**ICAF: Iterative Contrastive Alignment Framework for Multimodal Abstractive Summarization**

Zijian Zhang*\(^a\), Chang Shu*\(^a, b, c\), Youxin Chen\(^b\), Jing Xiao\(^b\), Qian Zhang\(^c\) and Lu Zheng\(^c, \dagger\)

\(^a\) Meituan-Dianping Group, Shanghai, China
\(^b\) Ping An Technology (Shenzhen) Co., Ltd, Shenzhen, China
\(^c\) University of Nottingham Ningbo China, Ningbo, China

zijian.zh96@gmail.com, \{scxcs1, qian.zhang, zheng.lu\}@nottingham.edu.cn, \{chenyouxin149, xiaojing661\}@pingan.com.cn

**Abstract**—Integrating multimodal knowledge for abstractive summarization task is a work-in-progress research area, with present techniques inheriting fusion-then-generation paradigm. Due to semantic gaps between computer vision and natural language processing, current methods often treat multiple data points as separate objects and rely on attention mechanisms to search for connection in order to fuse together. In addition, missing awareness of cross-modal matching from many frameworks leads to performance reduction. To solve these two drawbacks, we propose an Iterative Contrastive Alignment Framework (ICAF) that uses recurrent alignment and contrast to capture the coherences between images and texts. Specifically, we design a recurrent alignment (RA) layer to gradually investigate fine-grained semantical relationships between image patches and text tokens. At each step during the encoding process, cross-modal contrastive losses are applied to directly optimize the embedding space. According to ROUGE, relevance scores, and human evaluation, our model outperforms the state-of-the-art baselines on MSMO dataset. Experiments on the applicability of our proposed framework and hyperparameters settings have been also conducted.

**Index Terms**—multimodal abstractive summarization, recurrent alignment, contrastive learning

**I. INTRODUCTION**

The goal of Multimodal Abstractive Summarization (MAS) is to compress information from interacting modalities into a brief, simple, and legible summary [1]. With the growing number of videosharing platforms and multimodal data, MAS can help consumers rapidly obtain the information they need. Due to semantic gaps between vision and language is an effective strategy. As comprehension gap between vision and language. As seen in Figure 1, it is natural for human to focus more on the word *celebrating* that is caused by the semantic corresponding regions in the image (indicated by blue boxes). Since current approaches use different structures for multimodal data in vision and language [4], [5], i.e., the inconsistency between image and text models makes them less concerned with the deep semantics of each cross-modal corresponding counterpart segment. Large-scale text generation models based on transformers, such as UniLM [6], BART [7], or ProphetNet [8], have recently demonstrated outstanding performance on generation tasks [9]. Although leveraging and adapting these techniques to MAS is still an ongoing research topic that yet to be fully explored [10], there have been several attempts. For example, CtnR [11] aimed to unsupervisedly align various modalities while merely capturing shallow information. Using selected encoders to represent modalities and summarize information, Li et. al. [12] attempted to bridge the gap but struggle to produce image-text corresponding captions. Zhu et. al. [13] tackled MAS task with graph based intra modal reasoning method. Oscar [14] integrated the contrastive learning to learn cross-modal representation on image-text pairs, but it is contingent on the a priori condition that the object tags in image can be accurately detected and appear in the text. To summarize, existing attention-based or reasoning-based models have difficulty in capturing the fine-grained pairwise associations among large number of region-word fragment pairs because of the obvious heterogeneity gap between image and text features. In our opinion, using comparable model structures to unify the representation and shrink the semantic gaps between vision and language is an effective strategy.

