dimension of the token embedding D = 1024, and that of the encoded representation d = 768. The image size is unified to 224×224. Finally, we adopted the widely-used character-based ROUGE-1, ROUGE-2, and ROUGE-L [17], as evaluation metrics. Notably, all our experiments are conducted by 5 times, and the average performance on the testing set is reported for comparison.

5.2 On Model Comparison (RQ1)

To justify our V2P, we compared it with the following baselines.

- **Lead. [14]** This method directly deems the first P characters of the long text description as the summary. According to Table 1, we set P = 80.
- **LexRank [6].** This is an extractive text summarization method, which measures the sentence salience based on the graph-based centrality.
- **Seq2seq [11].** This is a dominant framework used for natural language generation tasks.
- **Pointer-Generator[24].** This is a hybrid model, which consists of a pointer network and a seq2seq based generator. The former facilitates the reproduction of information by copying words from the source text, while the latter enables the new word generation.
- **MASS [26].** This is a masked seq-to-seq pre-training method for encoder-decoder based language generation.
- **MMPG [14].** This is a multi-modal pointer-generator network for multi-modal product summarization, where the image is encoded by convolutional neural networks.
- **VG-BART (Dot-product) and VG-BART (Multi-head) [36].** These two methods employ BART as the backbone for multi-modal abstractive summarization, and incorporate the visual modality by the dot-product attention-based and multi-head attention based add-on layers over BART, respectively.

Table 2 shows the performance comparison among different methods over products of three categories (*i.e.,* Home Appliance, Clothing, and Cases&Bags) in terms of Rouge-1, Rouge-2, and Rouge-L. From this table, we have the following observation. 1) Our model consistently outperforms the other baseline methods over different product categories. To justify the improvement is statistically significant, we also conducted the significant test between our results and the second best results, and found that all the p-values are less than 0.01. This validates the superiority of our V2P over existing methods. 2) Our V2P surpasses both VG-BART (Dot-product) and VG-BART (Multi-head). This implies the advantage of injecting the visual modality via the attribute prompt extraction rather than the attention-based add-on layers. 3) Unexpectedly, although VG-BART (Dot-product) and VG-BART

[3]https://tinyurl.com/4h9fznkn.

Table 3: Ablation study results (%) of our proposed V2P. The best results are highlighted in boldface.

Model	Home Appliance			Clothing			Cases&Bags		
	Rouge-1	Rouge-2	Rough-L	Rouge-1	Rouge-2	Rough-L	Rouge-1	Rouge-2	Rough-L
V2P	**34.47**	**12.63**	**25.09**	**35.05**	**11.98**	**22.62**	**34.65**	**11.89**	**24.53**
V2P-w/o-Image	32.78	11.38	24.26	30.71	9.67	20.96	32.95	10.53	23.17
V2P-w/o-ReS	33.74	11.73	24.64	34.81	11.67	21.76	33.77	11.11	23.48
V2P-w/o-Robust	34.22	12.39	24.76	34.62	11.47	21.10	33.19	10.02	22.56
V2P-w-VGG	32.90	11.67	23.92	34.31	11.60	21.07	33.46	10.99	23.72
V2P-w-Res	33.15	11.90	24.12	34.49	11.67	21.13	19.44	4.58	13.34

Table 4: Performance of different models on vision-based prominent attribute prediction in terms of Precision (%), Recall (%), and F1-score (%), where the threshold (θ) for determining the predicted attributes of each product is set to 0.2**.**

Model (θ=0.2)	Home Appliance			Clothing			Cases&Bags		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Swin	18.53	**33.95**	**21.65**	**16.34**	**16.79**	**13.69**	7.36	10.85	**7.54**
ResNet101	17.35	22.26	17.06	14.32	15.20	12.68	3.97	**99.75**	7.33
VGG19	**21.25**	20.82	17.75	16.24	14.67	12.71	**8.21**	7.77	6.15

(Multi-head) are equipped with generative pre-trained language model, they still cannot surpass MMPG that does not adopt the generative pre-trained language model as the summary generation backbone in terms of several metrics. This sheds light on the importance of devising an effective vision injection scheme to fully take advantage of the pre-trained language model in the context of multi-modal product summary generation. And 4) our model achieves the largest improvement over the product category of Clothing. This may be due to that as compared with the home appliances, cases, and bags, the visual modality of clothing products contains more straightforward cues of products, and effectively exploring it can largely improve the model performance.

