Abstractive Sentence Summarization with Guidance of Selective Multimodal Reference

Anonymous Author(s)

ABSTRACT
Multimodal abstractive summarization with sentence output is to generate a textual summary given a multimodal triad – sentence, image and audio, which has been proven to improve users’ satisfaction and convenient our life. Existing approaches mainly focus on the enhancement of multimodal fusion, while ignoring the unalignment among multiple inputs and the emphasis of different segments in feature, which has resulted in the superfluous of multimodal interaction. To alleviate these problems, we propose a Multimodal Hierarchical Selective Transformer (mHsf) model that considers reciprocal relationships among modalities (by low-level cross-modal interaction module) and respective characteristics within single fusion feature (by high-level selective routing module). In details, it firstly aligns the inputs from different sources and then adopts a divide and conquer strategy to highlight or de-emphasize multimodal fusion representation, which can be seen as a sparsely feed-forward model - different groups of parameters will be activated facing different segments in feature. We evaluate the generalizability of proposed mHsf model with the pre-trained-fine-tuning and fresh training strategies. And Further experimental results on MSMO demonstrate that our model outperforms state-of-the-art baselines in terms of ROUGE, relevance scores and human evaluation.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence.

KEYWORDS
multimodal, abstractive summarization, selective routing

ACM Reference Format:
Anonymous Author(s). 2021. Abstractive Sentence Summarization with Guidance of Selective Multimodal Reference . In Proceedings of the 29th ACM International Conference on Multimedia (MM’21). ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnn.nnnnnn

1 INTRODUCTION
With the abundance of online multimedia information, there has been an increasing demand for short but effective messages. Many cross-media platforms rely on the text associated with the video or audio, such as news title. Hence, the advance of abstractive sentence summarization technology with multimodal inputs can avoid the waste of resources, reduce repetitive work and promote social development and human progress.

Generally, most existing summarization researches focus on unimodality (either sentences or images in isolation) [7, 8]. Recently, some researchers pay attention to summarizing multimodal inputs to the textual or multimodal outputs [7, 8]. Although the advances have been made in those researches, we find that two major problems are still in the existing methods: lack of cross-modal alignment and the emphasis of various segments in feature is the same. For the first pain point, the heterogeneities across modalities often increase the difficulty of analysing the interaction among language, image and audio. Figure 1 shows an example, the receptors for sentence, vision and audio streams may vary with variable receiving frequency, and then we may not obtain an optimal mapping between them. The caption, teammates and the celebrations in the figure have no connection with the text and audio. Hence, we need to align the features of multiple modals semantically, which is conducive to the expression of fusion features. All in all, the alignment of multiple inputs is the core issue we should solve. As for the second question (recall that the emphasis of various segments in feature is the same), it is a brand-new question but should be...
noticed. Like the sentence in Figure 1: the captain of the team is celebrating the victory with his teammates, we often focus on the objects caption and teammates that aligned with image or audio rather than other words. But in current methods, various semantic data share the same parameters. That is to say, they pay less attention to the deep semantics of each segment. This scheme can facilitate the unification of the model, but it is also a disaster for multimodal feature fusion. Even if the attention mechanism is used, it is still a simple weighted solution rather than the divide and conquer manner. A more ideal way is shown in Figure 1. We regard the aligned multimodal data as a cluster, and each cluster uses different sub-modules to represent its feature. Finally, the fusion features are concatenated together.

To address the above issues, we propose a Multimodal Hierarchical Selective Transformer (mHs) in this paper. It is an end-to-end model that extends the transformer [?] to learn representations directly from unaligned multimodal streams and divides different segments in various processing modules. The contributions are:

- A novel end-to-end multimodal abstractive summarization generation framework (mHsf) is proposed in this paper;
- Cross-modal interaction module is supported to directly represent fusion feature from unaligned multimodal streams;
- Selective routing module is proposed to consider various segments in different processing modules, which increase the number of parameters without loss of inference time.

2 RELATED WORK

Our research builds on previous works in following fields: multimodal fusion, text summarization and multimodal summarization.

