Hierarchical Cross-Modality Semantic Correlation Learning Model for Multimodal Summarization

Litian Zhang,[†] **Xiaoming Zhang,**[∗] **Junshu Pan**[†]

School of Cyber Science and Technology, Beihang University, Beijing, China
{litianzhang, yolixs, junshupan}@buaa.edu.cn

Abstract

Multimodal summarization with multimodal output (MSMO) generates a summary with both textual and visual content. Multimodal news report contains heterogeneous contents, which makes MSMO nontrivial. Moreover, it is observed that different modalities of data in the news report correlate hierarchically. Traditional MSMO methods indistinguishably handle different modalities of data by learning a representation for the whole data, which is not directly adaptable to the heterogeneous contents and hierarchical correlation. In this paper, we propose a hierarchical cross-modality semantic correlation learning model (HCSCL) to learn the intra- and intermodal correlation existing in the multimodal data. HCSCL adopts a graph network to encode the intra-modal correlation. Then, a hierarchical fusion framework is proposed to learn the hierarchical correlation between text and images. Furthermore, we construct a new dataset with relevant image annotation and image object label information to provide the supervision information for the learning procedure. Extensive experiments on the dataset show that HCSCL significantly outperforms the baseline methods in automatic summarization metrics and fine-grained diversity tests.

Introduction

With the rapid development of multimedia data on the Internet, multimodal summarization is a research direction worthy of attention and has broad development prospects. Therefore, some researches (Zhu et al. 2018; Li et al. 2020b; Zhu et al. 2020) focus on studying multimodal summarization with multimodal output (MSMO) to help readers improve reading efficiency and satisfaction. However, these approaches abstract the summary from the raw data directly, which is ineffective in learning the latent and vital information from both the text and image content.

Usually, the visual image and text article have heterogeneous structures. Directly mapping visual input and textual input as global vectors (Zhu et al. 2018; Li et al. 2020b; Zhu et al. 2020) is not effective to learn the important information for both modalities from each other, and even noisy information is added to decrease the performance of summarization. Previous experiments (Zhu et al. 2018) have shown that

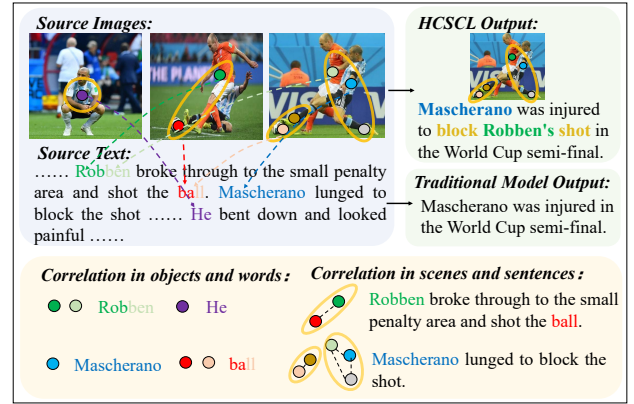∗Corresponding author.
†Equal contribution.

Figure 1: An example of the cross-modality semantics correlation in multimodal data.

multimodal input models may decrease summarization metric scores compared to text-only input models. Our experiments also show that some multimodal input model methods perform worse than traditional text-only input models. Therefore, one of the core problems of MSMO is how to effectively learn from each other modalities of data to obtain high-quality summaries.

Meanwhile, the correlation between the visual content and text article presents unique characteristics, providing clues to learning the important information from the two modalities complementarily to improve MSMO. As shown in Figure 1, the low-level objects in an image constitute the high-level semantics called scenes through the interaction between them. Therefore, by analyzing the objects and hence the scenes, we can know what the image describes from different levels. In the other data space, words are also the basic textual information in an article, while the combination of words, called sentences, present more abstract semantics information. Besides the intra-modal correlation, the semantics objects in the image and article are correlated in different levels. For example, in Figure 1, each person in the image may be related to a name in the article, and the football sport in the image is also described by a sentence in the article. By learning the inter-modal correlation, it can be known what is the important information in both

of the modalities. Moreover, even there are some incomplete descriptions in one modality, it can be learned from the other modality by exploiting the inter-modal correlation. As shown in Figure 1, we can generate the more complete information about the relation between the player "Mascherano" and the event "block Robben's shot". Therefore, by exploiting the hierarchical cross-modality correlation, we can extract the important information from both image and article more effectively.

However, there is still a great challenge to learn the hierarchical cross-modality correlation. First, different modalities have different feature spaces and structures among the elements. It is nontrivial to learn an effective representation to reflect both the different content and structure information. Second, much noisy information might exist, while some important information might be missed in one of the modalities. However, there is no explicit knowledge about the correlation between different modalities of data.