5.3 On Ablation Study (RQ2)

We introduced the following variant methods for ablation study.

- **V2P-w/o-Image**. To demonstrate the effect of the product image in summarizing the product, we designed this method that only uses the long text description of the product to generate the product summary. Namely, we only fed the long text description into the BART backbone.
- **V2P-w-VGG** and **V2P-w-Res**. To show the benefit of using Swin Transformer as the vision-based attribute prediction model, we replaced the Swin Transformer in our model with the VGG [25] and ResNet [9], respectively.
- **V2P-w/o-ReS**. To show the necessity of the representation-level supervision, we removed it from our V2P by setting $\lambda = 0$ in Eqn. (15).
- **V2P-w/o-Robust**. To verify the importance of the data augmentation-based robustness regularization, we omitted this module by setting $\beta = 0$ in Eqn. (15).

Table 3 shows the ablation study results of our proposed V2P over the three product categories in terms of Rouge-1, Rouge-2, and Rouge-L. From this table, we have the following observations. 1) Our V2P consistently outperforms V2P-w/o-Image over different product categories. This verifies the importance of taking into account the visual modality of products for their summary generation. Similarly, the improvement by our V2P over the Clothing category is the most significant, which implies that the images of clothing products do contain rich cues regarding the product properties again. 2) V2P exceeds both V2P-w/o-ReS and V2P-w/o-Robust, across different evaluation metrics, which indicates the necessity of the representation-level supervision and the data augmentation-based robustness regularization. And 3) V2P-w-VGG and V2P-w-Res perform worse than our V2P, which suggests the superior capacity of the Swin Transformer in the attribute classification of images. In fact, we also checked the performance of the Swin Transformer, VGG, and ResNet in our context of vision-based prominent attribute prediction. Table 4 shows the performance of different methods in terms of Precision, Recall, and F1-score. As can be seen, Swin Transformer performs superior over the VGG and ResNet across different product categories in terms of F1-score. This indicates that Swin Transformer well balances the precision and recall of the predicted prominent attributes.

5.4 On Case Study (RQ3)

To get an intuitive understanding of the product summary generation capability of our model, due to the limited space, we show a testing result of our model and its variant V2P-w/o-image that generates the product summary only based on the given long text description in Figure 4. As can be seen, the Rouge scores of our V2P is significantly higher than those of its variant V2P-w/o-Image. Looking into the generated summaries, we can learn that by incorporating the product's image, our V2P is able to capture the product's attributes, *e.g., cardigan, right spun, high softness, and slim*, while V2P-w/o-Image cannot. This intuitively demonstrates the necessity of considering the visual modality of the product.

5.5 On Sensitivity Analysis (RQ4)

We first checked our model's sensitivity towards the core hyperparameter in the first component, namely, the threshold θ for determining the prominent attributes. Due to the limited space, we show the performance of our model with different θ over the product category Cases&Bags in Table 5. As can be seen, the larger the θ, the lower the number of predicted attributes, so does the F1-score. This may be due to the fact that although the precision

Product Long Text Description:
Off season women's classic Pima cotton cardigan, navy blue, cotton, slim fit, long sleeve, round neck, suitable for different seasons. This cardigan is made of ring spun Pima cotton with high softness, slim scissors, slim cutting, long sleeve, rib knitted cuffs. Rib knit hem. Button placket.

Product Image:

GT Summary: This cardigan is made of ring spun Pima cotton with high softness. The slim fit is suitable for different seasons. Thread knitted hem effectively modifies the waist line, which is quite thin.

V2P: This cardigan is made of ring spun leather horse cotton with high softness. It is exquisite and fashionable. The fabric is neat and stylish. The slim version shows more temperament. The cardigan design modifies the facial lines, and the button cardigan is easy to wear and take off.
(Rouge-1: 56.20%, Rouge-2: 42.02%, Rouge-L: 54.00%)

V2P-w/o-Image: Made of high-quality Pima cotton fabric, it feels soft and delicate, has good air permeability and brings a comfortable wearing experience. The classic round neck design naturally fits the neck and is beautiful.
(Rouge-1: 21.90%, Rouge-2: 4.44%, Rouge-L: 15.93%)

Figure 4: Comparison between the summaries generated by our model and its variant V2P-w/o-image for a clothing product. The English texts are translated from the Chinese texts. GT: Ground-truth.

