Aspect-Aware Multimodal Summarization for Chinese E-Commerce Products

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, Bowen Zhou
JD AI Research
{lihaoran24, yuanpeng29, xusong28, wuyouzheng1, xiaodong.he, bowen.zhou}@jd.com

Abstract

We present an abstractive summarization system that produces summary for Chinese e-commerce products. This task is more challenging than general text summarization. First, the appearance of a product typically plays a significant role in customers' decisions to buy the product or not, which requires that the summarization model effectively use the visual information of the product. Furthermore, different products have remarkable features in various aspects, such as "energy efficiency" and "large capacity" for *refrigerators*. Meanwhile, different customers may care about different aspects. Thus, the summarizer needs to capture the most attractive aspects of a product that resonate with potential purchasers. We propose an aspect-aware multimodal summarization model that can effectively incorporate the visual information and also determine the most salient aspects of a product. We construct a large-scale Chinese e-commerce product summarization dataset that contains approximately 1.4 million manually created product summaries that are paired with detailed product information, including an image, a title, and other textual descriptions for each product. The experimental results on this dataset demonstrate that our models significantly outperform the comparative methods in terms of both the ROUGE score and manual evaluations.

Introduction

Commercial product advertisements, as a critical component of marketing management in e-commerce platforms, aim to attract consumers' interests and arouse consumers' desires to purchase the products. However, most product advertisements are so miscellaneous and tedious that the consumers cannot be expected to be patient enough to carefully read through them. As shown in Figure 1, advertisements are composed of numerous pictures with textual product descriptions. In this case, the potential consumers must browse the web page to retrieve the detailed product characteristics that they care about, which would hurt their consumption experience. A product summary can effectively provide customers with valuable information about the product in a short time, which is of practical value for e-commerce scenarios.

Figure 1: The product summarization task.

Compared with the conventional text summarization task, it is more challenging to generate summaries for e-commerce products. On one hand, the appearance of a product establishes customers' first impressions of the product, which has a critical impact on purchasing decisions. Thus, the summarization system must be able to adequately represent the useful information from product images. On the other hand, different products have different selling points. For example, the advantage of *compact refrigerators* is saving space and that of *environmentally friendly refrigerators* is high energy efficiency. Therefore, the summary should focus on the most distinctive aspects of the product, which can maximally promote consumers' interests.

Recently, there has been a surge of work proposed to generate summaries for e-commerce products (Wang et al. 2017; Chen et al. 2019). Daultani, Nio, and Chung (2019)

propose generating extractive summaries for product descriptions based on a coverage maximization algorithm. Khatri, Singh, and Parikh (2018) introduce a document-context-based sequence-to-sequence (seq2seq) model to produce summaries for product descriptions. Although applicable, these methods do not use the products' visual information. Moreover, they do not attempt to improve the quality of the summaries with respect to product aspects. In this paper, we propose an aspect-aware multimodal summarization model for Chinese e-commerce products, which can integrate the visual and textual information of a product to generate an aspect-aware summary. Specifically, as shown in Figure 1, our model takes a product image, a product title, and other product descriptions as the input, and generates a textual summary.

The seq2seq model is the dominant framework for natural language generation tasks, and the pointer-generator networks facilitate the seq2seq model by directly copying the words from the source text to the target, which has been proved useful for the abstractive text summarization task. In this work, we extend the pointer-generator networks into a multimodal framework, and we explore various strategies to use product image.

We tackle the common challenges for summarization tasks, including importance, non-redundancy, and readability, with respect to product aspects. For importance, we adopt an aspect-based reward augmented maximum likelihood (RAML) training method (Norouzi et al. 2016) that encourages the summarizer to produce a summary that covers salient product aspects. For non-redundancy, we introduce an aspect coverage mechanism to keep track of what aspects have been mentioned, which can discourage aspect repetition. For readability, we apply constrained decoding to enhance the coherence of summaries by avoiding aspect rebounding.

We construct CEPSUM, a **C**hinese **E**-commerce **P**roduct **SUM**marization dataset that contains approximately 1.4 million product-summary pairs in three categories, including *Home Appliances*, *Clothing*, and *Cases & Bags*. Experimental results on this dataset demonstrate that our models significantly outperform the comparative methods in terms of both the ROUGE score and manual evaluations.

Our main contributions are as follows.

- We design a summarizer that can automatically generate an aspect-aware textual summary for a Chinese e-commerce product by integrating textual and visual product information.
- We propose a multimodal pointer-generator network and explore various approaches to use product images in the product summarization task.
- We adopt aspect RAML training, aspect coverage, and aspect coherence strategies, which aim to improve importance, non-redundancy, and readability, respectively.
- We construct a large-scale corpus of Chinese e-commerce product summarization, which includes three subsets that correspond to three categories of products. The experimental results on this dataset demonstrate the superiority of our approach against other methods.