To tackle the challenge, we propose a novel Hierarchical Cross-Modality Semantic Correlation Learning model (HCSCL) to learn the intra- and inter-modality correlation for MSMO. In particular, two modality encoders are proposed to learn the intra-modal correlation for image and article, respectively. Then, a hierarchical fusion framework is proposed to learn the hierarchical correlation between image and article. A hierarchical attention method is proposed to combine the different levels of features learned by the hierarchical fusion framework to generate the summary. Furthermore, we construct a new dataset[1] with relevant image annotation to provide the supervision information for the learning procedure. Extensive experiments on the dataset show that HCSCL significantly outperforms the baseline methods in automatic summarization metrics and fine-grained diversity tests. Our main contributions are as follows:

- We propose a hierarchical learning model HCSCL to learn the intra- and inter-modality correlation in the multimodal data. To the best of our knowledge, this is the first work that guides multimodal summarization by exploiting the fine-grained semantics and their correlation information inside the multimodal data.

- We propose a multimodal visual graph learning method to capture the structure and the content information and reinforce the inter-modality interaction.

- We construct a large-scale multimodal summarization dataset with relevant image annotations and object labels to evaluate the performance of MSMO.

Related Work

Multimodal Summarization with Text Only Output. Unlike text summarization (Zhang et al. 2018; Xiao and Carenini 2019; See, Liu, and Manning 2017; Gao et al. 2019), multimodal summarization (UzZaman, Bigham, and Allen 2011) is defined as a task to generate a condensed summary from a multimodal input, such as text, image, and video. Several works focus on generating a better text summary with the help of images (Chen and Zhuge 2018; Li

---

[1]https://github.com/LitianD/HCSCL-MSDataset/

et al. 2018, 2020a). Palaskar et al. (2019) first release the How2 dataset for open-domain multimodal summarization. Liu et al. (2020) focus on reducing noise in longer text and video summarization and propose a forget gate method to control redundant information. Li et al. (2017) and Khullar and Arora (2020) aim to generate a text summary from more than three modalities: text-video-audio or text-image-audio-video.

Multimodal Summarization with Multimodal Output. Some other works generate a summarization containing both textual and visual content. Zhu et al. (2018) first propose to output a text summary and select the most relevant image. Then Zhu et al. (2020) add a multimodal loss function to improve the relevance between text summaries and images. Besides, Li et al. (2020b) and Fu, Wang, and Yang (2020) fuse the text-video feature and generate a text summary with a significant image from its associated video. Multimodal output summarization can facilitate readers to obtain crucial information efficiently. Most of these approaches fuse the different modalities of data directly, which neglects the latent correlation and heterogeneity of internal structure among them. Therefore, their performance of summarization is affected and even worse than text summarization.

Problem Formulation

Our HCSCL model takes a long article and the associated images as the input and generates a summarization containing both the textual summary and the most representative image. $X^t = \{S_1, S_2, ..., S_n\}$ is used to denote the textual input, which consists of n sentences, and $X^v = \{I_1, I_2, ..., I_p\}$ is used to denote the visual input associated p images. $S_i = \{x_{i,1}, x_{i,2}, ..., x_{i,m}\}$ means the sentence S_i has m words. The summarization can be divided into text output and visual output. The textual output is denoted as a sequence of words $Y^t = \{y_1, y_2, ..., y_t\}$, and the visual output is $Y^v = \{I\}$. In order to generate a multimodal summarization, the model can be formulated as an optimization problem as follows:

$$\arg\max_{\theta} HCSCL(Y^t, Y^v | X^t, X^v; \theta)$$

where θ is the set of trainable parameters in the model.

Model

The current multimodal summarization methods have two drawbacks: 1) They neglect the heterogeneity of internal structure between visual content and text. 2) They mainly consider the visual content as the whole, which ignores the hierarchical correlation between different modalities of data.

Therefore, in this work, we propose a hierarchical cross-modality semantic correlation learning model, as shown in Figure 2, to improve multimodal summarization. HCSCL comprises three modules: the Modality Feature Encoder is used to encode each modality, the Hierarchical Semantic Correlation Fusion module is used to learn the hierarchical intra- and inter-modality correlation, and the Multimodal Output Summarizer is used to generate the multimodal summary by exploiting the hierarchical correlation. Table 1 presents the key variables used in different modules.
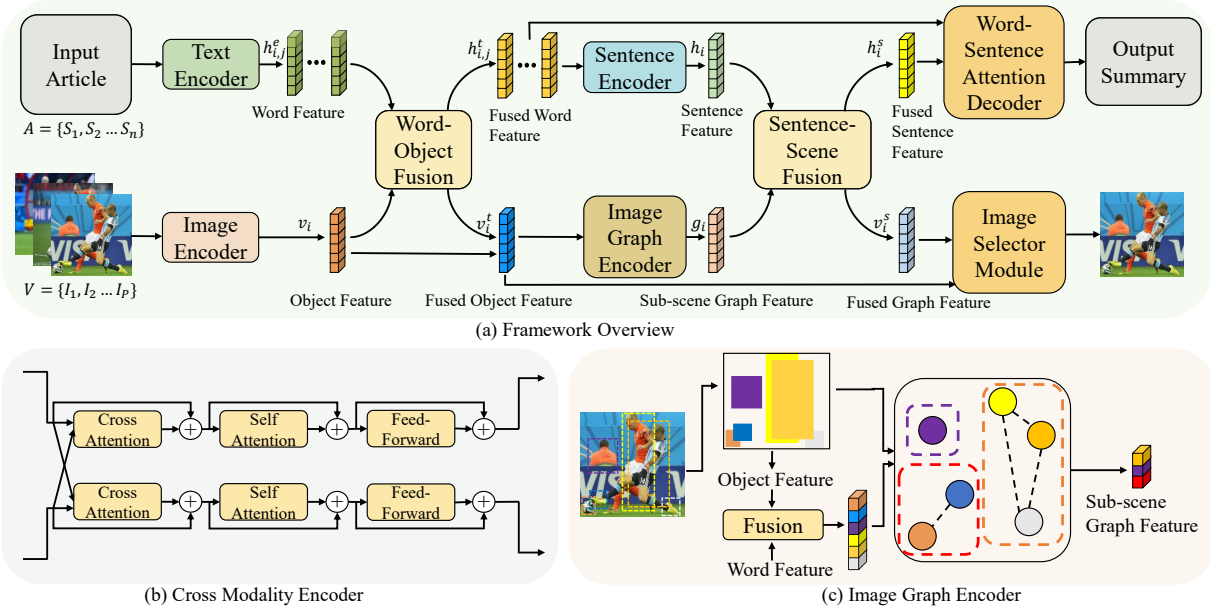
Figure 2: (a) is the overview of HCSCL. It consists of three parts: Modality Feature Encoder, Hierarchical Semantic Relation Fusion, and Multimodal Output Summarizer. (b) is the Cross Modality Encoder used for hierarchical fusion. (c) is the Image Graph Encoder to connect relevant objects into scenes.