2.1 Multimodal Fusion and Representation

Multimodal fusion is proposed to integrate multiple features, which can fuse multimodalities and be applied to the downstream tasks better. The Early Fusion (EF) [?] is one of the most popular technique but does not share any relevant information with itself. Then, Memory Fusion Network (MFN) [?] is proposed to address such shortcoming, which is a recurrent model consists of three modules to obtain dynamics. However, such method highly relies on attention mechanism. As is shown in the paper [?], Multi-attention Recurrent Network (MARN) is proposed to augment LSTM with a hybrid memory. Different from the above, some novel multimodal fusion techniques are Tensor Fusion Networks (TFN) [?], Low-rank Multimodal Fusion (LMF) [? ] and DeepCU [? ]. Those techniques perform multimodal fusion by utilizing the summarized information within other modalities as its average. However, few researchers pay attention to the problem of alignment of input from different sources when fusing multimodal inputs.

2.2 Text Summarization

[?] first propose a seq2seq (S2S) model to generate the summary from the paragraph. [?] and [?] incorporate the Copy mechanism and Pointer module into the seq2seq. Some researchers [?] focus on improving the semantic relevance by encouraging high similarity of their representation. [?] propose a discourse-aware hierarchical attention model and [?] present a contrastive attention mechanism to attend the irrelevant parts. Recently, with the development of pretrained then fine-tuning paradigm in nature language processing [? ]. Some researchers pay attention to improve the generation ability of pretrained model. [?] extend the mask mechanism (UniLM), so that the pretrained model can obtain better ability of sequential representation. GPT-2 [?] trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language tasks, such as summarization.

2.3 Multimodal Summarization

For the summarization task with multimodal data as input, the generated output can be unimodal or multimodal [? ]. Some researchers introduce available dataset for such task [? ]. [?] propose a seq2seq hierarchical attention-based technique, which combines the text and image features to produce the textual summary. But the actual performance is not useful. Recently, some methods of solving multimodal language task achieve huge process. In multimodal text classification task, Mult [?] has been proved to capture different modal information and apply in the field of generation. Furthermore, some pretrained methods are proposed in the field of multimodal processing. VL-BERT [?] adopts the powerful transformer model as the backbone, and extends it to take both visual and linguistic embedded features as considered. The UNiversal Image-Text Representation model (UNITER) is proposed by [? ], which can power heterogeneous downstream V+L tasks. Although the above research has made remarkable achievements in different fields, the ability of considering unaligned data have not been proved. Meanwhile, they regard the text and image as a whole rather than pay some attention to the various segments.

3 METHODOLOGY – MHSF

In this section, we describe our proposed Multimodal Hierarchical Selective Transformer (mHs, in Figure 2) for modeling unaligned multimodal streams. The core contributions are described in cross-modal encoder, which can fuse inputs with different sources (low-level cross-modal interaction) and respectively represent feature in suitable sub-modules (high-level selective routing).

3.1 Feature Representation

Given triple modal input \( \chi^{(S, I, A)} \), where \( S, I, A \) indicate the sentence, Image and Audio respectively. We use different processes for different modalities. In this paper, the Embedding methods are used to process the sequential modalities – sentence and audio. As is shown in Figure 2, text and audio are represented as \( e^{(S, A)} \in \mathbb{R}^{L(L(S,A) \times D(S,A))} \), where \( L \) represents the length of input and \( D \) means the dimension. The situation is completely different for images, because it has 3 dimensions in original image. We reshape the image \( X^I \in \mathbb{R}^{H \times W \times C} \) into a sequence of flattened 2D patches \( e^I \in \mathbb{R}^{L(I \times D_I)} \) by Patches Projection, which encouraged by the latest work of applying images to the transformer model [? ]. \((H, W)\) is the resolution of the original image, \( C \) is the number of channels, \( D_I = (P^2 \cdot C) \) indicates the projected dimension of image, where \((P, P)\) is the resolution of each image patch and \( L_I = HW/P^2 \) is the resulting number of patches.
3.2 Low-Level Cross-Modal Interaction

Cross-Modal Interaction module weighs a lot in mHSr, which can model and understand unaligned multimodal streams at low dimension. It is a transformer based module that merges multimodal timeseries including self/cross-modal attention, residual connection (Add in Figure 2) and layer normalization. Specially, self-attention is used to enhance the textual representation following the same learning the attention across the two modal features. Hence, the output of Cross-Modal Interaction module can be seen as three textual comprehension feature learned by sentence, image and audio, respectively. The details can be seen in Figure 2.