Background

In this section, we describe our baseline pointer-generator networks (See, Liu, and Manning 2017) and product aspects. The pointer-generator network is a seq2seq model with a copying mechanism that can copy tokens in the source sequence into the proper positions in the target sequence. The product aspects are some characteristics of a product, such as "energy efficiency" and "capacity" for *refrigerators*. Identifying the dominant aspects of a product is beneficial to generating an attractive summary.

Pointer-Generator Networks

Given a source sequence $\mathbf{x}$, the seq2seq model maximizes the conditional probability of a target sequence $\mathbf{y}$: $P(\mathbf{y}|\mathbf{x})$. The encoder and the decoder are the two primary components in the seq2seq model (Cho et al. 2014). The encoder reads a variable-length input sequence $\mathbf{x}$ and then converts $\mathbf{x}$ into an encoder hidden sequence $\mathbf{h}$ as follows:

$$h_i = f_{enc}(x_i, h_{i-1}) \tag{1}$$

where h_i is an encoder hidden state, and f_{enc} is an encoder function.

The decoder generates the hidden sequence $\mathbf{s}$ as follows:

$$s_t = f_{dec}(s_{t-1}, y_{t-1}, c_t) \tag{2}$$

where $s_t \in \mathbb{R}^n$ is a hidden state at timestep t, and f_{dec} is a decoder function. c_t is a context vector that is generated by weighted sum of the encoder hidden states based on attention (Bahdanau, Cho, and Bengio 2015) distribution α^t as follows:

$$e_{t,i} = u_a^T \tanh(W_a h_i + V_a s_{t-1} + b_a) \tag{3}$$

$$\alpha_t = \mathrm{softmax}(e_t) \tag{4}$$

$$c_t = \sum\nolimits_i \alpha_{t,i} h_i \tag{5}$$

For a seq2seq model without a copying mechanism, the probability distribution P_{gen} over all words in the target vocabulary is calculated as follows:

$$P_{gen}(w) = \mathrm{softmax}(W_b s_t + V_b c_t + b_b) \tag{6}$$

The pointer-generator network predicts words based on the probability distributions of two modules, namely, the generator and the pointer (Vinyals, Fortunato, and Jaitly 2015). The generator produces vocabulary distribution P_{gen} using Equation 6. The pointer copies a word y_t from the source sequence via pointing:

$$P_{copy}(w) = \sum\nolimits_{i:x_i=w} \alpha_{t,i} \tag{7}$$

The final distribution is a weighted sum of the vocabulary distribution and the attention distribution:

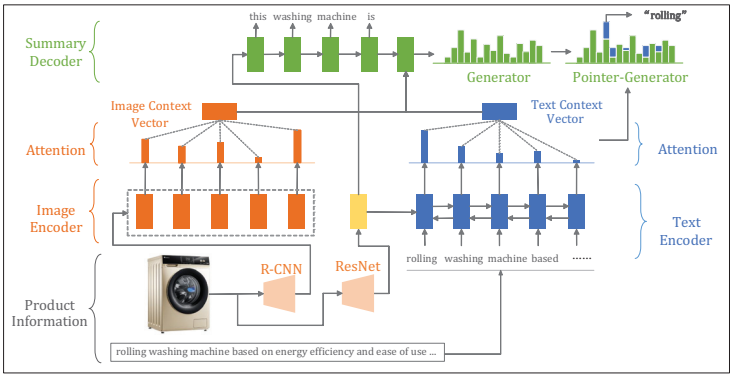
$$P(w) = \lambda_t P_{gen}(w) + (1 - \lambda_t) P_{copy}(w) \tag{8}$$

where $\lambda_t \in [0, 1]$ is the generation probability for timestep t:

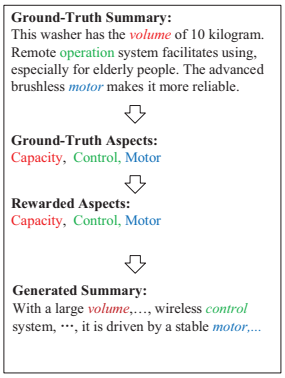
$$\lambda_t = \mathrm{sigmoid}(w_d^T c_t + u_d^T s_t + v_d^T y_{t-1} + b_d) \tag{9}$$

The loss function $\mathcal{L}$ is the average negative log likelihood of the ground-truth target word y_t for each timestep t:

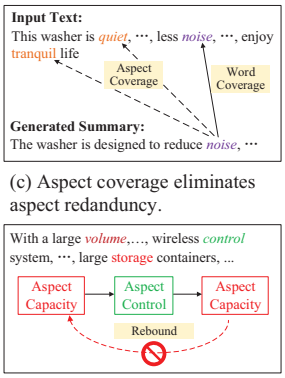
$$\mathcal{L} = -\frac{1}{T} \sum\nolimits_{t=0}^{T} \log P(y_t) \tag{10}$$



(a) Multimodal Pointer-Generator Network (MMPG). The global visual feature from ResNet is used to initialize the text encoder and summary decoder, and the local visual features from R-CNN are used to generate context representation.