	Var.	Description
Modality Feature	$h_{i,j}^0$	The jth word in the ith sentence
Encoder	v_i^0	The ith object feature for the article
Word-object	$h_{i,j}^1$	The jth word fused with object feature
Fusion	v_i^1	The ith object feature fused with word
	h_i^2	The ith sentence embedding for the article
Sentence-scene	v_i^2	The ith scene feature for the article
Fusion	h_i^3	The ith sentence fused with scene feature
	v_i^3	The ith scene feature fused with sentence

Table 1: Description of key variables.

Modality Feature Encoder

Text Encoder. We employ LSTM to encode an input article to a vector representation. Specifically, given an input article $X^t = \{S_1, S_2, \cdots, S_n\}$, the embedding of the words in the ith sentence $S_i = \{x_{i,1}, \cdots, x_{i,m}\}$ is learned as follows:

$$\overrightarrow{h_{i,j}^0} = LSTM(E[x_{i,j}], \overrightarrow{h_{i-1,j}^0}) \qquad (1)$$

$$\overleftarrow{h_{i,j}^0} = LSTM(E[x_{i,j}], \overleftarrow{h_{i+1,j}^0}) \qquad (2)$$

where $E[x_{i,j}]$ is the embedding vector of word $x_{i,j}$. $h_{i,j}^0 = [\overrightarrow{h_{i,j}^0}; \overleftarrow{h_{i,j}^0}]$ represents the hidden state of the jth word in the ith sentence. The ith sentence's word embedding is $\{h_{i,1}^0, h_{i,1}^0, ..., h_{i,m}^0\}$.

Image Encoder. Given a set of images $X^v = \{I_1, I_2, \cdots, I_p\}$ as the visual input, we apply a Faster R-CNN (Ren et al. 2016) initialized with ResNet-101 (He et al.

2016) to obtain object proposals for each image. For each object proposal, a triple set (o_i, l_i, a_i) is extracted. o_i is a feature vector extracted from the region of interest (ROI) pooling layer in the Region Proposal Network. l_i is a 4-dimensional bounding box location representation. a_i is a one-hot attribute class feature. Then we use the triple set to form a representation of the ith object as follows:

$$v_i^0 = Concat(o_i, W^l l_i, W^a a_i), \qquad (3)$$

where W^l and W^a are trainable embedding matrices. $\{v_1^0, v_2^0, ..., v_q^0\}$ is a set of q objects' features.

Hierarchical Semantic Correlation Fusion

Most existing approaches embed the whole image as a global vector, ignoring the internal and cross-modal correlation in the multimodal data. It is not effective to integrate the important information from different modalities. We propose a hierarchical semantic correlation fusion module, which can capture the hierarchical intra- and inter-modal correlation. It learns the important information in each modality by exploiting the intra-modal correlation at different grains and hence learns the important information in the multimodal content by exploiting the inter-modal correlation to reinforce each other, as Figure 2 shows. This module learns the correlation in two levels: word-object fusion and sentence-scene fusion. The word-object fusion component learns the correlation between words and objects to add visual information into the text entity features. In the sentence-scene fusion component, a visual relation graph is built on the objects to form a sub-scene graph fused with the sentences to learn the inter-modal correlation. Notably,

an attention-based cross-modality encoder (CME) (Tan and Bansal 2019) is employed to learn the inter-modality correlation, which enhances the critical information by each other.