Cross-Attention is utilized to interact several inputs with different sources. Given two unaligned modalities α and β, our goal is to align semantic feature via query/key/value attention. We have $e^α \in \mathbb{R}^{Lα \times D_α}$ and $e^β \in \mathbb{R}^{Lβ \times D_β}$. We then project them to query, key and value vectors, i.e. $Q^α = e^α W^Q$, $K^β = e^β W^K$ and $V^β = e^β W^V$, where $W^Q \in \mathbb{R}^{D_α \times D}$ and $W^K, W^V \in \mathbb{R}^{D_β \times D}$ are weights and $D$ is a projected dimension. We assume that a good way to fuse cross-modal information is providing a latent adaptation from one modality β to another target α:

$$H^\text{cross}_{β→α} = CA_{β→α}(e^α, e^β)$$

$$= \text{softmax} \left( \frac{Q^α (K^β)^T}{\sqrt{D}} \right) V^β$$

(1)

where $H^\text{cross}_{β→α} \in \mathbb{R}^{Lα \times D}$ has the same length as $Q^α$, but is meanwhile represented in the feature space of $V^β$. Similar to the above description, Self-Attention considers the input α from same source. It can be simply written as $H^\text{self}_{α→α} = SA_{α→α}(e^α, e^α)$.

After feature representation across different modalities respectively, the fusion feature can be represented as formal 2, where $LayerNorm$ denotes the function of layer normalization [? ]:

$$e^α \text{SA} = [e^{αS}, e^{αS}, e^{αS}]$$

$$e^{αS} = e^α + \text{LayerNorm}(H^\text{self}_{S→S})$$

$$e^{JS} = e^α + \text{LayerNorm}(H^\text{cross}_{T→S})$$

$$e^{AS} = e^α + \text{LayerNorm}(H^\text{cross}_{A→S})$$

(2)

3.3 High-Level Selective Routing

In the cognition of deep learning, models typically reuse the same parameters for inputs. For example, words with different parts of speech are processed by same parameters in existing methods. In order to consider the various semantic segments in feature (the second problem we want to solve), this paper propose a selective routing (SR) method to divide the features and process them with different sub-modules. This method design intuitive improved models with reduced communication and computational costs. Suppose that a set $(E_i(x))_{i=1}^N$ of $N$ sub-modules is defined in advance when facing the input representation $x$. The router variable $W_r$ produces logits $h(x) \in W_r \cdot x$ which are normalized via a softmax distribution over the available $N$ sub-modules at that layer. The gate-value for sub-module $i$ is given by

$$p_i(x) = \frac{e_i^r(x)}{\sum_j^N e_j^r(x)}$$

(3)

To simply the selective routing, we only select one sub-module for the chosen sub-feature. The top-1 gate values are selected for routing the token $x$. Hence, the output of selective routing is the weighted value by $p_i(x)$ and sub-module result $E_i(x)$, which can formulate as

$$y_i = p_i(x) E_i(x)$$

(4)

This only-one routing strategy is referred as a selective routing layer (rounter in Figure 2), whose benefits are two-fold: (1) The
router computation is reduced because only route a token or one dimension. (2) The sub-feature with similar semantics are routed together, which make each sub-module have respective ability to enhance the representation. Figure 3 shows an example of routing with ideal or extreme condition. Later, we will introduce how to design the auxiliary loss to make SR achieve the ideal situation.

In this paper, we take each sub-module as FFN (FeedForward Network) layer (E[i] = FFN[j]) [7] with separately different parameters. Some researches have proved that the transformer will stop from degeneration if do not make use of residual connection and FFN layer [7]. If the fusion features e^SIA is gained by cross-modal interaction, the current selective routing method is to select more suitable sub-module to represent various tokens at high-level. The formula can be simply written as:

\[ H^0 = e^{SIA} + \text{LayerNorm}\left(\sum_{j \in \mathcal{E}} p_i(e^{SIA}) \cdot \text{FFN}_i(e^{SIA})\right) \]  (5)