(b) Aspect RAML encourages salient aspects to appear in generated summary.



(c) Aspect coverage eliminates aspect redanduncy.

(d) Aspect coherence improves readability by avoiding aspect rebounding.

Figure 2: The framework of our model.

Table 1: Product aspect analysis for *refrigerators*.

Aspect	#keywords	Examples
Modes	28	air-cooled, direct-cooled
Energy efficiency	21	energy, consumption
Motor	9	motor, variable-frequency
Control	22	control, touch-sensitive
Freshness	69	deodorise, fresh
Appearance	94	double-door, fashionable
Noise	14	quiet, noise
Cleanliness	16	clean, antiseptic
Material	34	glass, stainless steel
Function	25	child lock, quakeproof

Product Aspects

Before modeling product aspects, we should acquire a practical approach to recognize product aspects. Based on a preliminary observation of our dataset, we find that a complete product description or a product summary may contain information about several aspects of the product. Meanwhile, each subsentence that is divided by punctuations including commas, periods, semicolons, question marks, and exclamation marks always describes a specific aspect. For such a subsentence, we can identify its aspect using some keywords, such as "consumption" and "energy" for the aspect of "energy efficiency". We show the product aspects and aspect keywords for *refrigerators* in Table 1. The way to mine the aspect keywords is introduced in the Dataset section.

Our Proposed Model

Overview

We first define the e-commerce product summarization task. The input includes a product image and textual information by concatenating the title and the product descriptions. The output is a product summary.

As shown in Figure 2, our proposed aspect-aware summarization model is based on the pointer-generator networks. To incorporate the visual features into the pointer-generator networks, we apply three strategies, including initializing the encoder with the global visual feature, initializing the decoder with the global visual feature, and generating context representations with the local visual features.

We model the importance, non-redundancy, and readability of the summary with respect to the product aspects. (1) We adopt an aspect-based RAML training to generate a summary covering salient product aspects. (2) We employ an aspect coverage mechanism to eliminate aspect redundancy. (3) We improve the readability by avoiding aspect rebounding.

Multimodal Pointer-Generator Networks

We first introduce two strategies for the visual features based on the model's hidden state initialization. For a general pointer-generator network, the encoder's initial hidden state h_0 and h_{n+1} are zero vectors, and the decoder's initial hidden state s_0 is initialized using the last hidden state for the backward LSTM $\overrightarrow{h}_n$ and forward LSTM $\overleftarrow{h}_1$:

$$\overline{h} = [\overrightarrow{h}_n; \overleftarrow{h}_1] \tag{11}$$

$$s_0 = \tanh(W_d \overline{h} + b_w) \tag{12}$$

For our multimodal pointer-generator networks, given a product image, we extract the activations from the last pooling layer of ResNet-101 (He et al. 2016) which is pre-trained on ImageNet (Deng et al. 2009) as the global visual features q, which we use to initialize the encoder and the decoder:

$$\overrightarrow{h}_0 = \tanh(W_{e1} q + b_{e1}) \tag{13}$$

$$\overleftarrow{h}_{n+1} = \tanh(W_{e2} q + b_{e2}) \tag{14}$$

$$s_0 = \tanh(W_f \overline{h} + V_f q + b_f) \tag{15}$$

The third strategy is enhancing the context representations using the local visual features. We extract the object proposal features v_i as the local visual features. We use the Faster R-CNN (Ren et al. 2017) that is initialized using ResNet-101, and then we retrain it using the Visual

Genome Dataset (Krishna et al. 2017). $v_i \in \mathbb{R}^{2048}$ is obtained from the ROI pooling layer in the Region Proposal Network. Finally, we choose the top 16 object proposals after non-maximum suppression (Neubeck and Gool 2006).

Specifically, beyond the attention over the words of the input sentence, our model can pay attention to different regions of the image, and we also apply the hierarchical attention (Libovický and Helcl 2017) to make our model pay distinct attention to textual and visual information.