Word-object Fusion. The framework of CME is shown in Figure 2(b). Instead of directly using the attention mechanism, we employ cross-modality attention to generate a fused representation for the multimodal content. CME consists of three parts: cross-attention layer, self-attention layer, and feed-forward layer.

To fuse the ith sentence's word embeddings $\{h_{i,1}^0, h_{i,1}^0, ..., h_{i,m}^0\}$ and objects' features $\{v_1^0, v_2^0, ..., v_q^0\}$, the cross-attention layer is defined as:

$$a_k = score(h_{i,j}^0, v_k^0) \tag{4}$$

$$\alpha_k = exp(a_k)/\sum_q exp(a_k) \tag{5}$$

$$CrossAtt_{h \to v} = \sum_q \alpha_k v_k^0 \tag{6}$$

where $h_{i,j}^0$ is a query word vector, v_k^0 is visual object vectors and $score$ is defined as product function to calculate the similarity value a_i. The three steps of CME is defined as follows:

$$h_{i,j}^{Cross} = CrossAtt_{h \to v}(h_{i,j}^0, \{v_1^0, v_2^0, ..., v_q^0\}) \tag{7}$$

$$h_{i,j}^{Self} = SelfAtt_{h \to v}(h_{i,j}^{Cross}, \{h_{i,j}^{Cross}\}) \tag{8}$$

$$\{h_{i,j}^{out}\} = FF(\{h_{i,1}^{Self}, h_{i,2}^{Self}, ..., h_{i,m}^{Self}\}) \tag{9}$$

where $h_{i,j}^{Cross}$ is the result after cross-attention layer, $h_{i,j}^{Self}$ is the result after self-attention and $FF(*)$ is the feed-forward layers. Residual connection and layer normalization are also added after each sub-layer. Three steps are repeated N_x times.

The inter-modal correlation between the textual words and the visual objects is learned by cross-modality attention. The learned multimodal representation is as follows:

$$h_{i,j}^1 = CME(h_{i,j}^0, \{v_1^0, v_2^0, ..., v_q^0\}) \tag{10}$$

$$v_i^1 = CME(v_q^0, \{h_{i,1}^0, h_{i,2}^0 ..., h_{i,m}^0\}) \tag{11}$$

where $h_{i,j}^1$ is the representation of the jth word in the ith sentence after the fusing procedure, and v_i^1 is the fused representation of the ith object. $\{h_{i,1}^1, h_{i,2}^1, ..., h_{i,m}^1\}$ is the word embeddings fused with object features and $\{v_1^1, v_2^1, ..., v_q^1\}$ is the object features fused with word vectors.

Sentence-scene Fusion. After obtaining the fused word features, for each sentence, we use LSTM to obtain a representation of the entire sentence. Then, the sentence representations are further fused with the visual feature by exploiting the correlation. Specifically, the representation of the ith sentence is computed as follows:

$$h_i^2 = LSTM(h_{i,1}^1, h_{i,2}^1, ..., h_{i,m}^1). \tag{12}$$

On the other side, a part of objects in an image is correlated, forming a scene to denote a more abstract concept or activity. The scenes are critical components of a summary. We propose an Image Graph Encoder to learn the scene representation, which can capture both the structure and the content information, as shown in Figure 2(c). First, based on the bounding box extracted by the image encoder, an Intersection over Union (IOU) score is calculated for every two objects. Next, a relation graph with adjacency matrix A is constructed, where $A_{ij} = 1$ if the IOU score exceeds the threshold value and $A_{ij} = 0$ else.

Then, we calculate the edge weight as follows. Given two target node feature v_i^1 and v_j^1, the feature score $s_{ij}^{feature}$ from node i to j is first calculated:

$$s_{ij}^{feature} = w_1^T \sigma(w_2 \cdot Concat(v_i^1, v_j^1)) \tag{13}$$

where w_1 and w_2 are learned parameters, σ is the activation function. For every directed edge of node i, we apply a softmax function over the IOU score s_{ij}^{IOU} and feature score $s_{ij}^{feature}$ to obtain the edge weight s_{ij}^{edge}:

$$s_{ij}^{edge} = exp(s_{ij}^{IOU} \cdot s_{ij}^{feature})/ \sum_{t \in N(i)} exp(s_{it}^{IOU} \cdot s_{it}^{feature}) \tag{14}$$

where $N(i)$ is the neighbors of i. Next, node feature is updated and combined with the connected node feature to formulate a sub-scene graph representation:

$$\tilde{v}_i^1 = \sigma(v_i^1 + \sum_{j \in N(i)} s_{ij}^{edge} A_{ij} w_3 v_j^1) \tag{15}$$

$$\{v_p^2\} = readout(\tilde{v}_1^1, \tilde{v}_2^1, ..., \tilde{v}_q^1) \tag{16}$$

where w_3 is the learned parameters, $readout$ is a node value aggregation function used to combine nodes' vectors to generate sub-scene graph vectors. $\{v_1^2, v_2^2, ..., v_p^2\}$ is a set of p scene vectors. We define the readout function as:

$$\{v_p^2\} = \frac{1}{N} \sum_{i \in N(i)} \tilde{v}_i^1 + Maxpooling(\tilde{v}_1^1, \tilde{v}_2^1, ..., \tilde{v}_q^1) \tag{17}$$