### 3.4 Masked Sequential Decoder

In the decode stage, the input sentence \{x_i\}_{i=1}^{LS} (for simplification, \(X^S\) and \(x, x^S\) and x are same meaning in this section) is first packed into \(X^0 = [x_1^0, ..., x_{LS}^0]\), and then encoded into contextual representations at different levels of \(X^{[l]} = [x_1^{[l]}, ..., x_{LS}^{[l]}]\) using an L-layer Transformer \(X^{[l]} = \text{Trans}_l(X^{[l-1]})\). For the \(l\)-th Transformer layer, the output of self-attention \(A^{[l]}\) is computed via:

\[
Q = X^{[l-1]}W_Q^Q \cdot K = X^{[l-1]}W^K_V \cdot V = X^{[l-1]}W^K_V
\]

\[
M_{ij} = \begin{cases} 0, \text{allow to attend} \\ \infty, \text{prevent from attending} \end{cases}
\]

\[
A^{[l]} = \text{softmax}\left(\frac{QK^T}{\sqrt{D}} + M\right)
\]

where the previous layer’s output \(X^{[l-1]} \in \mathbb{R}^{LS \times D}\) linearly projected to a triple of queries, keys and values using matrices \(W_Q^Q, W^K_V, W^K_V \in \mathbb{R}^{D \times D}\), respectively. The mask matrix \(M \in \mathbb{R}^{LS \times LS}\) determines whether a token can be attended to each other. In detail, it controls what context a token can attend to when computing its contextualized representation like Figure 4. Through the description of formula (2), the output with the fusion hidden state e^SIA can be presented by

\[ H^{[L]} = A^{[L]} + \text{LayerNorm}\left(\text{CA}(A^{[L]}, e^{SIA})\right) \]  (7)

The final prediction result \(O\) can be obtained by full connection layer and softmax function, which can be regarded as a conditional probability calculation of multimodal fusion features e^SIA, input text \(X^S\) and the masked sequential matrix \(M\)

\[
O = p(o_i|e^{SIA}, x^S, x_{t-1}, M) = \text{Softmax}\left(W \cdot H^{[L]} + b\right)
\]  (8)

### 3.5 Loss Function

Reconstruction Loss \(L_{\text{MFS}}(S, O)\) is the negative log-likelihood assigned by the (decoder) of mMFS to the textual input. The mFS is calculated by formula (3), which is the fraction of the router probability allocated for expert \(i\)

\[
L_{\text{MFS}}(S, O) = -\sum_i^N \log p(S_i = 0) \]  (9)

We do not expect \(L_{\text{MFS}}(S, O)\) to decrease to zero. However, we expect it to drive the mMFS model to produce such sentences that will increase the likelihood of the target words.

**Balancing Loss.** To encourage a balanced load across experts we add an auxiliary loss to \(L_{\text{MFS}}\). It drive the selective routing module to separate load-balancing and importance-weighting losses. Given \(N\) experts indexed by \(i = 1\) to \(N\) and a batch \(X\) with \(T\) tokens, the auxiliary loss is computed as the scaled dot-product between vectors \(f\) and \(P\),

\[
L_{SR} = \alpha N \sum_{i=1}^N f_i \cdot P_i \]  (10)

where \(f_i\) is the fraction of tokens dispatched to expert \(i\),

\[
f_i = \frac{1}{T} \sum x \in \mathcal{X} \text{I}[\text{argmax}_x p(x), i] \]  (11)

where \(P\) is calculated by formula (3), which is the fraction of the router probability allocated for expert \(i\)

\[
P_i = \frac{1}{T} \sum x \in \mathcal{X} p_i(x) \]  (12)
\(\alpha\) is a multiplicative coefficient for auxiliary loss, whose value is \(10^{-2}\). For each sub-layer in mHsf, this auxiliary loss is added to the total model loss during training. Then the total loss function can be defined as

\[
\mathcal{L} = L_{\text{mHsf}} + L_{\text{SR}} + \frac{1}{\alpha}||\text{w}||_2
\] (13)

### 4 EXPERIMENTAL RESULTS

#### 4.1 Datasets and Metrics

With the huge development of pre-train + fine-tune paradigm in both computer vision and nature language processing. This paper uses two strategies (Fresh Training and Pre-trained then Fine-tuning) to verify the effectiveness of mHsf on MSMO [7] dataset. It contains online news articles (text) with multiple image-caption (5 images on average) and one audio segment (72 secs on average) triads. There are 240,000 training triads and 30,000 triads for both valid and test sets. The Pre-trained + Fine-tuning strategy needs amounts of data. Hence, we integrates four image-text data pairs sets (COCO, Visual Genome, Conceptual Captions, and SBU Captions [8]) as one Huge sets and carries out large-scale pre-training step. The details about how to pre-train mHsf on the above training datasets will be given in the later analysis.