We compute the visual context vector c_t^{img} using the cross-modal attention strategy:

$$e_{t,i}^{img} = u_l^T \tanh(W_l v_i + V_l s_{t-1} + b_l) \tag{16}$$

$$\alpha_t^{img} = \text{softmax}(e_t^{img}) \tag{17}$$

$$c_t^{img} = \sum\nolimits_i \alpha_{t,i}^{img} v_i \tag{18}$$

We obtain the multimodal context vector c_t^{mm} by using the weighted sum of the textual context vector c_t in Equation 5 and the visual context vector c_t^{img} Equation 18. Following Li et al. (2018a), we adopt an image attention filter I_a to eliminate visual noise:

$$\beta_t^{txt} = \sigma(W_g s_t + V_g c_t + b_g) \tag{19}$$

$$\beta_t^{img} = \sigma(W_h s_t + V_h c_t^{img} + b_h) \tag{20}$$

$$I_a = \sigma(v_0^T s_0 + v_q^T q + v_s^T s_{t-1}) \tag{21}$$

$$\beta_t^{img} = I_a \cdot \beta_t^{img} \tag{22}$$

$$c_t^{mm} = \beta_t^{txt} W_m c_t + \beta_t^{img} V_m c_t^{img} \tag{23}$$

Finally, c_t in Equation 6 is replaced by c_t^{mm}.

Aspect RAML Training

RAML is a computationally efficient training paradigm to optimize task-specific reward. In this work, we adopt RAML to incorporate aspect-based reward into our summarization model. Our objective is to encourage our model to generate summary including important product aspects as shown in Figure 2(b).

Given a dataset $\mathcal{D} \equiv \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$, the RAML loss function is defined as follows:

$$\mathcal{L}_{\text{RAML}} = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \left\{ -\sum_{\bar{\mathbf{y}}\in\mathcal{Y}} q\left(\bar{\mathbf{y}}|\mathbf{x},\mathbf{y};\tau\right) \log p(\mathbf{y}|\mathbf{x}) \right\} \tag{24}$$

$$q\left(\bar{\mathbf{y}}|\mathbf{x},\mathbf{y};\tau\right) = \frac{1}{Z(\mathbf{x},\mathbf{y},\tau)} \exp\{r(\mathbf{x},\bar{\mathbf{y}},\mathbf{y})/\tau\} \tag{25}$$

$$Z(\mathbf{x},\mathbf{y},\tau) = \sum_{\bar{\mathbf{y}}\in\mathcal{Y}} \exp\{r(\mathbf{x},\bar{\mathbf{y}},\mathbf{y})/\tau\} \tag{26}$$

where $\mathcal{Y}$, $r(\mathbf{x},\bar{\mathbf{y}},\mathbf{y})$, τ are possible model outputs, reward function, and regularization parameter, respectively. The gradient of $\mathcal{L}_{\text{RAML}}$ is defined in terms of an expectation over samples from $q\left(\bar{\mathbf{y}}|\mathbf{x},\mathbf{y};\tau\right)$:

$$\nabla\mathcal{L}_{\text{RAML}} = E_{q(\bar{\mathbf{y}}|\mathbf{x},\mathbf{y};\tau)}\left[-\nabla \log p(\bar{\mathbf{y}}|\mathbf{x})\right] \tag{27}$$

Given a training sample $(\mathbf{x},\mathbf{y})$, RAML training first samples an output $\bar{\mathbf{y}}$ proportionally to the reward. Then, RAML optimizes log-likelihood on the sample. In this work, we sample auxiliary outputs from the exponentiated payoff distribution, $q\left(\bar{\mathbf{y}}|\mathbf{x},\mathbf{y};\tau\right)$ by stratified sampling. Particularly, given a sentence $\mathbf{y}$ of length ℓ, we first count the number of sentences within an edit distance d, where $d \in \{0,\ldots,2\ell\}$. Then, we reweight the counts by $\exp\{-d/\tau\}$ and normalize. Finally, we apply importance sampling by the weight $\exp\{(s(\mathbf{x},\bar{\mathbf{y}},\mathbf{y})+d)/\tau\}$, where the proposal distribution is Hamming distance sampling.

We define the reward $s(\mathbf{x},\bar{\mathbf{y}},\mathbf{y})$ as aspect precision:

$$s(\mathbf{x},\bar{\mathbf{y}},\mathbf{y}) = \frac{|\{\text{aspect}_i|\text{aspect}_i \in \bar{\mathbf{y}} \cap \mathbf{y}\}|}{|\{\text{aspect}_i|\text{aspect}_i \in \bar{\mathbf{y}}\}|} \tag{28}$$

Note that $s(\mathbf{x},\bar{\mathbf{y}},\mathbf{y}) = 0$ if there is not aspect in $\bar{\mathbf{y}}$.

Aspect Coverage

We employ an aspect coverage mechanism to eliminate aspect redundancy, as shown in Figure 2(c), which is expected to increase the recall for valuable aspects.