Finally, the sentence-scene fused features are generated by CME as follows:

$$h_i^3 = CME(h_i^2, \{v_1^2, v_2^2, ..., v_p^2\}), \tag{18}$$

$$v_i^3 = CME(v_i^2, \{h_1^2, h_2^2, ..., h_n^2\}). \tag{19}$$

where h_i^3 is the representation of the ith sentence after the fusing procedure, and v_i^3 is the representation of the ith scene. $\{h_1^3, h_2^3, ..., h_n^3\}$ is the set of sentence embedding fused with scene graph feature and $\{v_1^3, v_2^3, ..., v_p^3\}$ is the set of scene graph feature fused with sentence feature.

Multimodal Output Summarizer

The summarizer generates a text summary associated with the most relevant image. In the text summary generation, hierarchical attention is built to combine both word and sentence features. In the image selector, an object-scene gate mechanism is proposed to select an image as the visual output.

Text Summary Generation. In the hierarchical fusion module discussed above, visual objects and scenes features are respectively fused into words and sentences. In this summarizer, a hierarchical attention method is proposed to exploit the two levels of correlation in the decoder stage.

First, LSTM is used to decode sentence features, and the hidden state h_i' is generated as follows:

$$\{h_1', h_2', ..., h_t'\} = LSTM(h_1^3, h_2^3, ..., h_n^3) \qquad (20)$$

For the sentence level, the ith sentence weight β_i^{sent} is calculated as follows:

$$\beta_i^{sent} = \underset{i}{softmax}(score(h_i^3, h_{t-1}')) \qquad (21)$$

For the word level, the weight $\beta_{i,j}^{word}$ of the jth word in the ith sentence is calculated as follows:

$$\beta_{i,j}^{word} = \underset{i,j}{softmax}(\beta_i^{sent} \cdot score(h_{i,j}^1, h_{t-1}')) \qquad (22)$$

where h_i^3 and $h_{i,j}^1$ are sentence and word features respectively. Then the context vector c_t at timestep t and the word probability of y_t are calculated as follows:

$$c_t = \sum_{i=1}^{N} \sum_{j=1}^{M} \beta_{i,j}^{word} h_{i,j}^1 \qquad (23)$$

$$p(y_t|y_{1:t-1}) = softmax(V^T FF(h_t', c_t)) \qquad (24)$$

where V is the vocabulary weight matrix. The total loss of textual summary is calculated as follows:

$$L_{text} = -\sum_{t} \log p(y_t) \qquad (25)$$

Image Selector. The most relevant image should match the summary semantic in both the object level and the scene level. Therefore, the summary image is selected based on both the object features and the scenes graph features. First, the hidden state features h_t' are used to calculate the relevance score with object features v_i^1 and scene features v_j^3. Then, an object-scene gate mechanism is proposed to calculate the weights of the two types of features. The score of image I is calculated as follows:

$$s_i^{obj} = \sigma(FF(v_i^1 h_t')) \qquad (26)$$

$$s_j^{scene} = \sigma(FF(v_j^3 h_t')) \qquad (27)$$

$$\lambda^* = \sigma(FF(h_t')) \qquad (28)$$

$$s_I^{image} = \lambda^* \sum_{i \in I} s_i^{obj} + (1-\lambda^*) \sum_{j \in I} s_j^{scene} \qquad (29)$$

where $\sum_{i \in I} s_i^{obj}$ is the sum of all scores of the object features in image I, $\sum_{j \in I} s_j^{scene}$ is the sum of all scores of the scene graph features in image I, and λ^* is a balance weight between them. The image with the highest score is considered as the image output. The image summary loss is calculated as follows:

$$L_{image} = -\log(softmax(s_I^{image})) \qquad (30)$$

Finally, the total loss is calculated as follows:

$$L = L_{text} + \lambda L_{image} \qquad (31)$$

	Train	Valid	Test
Document Number	52656	5154	5070
Average Words Number	953.72	956.29	970.21
Average Sentences Number	19.52	19.57	19.86
Average Summary Length	36.66	36.33	36.36
Average Caption Length	25.07	25.10	25.10
Average Objects Number	7.25	7.16	7.32
Average Scenes Number	4.74	4.70	4.80

Table 2: Corpus statistics of our dataset.

Experimental Setup

Dataset

There is currently one dataset (Zhu et al. 2018) for MSMO tasks. However, this dataset is marked only on the test set due to the unsupervised image selection method. It is not suitable for our HCSCL, which needs the image objects information and annotating the most relevant image. Therefore, we construct a large-scale Chinese summarization dataset with complete image annotations and fine-grained object information. Specifically, documents and summaries are selected from TTNews (Hua, Wan, and Li 2017) and THUCNews (Sun et al. 2016), including sports, entertainment, current politics, society, technology, etc. For each selected document, we search it in Baidu Image Searcher[2], and then about ten images with their captions are crawled from the website. Next, we delete noisy images such as gifs, thumbnails, and advertisements.