To illustrate the effectiveness of mHsf model, we conduct the comparative experiments on Word-Overlap based, Embedding based metrics and human evaluation. The overall metrics are evaluated between the summary generated by mHsf and ground-truth. 1) ROUGE [7]: It is the standard metric, which calculates the ROUGE scores between the generated texts and reference. Due to different values of \(N\) in n-gram, ROUGE can be divided into ROUGE-1, ROUGE-2 and ROUGE-L.

2) Relevance [7]: We use embedding-based metrics to evaluate summary relevance, which are better correlated with human judgement than overlap-based metrics. In detail, Embedding Average, Embedding Extrema and Embedding Greedy are used to measure the effect of the generated summary.

3) Human: Eight native speakers were invited to judge the generated results with ground truth, whose ratio of male to female was 1:1. The metrics are human fluency (F) and human relevancy (R) scores, which are given on a scale 0-4.

#### 4.2 Implementation Details

We set word embedding size to 512 and transformer hidden sizes \(D\) to 1024. We use the limited vocabulary size to 30,004 (30,000 words and 4 special symbols) to reduce the cost of calculating. We also use dropout with probability equaling 0.3. The number of decoder layers \(L\) in both mHsf and baselines are 6. The batch size is up to 128 limited by the GPU (Nvidia 3090 with 24GB VRAM) and the overall parameters are trained 30 epochs for all baselines (only 10 epochs for fine-tuning baselines) and our model (20 epochs for pre-training step and fine-tuning 10 epoch also). We halve the learning rate when development performance worsens. We also use L2 regularization to constrain loss function like formula (13). The number of experts in mHsf is setup 3. In the post-processing process, we use beam search with size 3 to get more semantic combination. Finally, metrics ROUGE, Relevance and Human evaluation are calculated in the test set, which uses same python package for fairness.

### Table 1: Comparisons of parameters scale and inference speed with or without experts.

| Model       | Parameters | GPU Memory | Inference speed |
|-------------|------------|------------|-----------------|
| mHsf (1)    | 24.1M      | 14,754 MB  | 9.71 it/s       |
| mHsf (3)    | 29.5M      | 16,788 MB  | 9.03 it/s       |

#### 4.3 Quantitative Results

It should be noted that this paper compares the performance of mHsf model and baselines from two perspectives. One is whether the input is multimodal data, hoping to show that multimodal input can provide more reference information for the generated textual summary. The other one is whether to pre-trained, we hope to verify the importance of pre-trained model for specific tasks once again. Meanwhile, we hope to show that the mHsf model is not only useful for de novo training, but also for pre-training part of parameters firstly and then fine-tuning on specific tasks.

##### 4.3.1 Comparisons with Fresh Training Methods

The mHsf model proposed in this paper is trained by the start of random parameters, which named as Fresh Training. In order to evaluate the mHsf’s performance on the multimodal abstractive sentence summarization task, some baselines should be compared with. As the description of related work, we divide these methods into two categories: 1) single-modal input (text). Traditional methods such as S2S (sequence-to-sequence model using RNN) [7], S2S+Attn (S2S model with attention mechanism) and PointerNet [7]. For 2) multimodal input. Doubly-Attn [7] is the commonly used method, which uses various attention methods to focus on the commonness among modalities and generate the final summary. In order to show that the advancement of the proposed mHsf is not due to the excellence of the transformer, we also compare MultiT [7]. The part of baselines are advised by the open source website - NLPedia\(^1\).

For the Fresh Training version, the comparisons between mHsf and baselines are shown in Table 2 (top subtable). Our model outperforms all baselines on Rouge. Embedding metrics and Human evaluation (only exception: Embedding Extrema). On the whole, considering multimodal data as input is better than only considering textual input. On the one hand, multimodal-based method is proposed recently; On the other hand, multimodal inputs can provide more reference for the generation of final summary. Therefore, it is inevitable that the final metrics will be better. mHsf model is better than most baselines, which shows the effectiveness of the proposed scheme. At the same time, the result is better than MultiT showing that the proposed method can better fuse multimodal reference and has better ability of generation, rather than the superiority of transformer model. It is mentioned that the selective routing proposed in this paper can extend the scale of model’s parameters without losing the time of inference. As is shown in Table 1, using SR (mHsf with \(N=3\) sub-modules) increases the model parameters nearly by 49% compared with non-SR (mHsf with \(N=1\) sub-module), while the inference time is only lost by 7%.