**Lack of awareness of image-text matching.** Natural matches exist between each other due to the pairing of...
images and sentences in the existing dataset [15], [16]. While being beneficial to the training process, this decreases the generalization of the models and inhibits further model performance improvements [17]. If the words caption and teammates (yellow surrounding) are not taken into account, the sentences that correspond to two images in Figure 1 may believe that the bottom image is a better one and compatible with the phrases. Li et. al. [18] take image-text matching as a prior step before multimodal summarization to solve this problem. In essence, this prevents the model from distinguishing useful data from those are irrelevant. Entity matching is also considered in the abstractive generation via bottom-up attention [19], where entities should be well described otherwise performance would suffer. In a fine-grained task, i.e., cross-modal retrieval [20], hard triplet loss [21] is used to enhance the corresponding representation across vision and language. With similar ideas, MultimodalSum [22] proposes to generate summary from multiple reviews in a self-supervised manner. In general, these methods have focused on the awareness of matching, but have not been attempted in multimodal summarization, especially the semantic matching from bottom up.

Motivated by the potential of solving these issues, this paper seeks to guide multimodal summarization with iterative alignment and progressively matching corresponding image-text pairs through contrastive learning, termed ICAF. Specifically, we first propose a recurrent alignment (RA) layer that investigates the correspondence between vision and language for summarization via two modules: 1) an iterative cross-modal attention module (CAM) to align fragments across image and text; and 2) a renovation addition module (RAM) to integrate the current alignment with history knowledge. The iterative alignment approach may gradually update cross-modal attention in order to amass clues for identifying matching semantics and improving cross-modal information interaction. However, employing only attention-based mechanism to align cross-modal semantics yields inadequate results. As a solution, contrastive losses [23] are applied after each RA layer to keep the embedding space constant and maintain distance from other references. Aggregating multiple contrastive objectives may also be used to directly supervise the learning of image-text correspondences, while accelerating model convergence and increasing performance at the same time. By taking aforementioned proposed strategies, we show that ICAF can produce a more illustrative summary and achieve the state-of-the-art results empirically. In conclusion, the contributions of this paper are:

- a modified recurrent alignment layer in the encoder is proposed to handle the complexity of semantics, which incorporates CAM for aligning information and RAM for aggregating knowledge.
- two auxiliary contrastive losses are used in encoder, with image-to-text and text-to-image objectives, respectively.
- we thoroughly verify our ICAF on benchmark dataset and evaluate the settings of hyperparameters and novel units in details.

II. METHODOLOGY

In this section, we will elaborate on the details of our proposed ICAF for multimodal abstractive summarization task. First, we introduce our way of multimodal (Image&Text) feature representation in Section II-A. Then, in Section II-B, we show the structure of RA and how it can be implemented into the ICAF as a module (or layer). The interactive contrastive based InfoNCE losses between text and image embeddings, are explained in Section II-C. Finally, the total objective function is described in Section II-D.

A. Image&Text Representation

**Text representation** To achieve the fine-grained connection between vision and language, we extract word-level features by combining textual tokens with positional embeddings. Specifically, for a given sentence with $N$ words, we firstly project the sentence with a contiguous learnable embedding matrix $W_{tok} \in \mathbb{R}^{V \times D}$, where $V$ and $D$ are vocabulary size and embedding dimension respectively. Together with statistic positional embedding [24], the text representation is denoted by $T \in \mathbb{R}^{N \times D}$.

**Image representation** Conventionally CNN-based image representation may not work well in the transformer module [25], [26]. Encouraged by [27], we reshape the original image $I \in \mathbb{R}^{C \times H \times W}$ ($C$, $H$, $W$ denote number of channels, height, and width, respectively) into a sequence of flattened 2D patches $I \in \mathbb{R}^{M \times D'}$ using patches projection. $M$ denotes the number of regions (or patches) and $D'$ denotes embedding dimension for vision representation. In order to maintain the comparability of image and text features in the subsequent operations, we map the original dimension $D'$ to $D$ by learnable projection.