Image annotation from large amounts of data and images is a time-consuming and laborious task. Since the images we crawled have image captions, we use automatic selection and manual selection methods for image annotation. A Bert (Devlin et al. 2018) and ESIM (Chen et al. 2016) semantic matching model pretrained on LCQMC corpus (Liu et al. 2018) is used to infer the relevance of article summary and image caption. Then, for automatic selection, we select the three images with the highest relevance in each document and discard the remaining images. For manual selection, five volunteers are employed to select one of the three images as the most relevant image. In addition, for each image, a Faster R-CNN feature extractor (Anderson et al. 2017) is applied to extract the object features, bounding box locations, and attribute classes. Then, the bounding box locations are used to calculate the IOU between objects to generate the adjacency matrices of relation graphs. Finally, a multimodal summarization dataset with complete image annotations and image object features is constructed, including 52656 for training, 5154 for validation, and 5070 for testing. More details about our dataset is illustrated in Table 2.

Baseline Models

We compare our model in three categories of baselines, a total of nine models.

Multimodal Summarization Output Models. **MSMO** (Zhu et al. 2018) is the first model on multimodal summarization with multimodal output task, which applies attention

---

[2]https://image.baidu.com/

	R-1	R-2	R-L	B-1	B-2	B-3	B-4	IP
Traditional Textual Model								
PG (See, Liu, and Manning 2017)	42.54	28.25	39.95	38.26	30.78	24.55	19.92	-
S2S (Luong, Pham, and Manning 2015)	30.13	14.40	28.61	26.36	17.55	11.07	7.44	-
TextRank (Mihalcea and Tarau 2004)	22.22	10.39	18.49	28.32	8.48	4.47	2.78	-
Multimodal Summarization Model								
MAtt (Li et al. 2018)	41.30	24.70	37.55	37.41	28.44	21.13	16.02	-
HOW2 (Palaskar et al. 2019)	39.53	20.28	35.55	35.44	24.42	17.81	13.71	-
MSE (Li et al. 2020a)	42.99	28.78	**41.79**	39.02	31.51	25.21	20.50	-
Multimodal Summarization Output Model								
MSMO (Zhu et al. 2018)	42.89	28.25	39.86	39.06	31.26	24.78	19.93	33.50
MOF (Zhu et al. 2020)	42.41	28.10	39.60	38.04	30.55	24.38	19.71	33.16
VMSMO (Li et al. 2020b)	42.68	28.35	41.34	38.75	31.20	24.75	19.99	32.86
our models								
HCSCL text output only	42.11	27.57	39.52	37.90	30.26	23.84	19.13	-
HCSCL multimodal output	**43.64**	**29.00**	40.94	**39.64**	**31.91**	**25.40**	**20.54**	**40.98**

Table 3: Rouge, BLEU and IP scores comparison with summarization baselines.

to text and images during generating textual summary and uses coverage mechanism to select image. **MOF** (Zhu et al. 2020) is the model based on MSMO, which considers image accuracy as another loss. **VMSMO** (Li et al. 2020b) is a video-based news summarization model, which proposes a Dual-Interaction-based Mutilmodal Summarizer.

Multimodal Summarization Models. **MAtt** (Li et al. 2018) is an attention-based model, which utilizes image filtering to select visual content to enhance sentence features. **HOW2** (Palaskar et al. 2019) the first model proposed to generate a textual summary by exploiting video content. **MSE** (Li et al. 2020a) is a model which focuses on enhancing the ability of the encoder to identify highlights of the news.

Traditional Textual Models. **PG** (See, Liu, and Manning 2017) is a sequence-to-sequence framework combined with attention mechanism and pointer network. **S2S** (Luong, Pham, and Manning 2015) is a standard sequence-to-sequence architecture using a RNN encoder-decoder with a global attention mechanism. **TextRank** (Mihalcea and Tarau 2004) is a graph-based extraction summarizer which represents sentences as nodes and uses edges to weight similarity.

Evaluation Metrics

As a widely used evaluation metric in text summarization, ROUGE (Lin 2004) and BLEU (Papineni et al. 2002) are applied to evaluate the quality of the generated text summary. Besides, image precision (IP) is used to evaluate the quality of the output image (Zhu et al. 2018), which is as follows:

$$IP = \frac{1}{N} \sum_i I(ann_i^{img} = rec_i^{img}) \qquad (32)$$

where ann^{img} and rec^{img} denote the annotated image and the output image respectively.

Implementation Detail

The experiments are conducted in Pytorch on NVIDIA Tesla V100 GPUs. We freeze the basic version of pre-trained Bert-base-Chinese (Devlin et al. 2018) for the original text em-

bedding, which has 12 layers, 12 heads, 768 hidden dimensions, and 110M parameters. The Faster R-CNN feature extractor (Anderson et al. 2017) is used for image object detection, and objects with confidence greater than 0.55 are selected. Each object has a 2048-dimensional representation. The IOU threshold is set to 0.2. To train the model, we employ Adam (Kingma and Ba 2014) as the optimizer, and the batch size is set at 16. The initial learning rate is set to $5e^{-4}$ and multiply by 0.8 every six epochs.