---

\(^{1}\)http://explainaboard.nlpedia.ai/.
4.3.2 Comparisons with Pre-train & Fine-tune Methods. Large-scale pretraining and task-specific fine-tuning is now the standard methodology for many tasks in computer vision and natural language processing. Considering the mixed huge sets mentioned above, we only extract the former module (cross-modal encoder) in mHsf to fuse multimodal reference and connect them with the backend of VLBERT to complete the task of text-image retrieval. Following that, we pre-train the mixed mHsf model on huge sets with 20 epochs. During the fine-tuning step, we finely update the whole mHsf’s parameters on MSMO dataset (only 10 epochs). While part of its parameters (encoder’s parameters) are already pre-trained and the masked sequential decoder should be trained at the beginning of random parameters (named as fresh training). For the sake of fairness, this paper updates 10 epochs for the textual pre-trained models GPT-2 2, UniLM 3 and multimodal pre-trained models VLBERT 4, UNITER 5.

For the Pre-Trained & Fine-tuning version, the comparisons between mHsf and baselines are shown in Table 2 (bottom subtable). It can be seen that no matter compared with the single-stream pre-trained model or the multi-streams pre-trained model, the mHsf model has a great improvement in both metrics and human evaluation. On the one hand, it may be that the pre-trained parameters are directly used in the above existing models and then result in the worse performance than mHsf, which we did some ablation studies to verify this suspicion. On the other hand, mHsf’s powerful decoder provides strong support for the generation of final summary. Meanwhile, compared with the Fresh Training strategy, the performance of the Pre-trained & Fine-tuning strategy in the mHsf model is improved by nearly 13%, which proves the importance of pre-trained mode again.

### Table 2: Performance of mHsf and Baselines on MSMO Dataset with ROUGE/Relevance/Human.

| Type                  | Resource       | Methods         | ROUGE   | Relevance | Human |
|-----------------------|----------------|-----------------|---------|-----------|-------|
|                       | text-only      |                 | R-1     | R-2       | R-L   | Average | Extrema | Greedy | F     | R     |
| Fresh                 |                | S2S             | 29.84   | 11.05     | 26.68 | 0.187   | 0.171   | 0.259 | 2.98  | 3.01  |
| Training              |                | S2S+Attn        | 32.32   | 12.48     | 29.65 | 0.206   | 0.182   | 0.287 | 3.24  | 3.36  |
|                       |                | PointerNet      | 34.78   | 13.1      | 32.24 | 0.231   | 0.208   | 0.31  | 3.21  | 3.57  |
|                       |                | Doubly-Attn     | 41.78   | 27.81     | 40.28 | 0.362   | 0.269   | 0.389 | 3.13  | 3.27  |
|                       |                | MuT             | 47.28   | 30.11     | 45.21 | 0.392   | 0.287   | 0.401 | 3.32  | 3.41  |
|                       |                | ours            | 49.11   | 30.97     | 46.37 | 0.416   | 0.278   | 0.422 | 3.58  | 3.64  |
| Pre-Trained & Fine-tuning |                | GPT-2           | 39.16   | 28.16     | 37.01 | 0.379   | 0.273   | 0.381 | 3.41  | 3.37  |
|                       |                | UniLM           | 43.82   | 30.78     | 38.68 | 0.392   | 0.282   | 0.399 | 3.47  | 3.36  |
|                       |                | VLBERT          | 49.72   | 30.09     | 43.11 | 0.397   | 0.285   | 0.412 | 3.32  | 3.31  |
|                       |                | UNITER          | 52.17   | 32.15     | 46.78 | 0.421   | 0.321   | 0.443 | 3.37  | 3.51  |
|                       |                | ours            | 56.11   | 36.97     | 49.71 | 0.452   | 0.369   | 0.478 | 3.52  | 3.61  |

### 4.4 Ablation Studies

As the description of methodology and quantitative analysis, the low-level cross-modal interaction, high-level selective routing and masked sequential module are likely to contribute to the advancements. Hence, we pay attention to those modules in ablation studies. The low-level cross-modal interaction module is replaced by a simple multimodal fully connection projection to verify the effectiveness of this module, which is marked as mHsf-CMI. The selective routing process is removed, and the concatenation of multimodal feature is directly input into the decoder. This operation is marked as mHsf-SR. We can also set the number of sub-modules to 1 and sign as mHsf-\(\downarrow SR_1\). Finally, the effectiveness of the proposed masked sequential strategy is verified by removing the mask strategy, which makes the model degenerate into a pure transformer model (mHsf-mask). The comparison results are shown in Table 3.

