# Product Summarization Extraction Model with Image Information

**Jingru Lin[1], Xinqian Hu[2], Yutao Chen[2], Xiangfei Dai[2], Wen Jiang[2]**
**[1]Class of School of Informatics**
**[2]Class of Artificial Intelligence Research Institute**
**23020241154420, 33320241150452, 33320241150450, 33320241150451, 23020241154402**

## Abstract

With the rapid development of the information age, we are faced with a huge amount of text and image data, how to quickly and accurately extract the key content from the complicated information has become an urgent problem. Product summarization refers to the concise and condensed description of a product on an e-commerce platform or other sales occasions, so that consumers can quickly understand the information and characteristics of the item. However, traditional product summarization mainly relies on text information and ignores the importance of image information. Therefore, combining image information to improve the quality and efficiency of product summarization has become a challenging and practical research topic. The aim of this paper is to design and train a product summarization model that incorporates image information.In the product summarization model, text and images are encoded using BART and Resnet50 models respectively, and these two encodings are fused using splicing, so that the model can recognize the features of text and images at the same time. Finally, the multimodal feature representations of all input sentences are fed into the extractive text summarizer to determine whether the sentence is a summary sentence or not to extract the final summary.

## Introduction

### Research Background and Significance

With the rapid development of e-commerce and the Internet, online shopping has become an indispensable part of modern life. People are increasingly inclined to purchase clothing products on online platforms because they can shop anytime, anywhere, and have a wide range of choices. However, the overwhelming amount of online information also brings difficulties in selecting clothing.As shown in Figure 1, there is a lot of clothing information provided on shopping websites, and various information are scattered in different locations. Consumers often choose products that are not suitable for themselves due to not reading carefully. Therefore, how to effectively summarize a large amount of clothing information has become an urgent problem to be solved today.

Traditional text summarization only focuses on textual information and ignores the importance of image information. With the development of image processing technology and artificial intelligence technology, product summarization systems can also achieve more accurate and personalized services. Product summary refers to a concise and condensed description of a product on e-commerce platforms or other sales occasions, enabling consumers to quickly understand the important information and features of the product. Through multimodal fusion technology, the system can extract text and image features, automatically form summaries, and summarize clothing styles, colors, fabrics, and other attributes to provide customers with more accurate information.

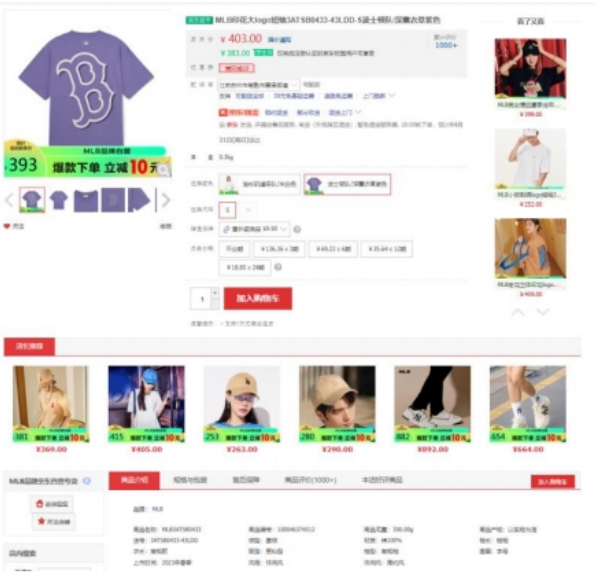

Figure 1: Clothing product introduction

### Main Work

This article mainly uses the CEPSUM dataset to preprocess and extract features from text and images. The BART model captures text semantic features and uses the deep residual network ResNet50 model to capture image visual features. Then, design a multimodal fusion module to fuse the two features to obtain more comprehensive information, generate concise and accurate text summaries. The main research work of this article is as follows:

- Build a multimodal fusion product summary model using BART model and ResNet50 model as the basic structure. The model combines text and image features to determine whether the input sentence is a summary sentence, and ultimately extracts the product summary. The model is tested based on the CEPSUM2.0 dataset and evaluated for system performance using ROUGE.

## Related Work

### Text Summarization

Text summarization is a type of automatic text generation that analyzes a given document or set of documents to extract key information,ultimately producing a concise summary.The sentences in the summary can either be directly taken from the original text or newly formulated.