B. Recurrent Alignment Layer

The encoder of ICAF consists of multiple RA layers that aim to align fragmental features (token or patch embeddings) by refining the current state and previous knowledge in a recurrent manner. The former seizes the corresponding information between image and text features and the latter discards trivial knowledge. RA can be regarded as a block (see Figure 2) that takes two inputs, i.e., $T$ and $I$. For current state, RA focuses on the alignment among input pair by CAM. For previous knowledge, RA concerns more about how to update the state by RAM.

For generalization, we denote the two input sets of features as $X = \{x_i | i = 1, \ldots, n\}$ and $Y = \{y_i | i = 1, \ldots, m\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}^d$ have the same hidden dimension. Note that $X$ can be either $T$ or $I$, while $Y$ is the other. The visualization flows described below is also shown in Figure 3.

**Cross-modal Attention Module (CAM)**. The CAM aims to account for global information in $Y$ for each local feature $x_i$ in $X$. To achieve this goal, we first compute the cosine similarity between each pair $(x_i, y_j)$ followed by a standard self-attention module to enhance inter information:

$$s_{ij} = \frac{x_i^T y_j}{||x_i|| \cdot ||y_j||}$$ (1)
As in [4], we further normalize the similarity score $s$ and filter high relevance items by adding an adjustable factor $\gamma$:

$$s_{ij} = \text{relu}(s_{ij} + \gamma) / \sqrt{\sum_{i=1}^{n}(s_{ij} + \gamma)^2}$$  \hspace{1cm} (2)

where $\text{relu} = \max(x, 0)$.

To seize the alignment information from $Y$ to $X$, attention-based mechanism [28] is applied by:

$$c_i^x = \sum_{j=1}^{m} \alpha_{ij} y_j$$  \hspace{1cm} (3)

$$s.t. \quad \alpha_{ij} = \frac{\exp(\lambda s_{ij})}{\sum_{j=1}^{m} \exp(\lambda s_{ij})}$$

where $\lambda$ is the inverse temperature of the softmax function [29] to adjust the smoothness of the attention distribution.

We define $C_x = \{c^x_i| i = 1, ..., n\} \in \mathbb{R}^{n \times d}$ as X-target alignment features. Each element $c^x_i$ captures related semantics contributed by the reference feature $Y$.

**Renovation Addition Module (RAM).** To refine the alignment feature for the next layer in the encoder, which updates the current state by previous knowledge, RAM is to aggregate the input feature $X$ with alignment feature $C_x$ dynamically:

$$x_i^* = f(x_i, c^x_i)$$  \hspace{1cm} (4)

where $f()$ is an aggregating function. We adopt a gate mechanism [30] for $f()$:

$$\alpha_i = \text{sigmoid}(W_\alpha [x_i, c^x_i] + b_\alpha)$$

$$u_i = \tanh([W_u x_i, c^x_i] + b_u)$$

$$x_i^* = (1 - \alpha_i) \ast x_i + \alpha_i \ast u_i$$  \hspace{1cm} (5)

where $W_\alpha, W_u, b_\alpha, b_u$ are learnable parameters. $u_i$ is a fused feature, which enhances the interactive information between previous knowledge $x_i$ and the current state $c^x_i$. And $\alpha_i$ is a forgetting coefficient that is able to filter trivial information in previous knowledge $x_i$ and focus more on the alignment information shared from $Y$.

**Recurrent Alignment (RA) Layer.** We integrate CAM and RAM into a RA layer, whose $k^{th}$ layer outputs are:

$$C_{k}^x = \text{RA}(X_{k-1}^*, Y), \quad k = 1, ..., K$$  \hspace{1cm} (6)

where the details of $C_k^x$ and $X_k^*$ are shown in Eq. 3 and Eq. 5.
C. Iterative Contrastive Alignment Framework

Considering mini-batch $B = \{ (T, I) \}$ with $|B|$ pairs in the whole dataset $D$, the multimodal abstractive summarization task can be framed as a seq2seq learning problem with extra iterative objectives in this work (more details in Figure 2). The ICAF with a modified encoder and a vanilla transformer decoder is the backbone architecture, where the model receives the utterances and image as input and outputs a corresponding textual summary $\{<sos>,...,w_i,...,<eos>\}$. The encoder is stacked with K-layers RA, interacting cross-modal features (soft-alignment) and passing to the next step.