Table 5: Performance (%) of our model with different θ. Count: the average number of predicted attributes.

Model	Count	F1-score	Rouge-1	Rouge-2	Rouge-L
θ=0.1	**31.38**	**10.35**	34.11	11.00	22.65
θ=0.2	11.09	7.54	**34.65**	**11.89**	**24.53**
θ=0.3	4.11	4.68	31.36	10.14	22.34
θ=0.4	1.71	2.76	29.58	9.12	21.18
θ=0.5	0.57	1.43	28.88	8.72	20.63

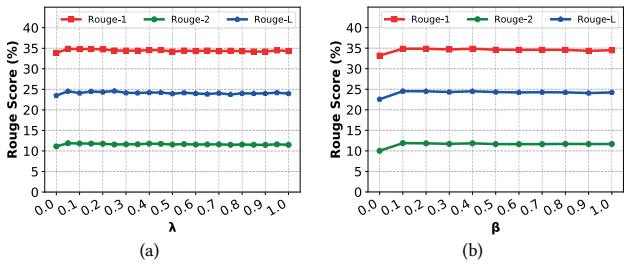

Figure 5: Sensitivity analysis on: (a) λ, and (b) β.

can be increased with the larger θ, the recall will be lower, and hence decreases the F1-score. In addition, we noticed that our model achieves the best Rouge scores when θ = 0.2. One possible explanation is that when θ is larger than 0.2, fewer attributes can be outputted as prompts, and thus hurt the Rouge scores of our model. On the contrary, when θ is too small, *i.e.*, 0.1, too many attributes would be yielded as prompts, which inevitably contain a few noisy ones, and hence harm the Rouge scores.

Meanwhile, we evaluated the sensitivity of our model in terms of the two key hyperparameters (*i.e.,* λ and β) in the objective function Eqn.(15) for optimizing the second component over the category Cases&Bags. Since our model performs best when λ = 0.05 and β = 0.1, we varied the values of λ and β from 0 to 1 at the step of 0.05 and 0.1, respectively. As can be seen from Figure 5, for both λ and β, our model performs worst when they are set to zero. This confirms the importance of the representation-level supervision and the data augmentation-based robustness regularization. In addition, we observed that our model performs relatively stably when λ and β are not zero. This implies that our model is not sensitive to the hyperparameters when they vary around the optimal values.

6 CONCLUSION AND FUTURE WORK

In this work, we present a vision-to-prompt based multi-modal product summary generation scheme, which seamlessly unifies the heterogeneous multi-modal data (*i.e.*, the long text description and image) of the product into the same embedding space of a GPLM. Extensive experiments on the large-scale Chinese dataset CEPSUM, involving around 1.4 million products of three categories (*i.e.,* Home Appliance, Clothing, and Cases&Bags), demonstrate the superiority of our model over existing cutting-edge methods. In particular, we notice that our model achieves the most significant improvement over the Clothing category. This is reasonable as the summaries of clothing products are more likely to contain the vision-related attributes, like color and pattern, as compared with those of the home appliances, cases, and bags. Meanwhile, the ablation study justifies the importance of incorporating the product's visual modality, representation-level supervision and data augmentation-based robustness regularization. Moreover, we also show the benefit of using Swin Transformer instead of VGG or ResNet in the vision-based prominent attribute prediction. In the future, we plan to adapt more advanced generative pre-trained language models to solve the multi-modal product summary generation task.

7 ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.: U1936203; Alibaba Group through Alibaba Innovative Research Program; and the Major Basic Research Project of Natural Science Foundation of Shandong Province, No.: ZR2021ZD15.

REFERENCES

[1] Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning Representations by Maximizing Mutual Information Across Views. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 15509–15519.

[2] Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards Knowledge-Based Personalized Product Description Generation in E-commerce. In *Proceedings of the The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 3040–3050.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the International Conference on Machine Learning*, Vol. 119. PMLR, 1597–1607.

[4] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. *CoRR* abs/2107.06278 (2021).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 4171–4186.

[6] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479.