Considering the different pre-trained datasets may have a huge impact on final results. In this paper, UniLM in single-stream and VLBERT in multi-streams are taken as examples to retrain the model instead of using the existing pre-trained parameters. Table 4 shows the results of VLBERT by using the multimodal data in huge sets and UniLM by using only the textual data in huge sets (where the scale is reduced by 32% compared with the original paper).

### Table 3: Ablation results where we measure the impact of CMI, SR and masked sequential strategy.

| Methods         | ROUGE   | Relevance |
|-----------------|---------|-----------|
|                 | R-1     | R-2       | R-1     | Average | Extrema | Greedy |
| mHsf            | 49.11   | 30.97     | 46.37   | 0.416   | 0.278   | 0.422  |
| mHsf-\(\downarrow CMI\) | 36.32   | 23.78     | 34.11   | 0.351   | 0.223   | 0.343  |
| mHsf-\(\downarrow SR_1\) | 43.71   | 27.82     | 39.16   | 0.392   | 0.280   | 0.401  |
| mHsf-mask       | 39.28   | 26.12     | 37.81   | 0.381   | 0.267   | 0.391  |

---

2Gpt-2 source code and pre-trained params from https://github.com/openai/gpt-2.
3UniLM source code from https://github.com/microsoft/unilm.
4UniLM pre-trained params from https://github.com/jackroos/VL-BERT.
5UNITER pre-trained params from https://github.com/ChenRocks/UNITER.
Li Na is destined for more grand slams and a place in the Hall of Fame after the Australian Open win cemented her position as the most influential female player of the last decade. Li’s final victory over Slovak Dominika Cibulkova gave her a second major following her 2011 French Open title, a triumph that made her Asia’s first singles grand slam champion.

Li was worthy of a place alongside her in the Hall of Fame. It’s not only about winning grand slams its about the influence that you have in tennis. Stacey said she was the most influential women’s tennis player in the last 10 years, with what she has done for global tennis so absolutely 100 percent. She is right up there with them too. There was always a little gap before you said Li Na’s name but now I think she is right up there with all of them after the type of tennis she played at the Australian Open.

**Table 4: Ablations with the impact of pre-training corpus.** ↓ means the decline of metrics, while ↑ means the opposite

| Methods | ROUGE R-1 | ROUGE R-2 | ROUGE R-L | Relevance Average | Relevance Extrema | Relevance Greedy |
|---------|-----------|-----------|-----------|-------------------|-------------------|------------------|
| mHsf    | 56.11     | 36.97     | 49.71     | 0.452            | 0.369             | 0.478            |
| UniLM ↓ | 41.67     | 28.92     | 37.19     | 0.382            | 0.261             | 0.382            |
| VLBERT ↑ | 51.07     | 31.19     | 45.42     | 0.412            | 0.292             | 0.418            |

**4.5 Case Studies**

In the section, we show some successful cases from the MSMO dataset. The details (both the results of baselines and mHsf) are shown in Figure 5, which includes the original textual paragraphs, image data and the generated summary. It can be seen intuitively that the summary gains textual references (words with gray background) from the original text. In the cross-modal interaction processing, related words (signed as red) and corresponding patches (As is shown in Figure 5 (b), the bright areas indicate the focus of attention) are captioned by attention mechanism. In this paper, we set the number of sub-modules to 3. As is shown in Figure 5 (c), green, pink and yellow shadows show the selective parts of mHsf model. Intuitively, FFN-2 is most like to process target or highlight patches by one feed-forward layer. However, the other two sub-modules use different feed-forward layers (FFN) to focus on the background parts, which meet the description of selective routing module. Compared with the textual summary generated by baselines, the result generated by mHsf is longer (contains more information) and has no obvious mistakes (the blue part indicates possible errors in the result).

**5 CONCLUSION**

This paper address a multimodal abstractive summarization task, that is, how to generate a textual summary from multimodal data. The proposed mHsf module with low-level cross-modal interaction & high-level selective routing (hierachical) outperforming baselines on the MSMO dataset shows that 1) Cross-modal encoder can effectively fuse multimodal streams and gain additional information from each modalities; 2) The masked sequential decoder can generate topic-related, fluent and readable summary.