With the development of models such as RNN(Kawakami 2008), LSTM(Hochreiter and Schmidhuber 1997) and Transformer(Vaswani et al. 2017),text summarization models have also been further improved.At present, Transformer-based models such as BERT(Devlin et al. 2019), BART(Lewis et al. 2020), and GPT(Elounda et al. 2023) have achieved remarkable achievements in the field of text processing.Jacob Devlin and others have discussed the definition of pre-trained models and the reasons for using them,introducing BERT, a pre-trained model that understands semantic information.Pre-training involves two tasks that allow the model to learn deep textual information, which is then fine-tuned with a good initial state for various downstream tasks.The BERT model uses a bidirectional Transformer-based pre-trained language model with strong contextual understanding capabilities(Sutskever, Vinyals, and Le 2014). However, the BERT model is complex, has a large number of parameters, and consumes more time and resources for training and inference. Mike Lewis and others proposed the BART model, a denoising autoencoder for pre-training sequence-to-sequence models. BART uses self-attention mechanisms, enabling the model to better understand context and generate more accurate and consistent outputs. Currently, compared to traditional extractive summarization, generative summarization shows a positive development trend. It can generate semantic representations that include more contextual structural information and are not limited to the words in the original text.

### Image Processing

Traditional text summarization focuses solely on textual information, neglecting the importance of image data. The acquisition of image information primarily relies on image encoding, which involves extracting features from the images to merge data from textual and visual modalities.

Since the inception of the ImageNet competition,it has greatly promoted the development of deep learning in the field of computer vision,promoted the establishment of large-scale data sets,the innovation of deep neural net-work architecture, and the application of transfer learning and other technologies, significantly improving image recognition and classification accuracy. AlexNet(Krizhevsky, Sutskever, and Hinton 2012) achieved a breakthrough in the ImageNet competition,marking the rise of deep learning in the field of computer vision.VGG(Simonyan and Zisserman 2014) further improved the accuracy of image recognition through a deeper network structure. When integrating features at different levels, deep networks face some challenges as the depth of the network increases.Although normalized initialization and intermediate normalization solve the gradient vanishing and exploding problems, allowing deep networks to converge, as the number of network layers increases,the training accuracy degrades.To solve this problem, He Kaiming et al. proposed a network called the"Deep Residual Learning Frame-work" ResNet(He et al. 2016) solved the degradation problem in deep network training by introducing residual connections,promoting the development of deeper networks.

### Multi-Modal Fusion

Multimodal fusion refers to the technical method of combining data from different modalities, including but not limited to images, audio, and text, to enhance information processing and understanding capabilities. Multimodal data typically possess different physical properties and informational characteristics. The main methods of multimodal fusion include feature-level(Baltrusaitis, Ahuja, and Morency 2018), decision-level, and attention mechanism fusion(Huo et al. 2021), among others. Currently, the more mainstream fusion methods are based on attention mechanisms and tensor-based modality fusion. Attention mechanism-based fusion possesses adaptability and flexibility, allowing the model to automatically adjust the weights of different modalities according to task requirements. Tensor-based modality fusion methods effectively utilize the multi-dimensional structure of data to capture complex relationships, although their computational complexity is relatively high.

## Proposed Solution

This paper mainly uses the CEPSUM dataset to preprocess and extract features from text and images. It leverages the BART model to capture textual semantic features and the deep residual network ResNet50 model to capture visual features from images. Then, a multimodal fusion module is designed to integrate these two types of features to obtain more comprehensive information, thereby generating concise and accurate text summaries.The specific structure of our model is shown in Figure 2,the Model architecture diagram.

### BART model

The BART model(Liu and Lapata 2019) consists of two parts: the Encoder and the Decoder. It utilizes the original Transformer Encoder-Decoder architecture. The specific structure of the Transformer is shown in Figure 3, the Transformer architecture diagram.

The Encoder part of the BART model is a multi-layer Transformer Encoder, primarily used for processing noisy text and encoding it into a hidden representation. The Decoder part is a multi-layer Transformer Decoder, responsible