[7] Yue Feng, Zhaochun Ren, Weijie Zhao, Mingming Sun, and Ping Li. 2021. Multi-Type Textual Reasoning for Product-Aware Answer Generation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1135–1145.

[8] Yu Gong, Xusheng Luo, Kenny Q. Zhu, Wenwu Ou, Zhao Li, and Lu Duan. 2019. Automatic Generation of Chinese Short Product Titles for Mobile Display. In *The Innovative Applications of Artificial Intelligence Conference*. AAAI Press, 9460–9465.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 770–778.

[10] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

[11] Chandra Khatri, Gyanit Singh, and Nish Parikh. 2018. Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks. *CoRR* abs/1807.08000 (2018).

[12] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 1–15.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 7871–7880.

[14] Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Aspect-Aware Multimodal Summarization for Chinese E-Commerce Products. In *The Innovative Applications of Artificial Intelligence Conference*. AAAI Press, 8188–8195.

[15] Qintong Li, Piji Li, Xinyi Li, Zhaochun Ren, Zhumin Chen, and Maarten de Rijke. 2021. Abstractive Opinion Tagging. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 337–345.

[16] Jinzhi Liao, Xiang Zhao, Xinyi Li, Lingling Zhang, and Jiuyang Tang. 2021. Learning Discriminative Neural Representations for Event Detection. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 644–653.

[17] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Meeting of the Association for Computational Linguistics*.

[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. *CoRR* abs/2103.14030 (2021).

[19] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. 3111–3119.

[20] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543.

[21] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alex Alemi, and George Tucker. 2019. On Variational Bounds of Mutual Information. In *Proceedings of the International Conference on Machine Learning*. PMLR, 5171–5180.

[22] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing*. ACL, 2401–2410.

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (2020), 140:1–140:67.

[24] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 1073–1083.

[25] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.).

[26] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *Proceedings of the International Conference on Machine Learning*, Vol. 97. PMLR, 5926–5936.

[27] Xue Song, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. 2021. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia* (2021).

[28] Hoang Van, Vikas Yadav, and Mihai Surdeanu. 2021. Cheap and Good? Simple and Effective Data Augmentation for Low Resource Machine Reading. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2116–2120.

[29] 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 *Proceedings of the Annual Conference on Neural Information Processing Systems*. 5998–6008.

[30] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. 2021. M2tr: Multi-modal multi-scale transformers for deepfake detection. *arXiv preprint arXiv:2104.09770* (2021).

[31] Jason W. Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. ACL, 6381–6387.

[32] Joan Xiao and Robert Munro. 2019. Text Summarization of Product Titles. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Vol. 2410. CEUR-WS.org.

[33] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. 2021. Self-Supervised Learning with Swin Transformers. *CoRR* abs/2105.04553 (2021).

[34] Guohai Xu, Yan Shao, Chenliang Li, Feng-Lin Li, Bin Bi, Ji Zhang, and Haiqing Chen. 2021. AliMe DA: A Data Augmentation Framework for Question Answering in Cold-start Scenarios. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2637–2638.

[35] Min Yang, Qiang Qu, Ying Shen, Qiao Liu, Wei Zhao, and Jia Zhu. 2018. Aspect and Sentiment Aware Abstractive Review Summarization. In *Proceedings of the International Conference on Computational Linguistics*. ACL, 1110–1120.

[36] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 3995–4007.

[37] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the International Conference on Machine Learning*, Vol. 119. PMLR, 11328–11339.

[38] Jianguo Zhang, Pengcheng Zou, Zhao Li, Yao Wan, Xiuming Pan, Yu Gong, and Philip S. Yu. 2019. Multi-Modal Generative Adversarial Network for Short Product Title Generation in Mobile E-Commerce. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 64–72.

[39] Xueying Zhang, Yunjiang Jiang, Yue Shang, Zhaomeng Cheng, Chi Zhang, Xiaochuan Fan, Yun Xiao, and Bo Long. 2021. DSGPT: Domain-Specific Generative Pre-Training of Transformers for Text Generation in E-commerce Title and Review Summarization. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2146–2150.

[40] Mengxue Zhao, Yang Yang, Miao Li, Jingang Wang, Wei Wu, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2022. Personalized Abstractive Opinion Tagging. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[41] Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging Lead Bias for Zero-shot Abstractive News Summarization. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1462–1471.