Word-based coverage mechanism (See, Liu, and Manning 2017) aims to avoid generating repetitive words by tracking the attention history and penalizing the repetitive attention to the same position of the source text as follows:

$$c_{t,i} = \sum_{\tau=1}^{t-1} \alpha_{\tau,i} \tag{29}$$

$$e_{t,i} = u_c^T \tanh(W_c s_{t-1} + V_c h_i + w_c c_{t,i}) \tag{30}$$

$$\mathcal{L}^w = \sum_{t=1}^{T}\sum_{i=1}^{N} \min(\alpha_{t,i}, c_{t,i}) \tag{31}$$

$$\mathcal{L} = \mathcal{L} + \mathcal{L}^w \tag{32}$$

For our aspect coverage model, we record the attention history for all the product aspects, and then the decoder is discouraged to pay the current attention to the aspects which have received enough attention before. Specifically, we maintain an aspect attention vector α_{t,a_j}, which is the sum of attention distributions over all the words belonging to aspect a_j:

$$\alpha_{t,a_j} = \sum_{x_k \in a_j} \alpha_{t,k} \tag{33}$$

Then we calculate aspect coverage vector c_{t,a_j} as sum of aspect attention distributions over all previous decoder steps:

$$c_{t,a_j} = \sum_{\tau=1}^{t-1} \alpha_{\tau,a_j} \tag{34}$$

Intuitively, c_{t,a_j} denotes the degree of aspect a_j covered by the target before time t, and c_{0,a_j} is a zero vector. Then for $x_i \in a_j$, $e_{t,i}$ in Equation 3 is updated as follows:

$$e_{t,i} = u_e^T \tanh(W_e s_{t-1} + V_e h_i + w_e c_{t,a_j}) \tag{35}$$

The aspect coverage loss is defined as follows:

$$\mathcal{L}^a = \sum_{t=1}^{T}\sum_{j=1}^{J} \min(\alpha_{t,a_j}, c_{t,a_j}) \tag{36}$$

$$\mathcal{L} = \mathcal{L} + \mathcal{L}^a \tag{37}$$

Aspect Coherence

We propose an aspect coherence strategy to enhance readability. As shown in Figure 2(d), the generated summaries first describe aspect "Capacity", and then describe aspect "Control", and later describe aspect "Capacity" again. We refer to this phenomenon as aspect rebounding, which influences the aspect-level coherence. To avoid aspect rebounding, we apply a constrained decoding during testing. Specifically, if aspect a_2 appears after aspect a_1, then for the word y_{a_1} corresponding to aspect a_1, we will set $P(y_{a_1}) = 0$, which means a_1 is not allowed to be generated again.

Improving Consistency

Consistency is an essential requirement for summarization (Cao et al. 2018; Li et al. 2018b). A product summary with false information may lead customers to buy the product they do not really need, which brings in bad user experience, or even destroys customers' trust in the e-commerce platform. It is mainly caused by wrong attribute words, such as "100 litre" for *Home Appliances* and "silk" for *Clothing*, which do not exist in the product information while appear in the generated summary. To avoid generating summaries inconsistent with the product descriptions, we first extract the attribute words from product specifications (key-value pairs like "material: glass", we extract value as attribute word), and then for each attribute word y_{att}, we set $P_{gen}(y_{att}) = 0$ in Equation 6 during testing. In other words, attribute words can present in the summary only through copying from the source.

Related work

Abstractive Text Summarization

Rush, Chopra, and Weston (2015) first apply the seq2seq model to abstractive sentence summarization. Chopra, Auli, and Rush (2016) and Nallapati et al. (2016) extend the seq2seq model. Gu et al. (2016) and Zeng et al. (2016) incorporate a copying mechanism into seq2seq learning and Gulcehre et al. (2016) propose a switch gate to control whether to copy from the source or generate a word by the decoder. See, Liu, and Manning (2017) propose a hybrid pointer-generator architecture with coverage to reduce repetition.

E-Commerce Product Summarization

Chen et al. (2019) generate personalized product descriptions. They design three product aspects, and the generated product description is restricted to focus on one aspect. Yang et al. (2018) produce summaries for the product reviews. They apply a mutual attention mechanism to integrate the sentiment and aspect information into the summarizer. Our work focuses on condensing a detailed product description into a summary, which aims to save time and improve the purchasing experiences for customers.

Daultani, Nio, and Chung (2019) propose a system to generate extractive summaries for product descriptions based on coverage maximization and various concepts. Khatri, Singh, and Parikh (2018) use a document-context-based seq2seq model to produce abstractive and extractive summaries for product descriptions, which outperforms standard seq2seq models. Intuitively, the customer will look at the product image before reading detailed product descriptions and decide to buy the product whose distinctive aspects they care about, and multimodal information show efficiency for summarization (Li et al. 2017; 2018a; Zhu et al. 2018; Li et al. 2019). Thus, we propose an aspect-aware multimodal abstractive summarizer, which can integrate visual and textual information of a product to generate an aspect-aware summary.

Table 2: Corpus statistics.