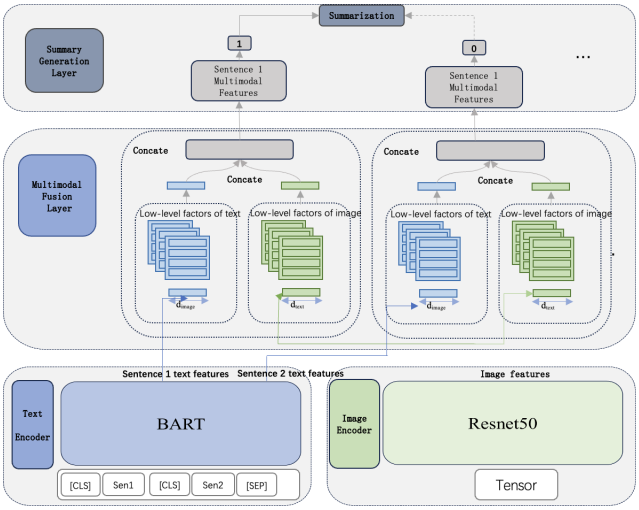


Figure 2: Model Architecture Diagram

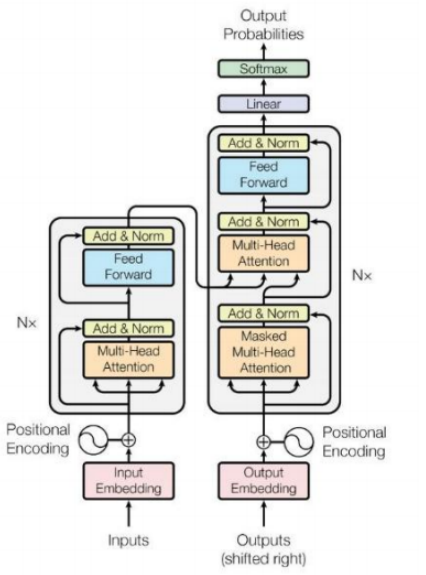


Figure 3: Transformer Architecture Diagram.

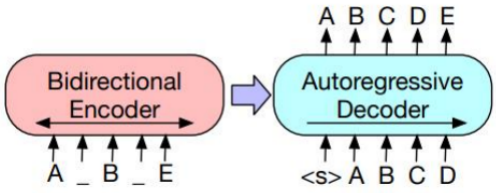


Figure 4: The BART model architecture.

for decoding the hidden representation produced by the Encoder into output text. This structure (as shown in Figure 4) enables the BART model to have both BERT's bidirectional language understanding capabilities and GPT's autoregressive generation capabilities.

The input to the Encoder does not need to be aligned with the output of the Decoder, thus allowing for arbitrary noise transformations. These noise transformations can help the model learn more robust representations because it is forced to understand and remember more contextual information in order to reconstruct the original document during the decoding phase. This process can also be seen as a variant of denoising autoencoders, where both the encoder and decoder (as shown in Figure 4 ) are based on Transformer models.

The pre-training of BART is divided into five tasks, with the main content of each task as follows:

1. Token Masking: Similar to the operation of the BERT model, BART randomly selects a portion of the input text's words and replaces them with the mask token [MASK]. The model needs to learn to infer these masked words from the context.
2. Token Deletion: BART also randomly omits some words, which means the model must not only predict the missing words but also determine their positions within the text.
3. Sentence Permutation: BART divides the input text into multiple sentences based on periods and randomly shuffles the order of these sentences. The model needs to understand the semantic information of the entire text to restore the correct order of the document.
4. Document Rotation: BART randomly selects a word as a starting point and rotates the text to begin at that word. The model needs to find the original beginning of the text to restore the order of the entire document.
5. Text Infilling: BART randomly selects multiple segments of the text and replaces these segments with mask tokens. The length of the segments is sampled according to a Poisson distribution, which requires the model to infer the replaced text content based on the context.

## Resnet model

Deep networks face several challenges when integrating features at different levels. Although normalized initialization and intermediate normalization have addressed the vanishing and exploding gradient problems, enabling deep networks to converge, further deepening of the network layers leads to a phenomenon of training accuracy degradation. This phenomenon is not due to overfitting but rather an increase in training error caused by the addition of layers. To address this issue, He et al. proposed a network called the "Deep Residual Learning Framework", which introduces residual connections allowing the network to learn multilayer features and effectively deepen the network's depth.

The fundamental building block of ResNet is the residual block. Each residual block contains multiple convolutional layers, as well as a skip connection that adds the input data directly to the output of the residual block, creating a residual learning strategy. Within the residual block, each stacked layer no longer directly fits H(x) (the desired output of the entire residual block), but fits F(x) (the residual between H(x) and the input x). Thus, the original mapping H(x) is

reconstructed as F(x) + x.F(x) + x can be implemented by a feedforward neural network, which is the line in Figure 5 that jumps from the input end to the output end, spanning one or more layers to perform the identity mapping. It is evident that this does not add extra parameters and computational load to the model.The core idea of residual learning is that, through skip connections, the network can preserve the original information in the input data and process and refine this information in subsequent layers through convolutional layers.