Results and Analysis
Overall Performance

To validate the effectiveness of our model, we compare our model with three kinds of baselines: Traditional Textual Model, Multimodal Summarization model, and Multimodal Summarization Output Model. Table 3 shows the result of comparative models on the dataset. We can obtain several conclusions from this table. First, it shows that HCSCL achieves state-of-the-art performance in almost all evaluation metrics. HCSCL outperforms baselines 1.51%, 0.76% in terms of Rouge-1, Rouge-2 and 1.48%, 1.27%, 0.75%, 0.20% in terms of BLEU-1, BLEU-2, BLEU-3, BLEU-4 and 22.32% in IP. It demonstrates the superiority of the hierarchical cross-modality correlation learning model. By exploiting the intra-modality correlation learning and the inter-modality feature aligning, the visual content can reinforce a specific part of the representation of text content, and the text content can reinforce to select the relevant image. Second, in Rouge-L, HCSCL (40.94) is slightly worse than the text output model MSE (41.79). After analyzing the cases, we find that the text outputs of HCSCL are more relevant to image semantics. Therefore, the longest common subsequence (R-L) is less matched with the text description, affecting this metric's evaluation.

Ablations Study

Four comparative experiments are designed to verify the effectiveness of word-object fusion, sentence-scene fusion,

	R-1	R-2	B-1	IP
HCSCL Word-Object only	41.70	25.37	38.01	32.31
HCSCL Sentence-Scene only	38.99	20.41	35.58	33.36
HCSCL *w/o* Sentence-Scene Fusion	41.05	24.04	37.44	31.94
HCSCL *w/o* Word-Object Fusion	38.35	19.94	34.97	30.83
HCSCL	43.64	29.00	39.64	40.98

Table 4: The ablation results for HCSCL. "Word-Object only" means a model only contains Text Encoder, Image Encoder, and Word-Object Fusion. "Sentence-Scene only" means a model only contains Sentence Encoder, Image Graph Encoder, and Sentence-Scene Fusion.

	HCSCL	PGN	MSMO	MSE	HOW2	MAtt	VMSMO	S2S
Ent.	4006	3972	3844	3778	3746	3691	3566	3489
Rel.	1911	1900	1836	1814	1801	1764	1702	1643

Table 5: The total number of name entities and relations in output text summaries.

and hierarchical structure. The result is shown in Table 4. We can see that with the hierarchical structure, the model has better performance. With the word-object fusion, the R-1, R-2, and B-1 are higher. With the sentence-scene fusion, the image selection metric IP is higher. However, when skipping the low-level word-object fusion module and using the sentence-scene fusion module directly, the model's performance worsens text evaluation metrics. It demonstrates that the low-level semantic correlation helps learn the high-level semantics and improve the summarization quality.

Fine-grained Semantic Diversity Analysis

To test fine-grained degree (Yang et al. 2020), we calculate the number of named entities and relations in the output text summaries. The pre-trained relation extraction model NEZHA (Wei et al. 2019) is trained on the LIC2021 Competition dataset[3], which extracts 24 types of name entities and 42 types of relations in our dataset. Then, the text summaries are input to the relation extraction model. Next, 24 name entities are divided into five categories: PER, ORG, LOC, WORK and OTHER, and 42 relations are divided into six categories: P-P, O-O, P-O, O-P, P-ATTR, and O-ATTR. The statistic result of the experiment is shown in Figure 3. Note that when the ROUGE is similar or slightly lower (as shown in Table 3), PGN can extract more name entities and relations than the other multimodal summarization models, which shows that the summaries generated by the previous multimodal summarization models have a certain loss of semantic information. Compared with the baselines, the summaries generated by HCSCL have the largest number of named entities and relations. From these results, we find that HCSCL can generate summaries with the richest semantic information. This is because HCSCL learns the correlation between the objects in the multimodal data, which is more effective in discovering the relationships and entities even if some critical information is missed in the text data.

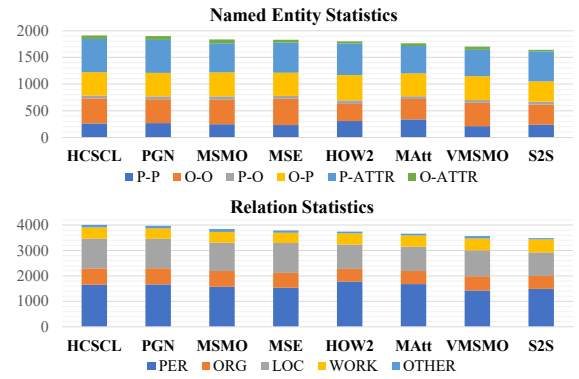[3]https://aistudio.baidu.com/aistudio/competition/detail/65

Figure 3: The histogram statistics of the number of named entities and the relations in output text summary.
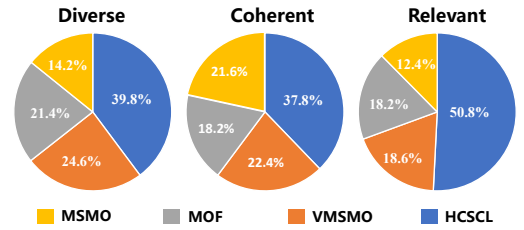

Figure 4: The pie charts for human evaluation.