Category	Home Appliances	Clothing	Cases&Bags
# Train	437,646	790,297	97,510
# Valid	10,000	10,000	5,000
# Test	10,000	10,000	5,000

Dataset

Dataset Construction

We collect the dataset from a large-scale e-commerce platform in China, including about 1.4 million instances from the *Home Appliances*, *Clothing*, and *Cases & Bags* categories. There are 119 kinds of product in *Home Appliances*, such as *air conditioner*, *water heater*, and *electric pot*; 86 kinds of product in *Clothing*, such as *jacket*, *dress*, and *T-shirt*; 33 kinds of product in *Cases & Bags*, such as *suitcase*, *handbag*, and *wallet*. Each instance in our dataset is a (product information, product summary) pair, and the product information contains an image, a title, and other product descriptions. The product summaries are generated by thousands of qualified writers from the "Discover" channel of the e-commerce platform. The professional auditing groups will strictly review the summaries, and only the high-quality summaries are kept. The average numbers of the characters in the source and the target are 303 and 79, respectively. Table 2 shows more details about our dataset.

Aspect Keywords Mining

The products in the *Clothing* and *Cases & Bags* categories share the same sets of aspect annotations, respectively. Due to the significant differences among products in the *Home Appliances* category, we divide these products into 16 subcategories, such as the *refrigerators* subcategory that includes *refrigerators*, *freezers*, and *electric wine cabinets*. We mine the aspect keywords using the Aspect Segmentation algorithm (Wang, Yue, and Zhai 2010), which is a boot-strapping algorithm automatically extending aspect keywords. For each category or subcategory, we first manually define aspects and their keywords by annotating 200 ground-truth summaries, and each remaining sentence is labeled with the aspect whose corresponding keyword appears in the sentence. Then, the Aspect Segmentation algorithm will calculate the dependencies between aspects and words using the Chi-Square statistic and add the words with high dependencies into the corresponding aspect keyword list. As a result, approximately 80% of the subsentence in our corpus can be covered by our aspect annotation, and almost all

of the uncovered subsentences are not related to any aspect, such as emotional expression. There are 10.6 product aspects on average for each category or subcategory. For each product aspect, there are 30.9 aspect keywords on average.

Experiment

Comparative Methods

We compare several text-based extractive and abstractive summarization methods.

Lead extracts the first 80 characters (according to the length of the ground-truth summary) as the summary.

LexRank (Erkan and Radev 2011) is an unsupervised graph-based ranking algorithm for extractive summarization.

Seq2seq is a standard text-based seq2seq model with an attention mechanism.

Pointer-Generator (PG) is a text-based seq2seq network model with an attention and copying mechanism.

MASS (Song et al. 2019) is the-state-of-the-art method for text summarization based on masked language modeling.

Experimental Settings

We set the sizes of the word embedding and the LSTM hidden state to 300 and 512, respectively. We set the initial learning rate for Adam to 5×10^{-4}. The mini-batch size is set to 16. During training, we test ROUGE-2 (Lin 2004) F1 score and perplexity using the development set for every 5,000 batches, and we halve the learning rate if model's ROUGE-2 score drops for 7 consecutive tests. We first train our models without coverage until they converge using an early stopping strategy, and then we add the coverage mechanism to further train the models. During testing, we use the beam search with a beam size of 10 to generate the summaries, and character-based trigram repetition avoidance (Paulus, Xiong, and Socher 2018) is applied.

Experimental Results

We use the ROUGE-1.5.5 toolkit (Lin and Hovy 2003) to evaluate the summaries. We report character-based ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) F1 scores.

Table 3 shows the performances of the multimodal pointer-generator (MMPG) model with different visual feature strategies. **EncInit (E)** denotes initializing the encoder with global visual features. **DecInit (D)** denotes MMPG model initializing the decoder with global visual features. **MMAtt (M)** denotes multimodal attention mechanism using local visual features.

We can conclude that the **MMPG** models outperform the **Seq2seq** and **PG** models, which corroborates the necessity of product images for this task. Among the different product categories, product images contribute the most to *Clothing*, followed by *Cases & Bags*, since customers may put more emphasis on the appearance for these categories. In addition, we find that **DecInit** is the most effective single strategy among the MMPG models and combining **DecInit** and **MMAtt** is slightly better than **DecInit**, while **MMAtt** significantly increases the parameters for the model. Thus, we adopt the **DecInit** strategy for our **MMPG** model in the following experiments.

Table 3: Experimental results for multimodal models (%).