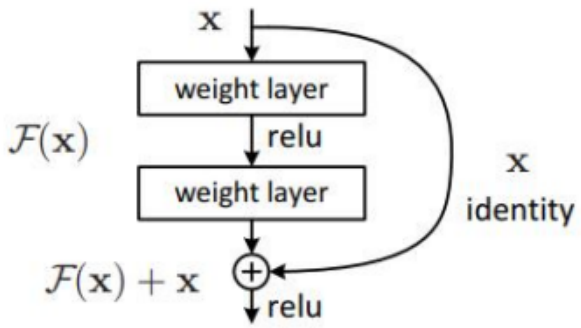


Figure 5: Residual Block Structure.

### Convolution-related calculations

The formula for calculating the width and height of the convolution output matrix is given by Here, denotes the floor function, which rounds down to the nearest integer.

$$\mathrm{w}_{out} = \frac{(w_{in} - k + 2p)}{s} + 1 \tag{1}$$

The formula for calculating the number of parameters in a convolutional layer is:

$$params = k_w \times k_h \times c_{in} \times c_{out} \tag{2}$$

The formula for calculating the number of floating-point operations (FLOPs) in a convolutional layer is:

$$FLOPs = c_{in} \times k \times k \times c_{out} \times w \times h \tag{3}$$

### Multimodal Fusion

Multimodal fusion methods mainly include feature-level, decision-level, and attention mechanism fusion(Huo G 2021), etc. The method used in this study is feature-level fusion, which includes simple concatenation fusion (Concate), tensor-based fusion (TFN), and low-rank tensor fusion methods (LMF), etc. The method used in this study is simple concatenation fusion (Concate). This fusion method concatenates features from different sources on the feature dimension. For example, suppose we have two feature vectors A and B, with A's dimension being (n, p) and B's dimension being (n, q), where n represents the number of samples, and p and q represent the feature dimensions of A and B, respectively. The Concate method concatenates A and B on the feature dimension to form a new feature vector C, with C's dimension being (n, p+q). Since the feature dimension becomes high after concatenation, a fully connected layer is usually followed to convert high-dimensional features into low-dimensional features. The weights of the fully connected layer can be learned through backpropagation, allowing the model to automatically learn better feature representations.

The TFN structure consists of three core components: the modal embedding sub-network, the tensor fusion layer, and the sentiment reasoning sub-network. The modal embedding sub-network is responsible for receiving unimodal features as input and converting them into informative modal embedding outputs; the tensor fusion layer uses the 3-fold Cartesian product of these modal embeddings to effectively simulate complex interactions between unimodal, bimodal, and trimodal; and the sentiment reasoning sub-network further processes sentiment reasoning based on the output results of the tensor fusion layer. Compared with TFN, Concate is simpler and more intuitive, retaining all feature information, but it may lead to increased model complexity and high training computational costs. TFN has a relatively high computational complexity and requires more computational resources.

Since the model used in this study is not complex and the dataset is not large, the chosen concatenation method is the simpler and more direct Concate.

## Experiments

### Dataset

We conduct experiments on the clothing subset of the CEPSUM 2.0 dataset (Li et al. 2020), which consists of 220k training samples, 10k validation samples, and 10k test samples. Each instance in the dataset consists of a pair of (product information, product summary), where the product information includes an image, a title, and additional product descriptions.

### Evaluation Method

We adopt ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy 2003) as the evaluation metric. ROUGE assesses the quality of summaries based on the overlap of n-grams between the generated summaries and reference summaries. It is a recall-oriented metric that evaluates how well the generated summary captures key elements of the reference summaries. Specifically, a set of expert-generated summaries is used to create a reference summary set, and the automatic summaries generated by the model are compared to these reference summaries. The overlap of fundamental units, such as n-grams, word sequences, and word pairs, is used to measure summary quality.

In this study, we primarily use ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) to evaluate the model performance.

### Technical Details

We implement the models using the PyTorch deep learning framework and train them on an NVIDIA GeForce RTX

4090 GPU. The pre-trained text model used is Randeng-BART-139M-SUMMARY, which is designed for processing Chinese text and is available on the Hugging Face platform. This model is implemented using the Python Transformers library. The BART model is trained using the AdamW optimizer (Loshchilov and Hutter 2017) with an initial learning rate of $10^{-4}$.

For image processing, we utilize the ResNet50 model, implemented using Python's torchvision library. The fusion method employed in this study is the Concate fusion approach, where image and text features are concatenated using `torch.cat`. The combined features are then passed through a fully connected layer for further integration.

## Comparison Experiments

To evaluate the performance of the proposed model, we compare it with several baseline methods. All experiments are conducted on the CEPSUM 2.0 dataset.