Human Evaluation

To further evaluate our model's performance, 100 multimodal summaries generated by MSMO, MOF, VMSMO, and HCSCL are selected for human evaluation. Then five graduate students are volunteered to choose the most Diverse, Coherent, and Relevant (how the image matches the textual summary). The evaluation is shown in Figure 4. Compared with MSMO, MOF, and VMSMO, the multimodal outputs generated by HCSCL are the most diverse, coherent, and relevant, obtaining 39.8%, 37.8%, and 50.8% votes. It further demonstrates that HCSRF can learn the most representative and abundant semantic information from the article.

Conclusion

In this paper, we propose a Hierarchical Cross-Modality Semantic Correlation Learning Model for multimodal summarization, which exploits the intra- and inter-modality correlation in the multimodal data to learn the summary information from both modalities complementarily. The experimental results on the well-designed dataset show that our model can generate the most diverse and coherent text summary with the most relevant image. The novelty of this work is to tackle the multimodal summarization by proposing a multimodal model to learn the heterogeneous structure of different modalities, and hence the correlation between them. This complements the current research, which is not effective to mine the latent and important information inside the multimodal content.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2017. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering.

Chen, J.; and Zhuge, H. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4046–4056.

Chen, Q.; Zhu, X.; Ling, Z.; Wei, S.; Jiang, H.; and Inkpen, D. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fu, X.; Wang, J.; and Yang, Z. 2020. Multi-modal Summarization for Video-containing Documents. *arXiv preprint arXiv:2009.08018*.

Gao, S.; Chen, X.; Li, P.; Ren, Z.; Bing, L.; Zhao, D.; and Yan, R. 2019. Abstractive text summarization by incorporating reader comments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6399–6406.

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.

Hua, L.; Wan, X.; and Li, L. 2017. Overview of the NLPCC 2017 shared task: Single document summarization. In *National CCF Conference on Natural Language Processing and Chinese Computing*, 942–947. Springer.

Khullar, A.; and Arora, U. 2020. MAST: Multimodal Abstractive Summarization with Trimodal Hierarchical Attention. *arXiv preprint arXiv:2010.08021*.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer Science*.

Li, H.; Zhu, J.; Liu, T.; Zhang, J.; Zong, C.; et al. 2018. Multi-modal Sentence Summarization with Modality Attention and Image Filtering. In *IJCAI*, 4152–4158.

Li, H.; Zhu, J.; Ma, C.; Zhang, J.; and Zong, C. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1092–1102.

Li, H.; Zhu, J.; Zhang, J.; He, X.; and Zong, C. 2020a. Multimodal Sentence Summarization via Multimodal Selective Encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5655–5667.

Li, M.; Chen, X.; Gao, S.; Chan, Z.; Zhao, D.; and Yan, R. 2020b. VMSMO: Learning to Generate Multimodal Summary for Video-based News Articles. *arXiv preprint arXiv:2010.05406*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, N.; Sun, X.; Yu, H.; Zhang, W.; and Xu, G. 2020. Multistage Fusion with Forget Gate for Multimodal Summarization in Open-Domain Videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 1834–1845.

Liu, X.; Chen, Q.; Deng, C.; Zeng, H.; Chen, J.; Li, D.; and Tang, B. 2018. Lcqmc: A large-scale chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1952–1962.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404–411.

Palaskar, S.; Libovický, J.; Gella, S.; and Metze, F. 2019. Multimodal abstractive summarization for how2 videos. *arXiv preprint arXiv:1906.07901*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Sun, M.; Li, J.; Guo, Z.; Zhao, Y.; Zheng, Y.; Si, X.; and Liu, Z. 2016. THUCTC: An Efficient Chinese Text Classifier. *GitHub Repository*.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

UzZaman, N.; Bigham, J. P.; and Allen, J. F. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces*, 43–52.

Wei, J.; Ren, X.; Li, X.; Huang, W.; Liao, Y.; Wang, Y.; Lin, J.; Jiang, X.; Chen, X.; and Liu, Q. 2019. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*.

Xiao, W.; and Carenini, G. 2019. Extractive summarization of long documents by combining global and local context. *arXiv preprint arXiv:1909.08089*.

Yang, X.; Gao, C.; Zhang, H.; and Cai, J. 2020. Hierarchical Scene Graph Encoder-Decoder for Image Paragraph Captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4181–4189.

Zhang, X.; Lapata, M.; Wei, F.; and Zhou, M. 2018. Neural latent extractive document summarization. *arXiv preprint arXiv:1808.07187*.

Zhu, J.; Li, H.; Liu, T.; Zhou, Y.; Zhang, J.; and Zong, C. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 4154–4164.

Zhu, J.; Zhou, Y.; Zhang, J.; Li, H.; Zong, C.; and Li, C. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9749–9756.