	Home App.		Clothing		Cases&Bags	
	R-1	R-2	R-1	R-2	R-1	R-2
Seq2seq	21.57	7.18	23.05	6.84	23.18	6.94
PG	31.31	10.93	29.11	9.24	31.11	10.27
MMPG+EncInit (E)	32.67	11.54	30.14	9.85	32.05	11.36
MMPG+DecInit (D)	32.88	11.88	**30.73**	10.29	**32.69**	**11.78**
MMPG+MMAtt (M)	32.76	11.67	30.67	10.13	32.59	11.58
MMPG+D+M	32.87	11.88	30.72	**10.31**	32.66	11.74
MMPG+E+D+M	**32.89**	**11.89**	30.71	10.30	32.68	11.77

Table 4: Diversity manual evaluation.

	Home App.	Clothing	Cases&Bags
PG	2.56	2.45	2.47
MMPG+EncInit	2.91	2.96	2.83
MMPG+DecInit	3.12	3.22	3.23
MMPG+MMAtt	2.94	3.11	3.18
Ground-Truth	3.29	3.25	3.21

The experimental results for other comparative methods are shown in Table 5. The **PG** model performs better than other text-based models, which suggests that the copy mechanism is essential. It is reasonable to copy the original text from the product description for human writers without product experience. Our proposed **MMPG** model with **DecInit** achieves significantly better results than all the text-based models, and the three aspect-based strategies lead to further improvements, which proves that aspect information plays a significant role in the product summarization task. When combining **Consistency** strategy, our model achieves +2.85% R-1, +1.89% R-2, and +1.17% R-L improvements on average over the **PG** model.

Further Analysis

Effectiveness of the Image

Product image can provide extra information for product summarization task. It is unclear whether the images can help our model generate more diverse summaries. Thus, we randomly sample 100 products for each category and employ 10 graduate students to evaluate diversity with a score from 1 (worst) to 5 (best). The results in Table 4 show that multimodal models outperform text-based model, especially for *Clothing* and *Cases& Bags*, for which people are inclined to focus on appearance of the product.

In addition, for these products, through manual verification, we find that approximately 5% of the product summaries that are generated by the **PG** model are inconsistent with the corresponding categories, and this drops to 2% for **MMPG+DecInit** model. This outcome demonstrates that the product image is a straightforward indicator for the product category.

Effectiveness of the Aspect

The ROUGE evaluation is weakly correlated with aspect information, thus conduct evaluations for the aspect-based im-

Table 5: Main results (%). Our **MMPG** models perform significantly better than other competitive methods by the 95% confidence interval in the ROUGE script.

	Home Appliances			Clothing			Cases&Bags		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-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
MMPG+Aspect	33.97	12.43	22.21	31.81	10.87	21.32	33.67	12.44	22.31
MMPG+Aspect+Consistency	**34.36**	**12.52**	**22.35**	**31.93**	**11.09**	**21.54**	**33.78**	**12.51**	**22.43**

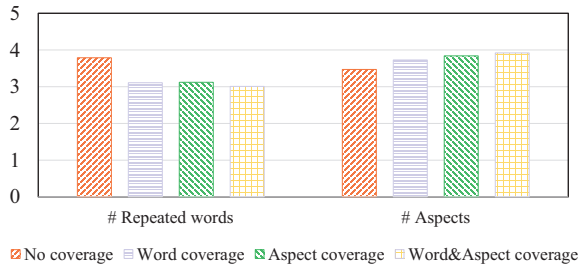


Figure 3: Coverage alleviates word and aspect redundancy.

portance, non-redundancy, and readability on the 100 products for each category. The results are shown in Table 6 and Figure 3. **AspRAML** is aspect-based reward augmented maximum likelihood training framework. **AspCov** and **AspCoh** denote aspect coverage and aspect coherence, respectively.

Importance Analysis Compared with the general text summarization task, retaining important aspects is an essential requirement for product summarization. We also manually review whether the important aspects that are mentioned in the product description appear in their respective summary. The results that are presented in Table 6 ("Imp." column) show that aspect-based RAML is conducive to identify the salient aspects.

Non-Redundancy Analysis We measure the non-redundancy with respect to the words and aspects. The results in Figure 3 show that the number of duplicated words is reduced and the number of aspect is increased.

Readability Analysis We assess the readability using a score from 1 (worst) to 5 (best). From Table 6 ("Read." column), we can find that when aspect rebounding is eliminated by our aspect-based constrained decoding, the readability is significantly improved.

Consistency Analysis

We manually evaluate whether the generated summaries are consistent with the product descriptions. From Table 7, we can find that over 90% of the summaries are faithful to the source, illustrating that our consistency strategy is effective.

Table 6: Manual evaluation for aspect (%). "Imp." and "Read." denote "Importance" and "Readability", respectively.

	Home Appl.		Clothing		Cases&Bags	
	Imp.	Read.	Imp.	Read.	Imp.	Read.
MMPG	78.00	3.05	76.00	3.26	73.00	3.06
+AspRAML	91.00	3.06	92.00	3.31	88.00	3.12
+AspCov	92.00	3.04	94.00	3.22	91.00	3.10
+AspCoh	93.00	3.26	94.00	3.64	92.00	3.34

Table 7: Consistency Analysis (%). "Asp." denotes "aspect".