**Baseline Models** We choose the following baseline models for comparison:

- **LEAD**: A simple extractive summarization method that selects the first few sentences of a document as the summary. This model provides a strong baseline but lacks any deep semantic understanding of the text.
- **Textrank** (Mihalcea and Tarau 2004): A graph-based extractive summarization method that ranks sentences based on their similarity, using PageRank to prioritize important sentences. While effective for many tasks, it relies purely on sentence-level similarities without capturing deeper semantic meaning.
- **BERTSUMExt** (Liu and Lapata 2019): An extractive summarization method based on BERT that encodes sentences and classifies them for inclusion in the summary. BERTSUMExt leverages pretrained BERT representations, but it still does not fully capture multimodal information, which may be beneficial for tasks involving both text and images.

**Fusion Methods** In addition to comparing different models, we also explore various multimodal fusion strategies for integrating text and image features. The following fusion methods are considered:

- **Add Fusion**: This method simply adds the text and image features element-wise. While it is straightforward, it may not capture complex relationships between the two modalities effectively.
- **Concate Fusion**: This method concatenates the text and image features, allowing the model to process both modalities together. We hypothesize that this approach better captures the interactions between text and image features, leading to improved summarization performance.

## Results and Analysis

The results of different models are summarized in Table 1. The proposed Concate Fusion model achieves the best performance among all models across all ROUGE metrics. Specifically, it significantly outperforms the LEAD model, with improvements of 4.24, 2.14, and 3.57 percentage points in ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively.

Compared to TextRank and BERTSUMExt, the Concate Fusion model also demonstrates consistent gains: 2.69 percentage points higher in ROUGE-1, 1.49 in ROUGE-2, and 2.76 in ROUGE-L over TextRank; and 2.18, 1.36, and 3.12 percentage points higher over BERTSUMExt.

These results validate the effectiveness of the proposed multimodal fusion approach, illustrating that combining image and text features significantly enhances summarization quality.

Table 1: Comparative experimental results of different models

| Model/Fusion Method | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| LEAD | 20.92 | 6.58 | 13.95 |
| Textrank | 22.47 | 7.23 | 14.76 |
| BERTSUMExt | 22.98 | 7.36 | 14.40 |
| Add Fusion | 23.32 | 6.77 | 14.66 |
| Concate Fusion (Ours) | **25.16** | **8.72** | **17.52** |

# Conclusion

This paper integrates image information into text information and implements a product summary model that combines image information. The model uses BART and Resnet50 models to encode text and images respectively, and the two codes are fused in a concatenated manner. The text and image features are spliced along the last dimension, and the two-dimensional sequence is reshaped and converted into one dimension. The reshaped features are processed through a fully connected layer, so that the model can recognize the features of text and image at the same time. Finally, the multimodal feature representation of all input sentences is input into the extractive text summarizer to determine whether the sentence is a summary sentence, so as to extract the final summary. The experimental results show that the model is better than the baseline system that only uses text features.

## References


Baltrusaitis, T.; Ahuja, C.; and Morency, L. P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Devlin, J.; Chang, M. W.; Lee, K.; and et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Elounda, T.; Manning, S.; Mishkin, P.; and et al. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arxiv preprint arxiv:2303.10130*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Huo, G.; Zhang, Y.; Gao, J.; and et al. 2021. CaEGCN: Cross-attention fusion based enhanced graph convolutional network for clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(4): 3471–3483.

Huo G, G. J., Zhang Y. 2021. Cross-attention fusion based enhanced graph convolutional network for clustering. 3471–3483.

Kawakami, K. 2008. *Supervised sequence labelling with recurrent neural networks*. Master's thesis.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25.

Lewis, M.; Liu, Y.; Goyal, N.; and et al. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Li, H.; Yuan, P.; Xu, S.; Wu, Y.; He, X.; and Zhou, B. 2020. Aspect-Aware Multimodal Summarization for Chinese E-Commerce Products. *AAAI*, 34(05): 8188–8195.

Lin, C.-Y.; and Hovy, E. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 150–157.

Liu, Y.; and Lapata, M. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3730–3740. Stroudsburg, PA: Association for Computational Linguistics.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *ICLR*.

Mihalcea, R.; and Tarau, P. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. Stroudsburg, PA: Association for Computational Linguistics.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in neural information processing systems*.

Vaswani, A.; Shazeer, N.; Parmar, N.; and et al. 2017. Attention Is All You Need. *arXiv*, abs/1706.03762.