	Home Appl.	Clothing	Cases&Bags
MMPG+Asp.	71	66	62
+Consistency	92	91	93

Conclusions

We address a product summarization task, namely, how to generate a summary for a product using its information, including a product image, a product title, and other detailed descriptions. Our proposed model can effectively leverage the visual information in product images and capture the most salient aspects of products. We implement three strategies to integrate the visual information into our model and prove that product images are necessary for our task. We adopt an aspect RAML to maintain the salient product aspects, and we propose an aspect coverage mechanism to improve the non-redundancy. Furthermore, our model produces more readable summaries by promoting aspect coherence. In addition, we introduce an effective method to ensure consistency. We provide a large-scale Chinese e-commerce product summarization corpus, and the experimental results on this dataset show our proposed aspect-aware multimodal pointer-generator model performs significantly better than the baselines. Our dataset and code are available[1].

Acknowledgments

This work is partially supported by Beijing Academy of Artificial Intelligence (BAAI). We would like to thank Yue

[1]https://github.com/hrlinlp/cepsum

Wang and Tiangang Zhu for helpful discussions.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Cao, Z.; Wei, F.; Li, W.; and Li, S. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of AAAI*, 4784–4791.

Chen, Q.; Lin, J.; Zhang, Y.; Yang, H.; Zhou, J.; and Tang, J. 2019. Towards knowledge-based personalized product description generation in e-commerce. In *KDD*.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, 1724–1734.

Chopra, S.; Auli, M.; and Rush, A. M. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of NAACL*, 93–98.

Daultani, V.; Nio, L.; and Chung, Y. 2019. Unsupervised extractive summarization for product description using coverage maximization with attribute concept. In *ICSC*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Erkan, G., and Radev, D. R. 2011. Lexrank: Graph-based lexical centrality as salience in text summarization. *CoRR* abs/1109.2128.

Gu, J.; Lu, Z.; Li, H.; and Li, V. O. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of ACL*, 1631–1640.

Gulcehre, C.; Ahn, S.; Nallapati, R.; Zhou, B.; and Bengio, Y. 2016. Pointing the unknown words. In *Proceedings of ACL*, 140–149.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Khatri, C.; Singh, G.; and Parikh, N. 2018. Abstractive and extractive text summarization using document context vector and recurrent neural networks. *CoRR* abs/1807.08000.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* 32–73.

Li, H.; Zhu, J.; Ma, C.; Zhang, J.; and Zong, C. 2017. Multimodal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of EMNLP*.

Li, H.; Zhu, J.; Liu, T.; Zhang, J.; and Zong, C. 2018a. Multimodal sentence summarization with modality attention and image filtering. In *Proceedings of IJCAI*, 4152–4158.

Li, H.; Zhu, J.; Zhang, J.; and Zong, C. 2018b. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of COLING*, 1430–1441.

Li, H.; Zhu, J.; Ma, C.; Zhang, J.; and Zong, C. 2019. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *TKDE*.

Libovický, J., and Helcl, J. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *ACL*.

Lin, C.-Y., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL*, 150–157.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Nallapati, R.; Zhou, B.; dos Santos, C.; Gulcehre, C.; and Xiang, B. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.

Neubeck, A., and Gool, L. V. 2006. Efficient non-maximum suppression. In *ICPR*, 850–855.

Norouzi, M.; Bengio, S.; Chen, Z.; Jaitly, N.; Schuster, M.; Wu, Y.; and Schuurmans, D. 2016. Reward augmented maximum likelihood for neural structured prediction. In *NeurIPS*.

Paulus, R.; Xiong, C.; and Socher, R. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*.

Rush, A. M.; Chopra, S.; and Weston, J. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of EMNLP*, 379–389.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of ACL*, 1073–1083.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of ICML*, 5926–5936.

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. In *NeurIPS*, 2692–2700.

Wang, J.; Hou, Y.; Liu, J.; Cao, Y.; and Lin, C. 2017. A statistical framework for product description generation. In *Proceedings of IJCNLP*, 187–192.

Wang, H.; Yue, L.; and Zhai, C. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of KDD*.

Yang, M.; Qu, Q.; Shen, Y.; Liu, Q.; Zhao, W.; and Zhu, J. 2018. Aspect and sentiment aware abstractive review summarization. In *Proceedings of COLING*, 1110–1120.

Zeng, W.; Luo, W.; Fidler, S.; and Urtasun, R. 2016. Efficient summarization with read-again and copy mechanism. *arXiv:1611.03382*.

Zhu, J.; Li, H.; Liu, T.; Zhou, Y.; Zhang, J.; and Zong, C. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of EMNLP*, 4154–4164.