$$C_k^T, T_k = RA(T_{k-1}, I)$$

$$C_k^I, I_k = RA(I_{k-1}, T)$$

(7)

where $C_k^T$ and $C_k^I$ indicate step-wise alignment features of text $T$ and image $I$, respectively. $T_k$ and $I_k$ are the $k^{th}$ iterative knowledge in the encoder. In the beginning, $T_0 = T$, $I_0 = I$ represent the fragments embedding detailed in Section II-A. Following the motivation aforementioned, we expect that the corresponding image and text pairs have high consistency, while the irrelevant pairs are hard to capture, especially those fine-grained interplay. To achieve this goal, we accumulate the contrastive losses advised by infoNCE [23] in each RA layer directly.

$$L_{T2I}^B = - \frac{1}{|B|} \sum_{i=1}^{|B|} \sum_{k=1}^K \log \frac{e^{\tau \text{sim}(f_{pl}(C_k^T, I_i), f_{pl}(I_k))}}{\sum_{j=1}^L e^{\tau \text{sim}(f_{pl}(C_k^T, I_i), f_{pl}(I_j))}}$$

$$L_{I2T}^B = - \frac{1}{|B|} \sum_{i=1}^{|B|} \sum_{k=1}^K \log \frac{e^{\tau \text{sim}(f_{pl}(C_k^I, T_i), f_{pl}(T_k))}}{\sum_{j=1}^L e^{\tau \text{sim}(f_{pl}(C_k^I, T_i), f_{pl}(T_j))}}$$

(8)

where $\tau$ is a temperature hyperparameter and $\text{sim}(\cdot)$ is the cosine similarity between two input vectors. Function $f_{pl}$ denotes the pooling strategy, i.e., mean/max pooling. The pairs $(C_{k,i}^T, I_i)$ and $(C_{k,i}^I, T_i)$ are positive samples of each other. For $L_{T2I}^B$ and $L_{I2T}^B$, these contrastive learners interact within the encoder to emphasize on multimodal pairs while keep text&image which have similar content but opposite labels away from each other in the embedding spaces. In this way, ICAF is able to optimize the contrastive loss directly by supervising the learning of image-text correspondences at each recurrent alignment layer, which is expected to help the model yield higher-quality alignment in each step.

Following the encoder with multiple RA layers, only the text-target alignment features $C_k^T$ is feedforwarded to the decoder, as the Key and Value. Taking the predicted tokens as Query, the standard transformer decoder layer is able to reconstruct the target summary. Compared with the input text, it is shorter but integrates multimodal information better.

D. Loss Function

The main goal of ICAF is to generate the target summary from input sentences and image, based on seq2seq style framework. Therefore, the reconstruction loss is used to learn the optimal model parameters $\theta$, minimizing the negative log-likelihood:

$$L_{\text{gene}}^B = - \sum_{B} \sum_{i} \log p(w_i|w_{1:i-1}, T, I; \theta)$$

(9)

where $w$ denotes the words in the target summary and stops when special token $<eos>$ is generated.

With the auxiliary contrastive losses, the total loss function is defined as Eq. 10 and $|| \cdot ||_2$ denotes the L2 normalization for model parameters $\theta$.

$$L = - \frac{1}{|D|} \sum_{B \in D} (L_{\text{gene}}^B + \beta_1 L_{I2T}^B + \beta_2 L_{T2I}^B) + || \theta ||_2$$

(10)

where $\beta_1$ and $\beta_2$ are optional static parameters, which include three types (see Figure 4): a) Random, b) Increase and c) Decrease. Their values range from 0 to 0.3.

III. EXPERIMENTS

A. Dataset and Metrics Description

To validate the performance of ICAF on MSMO [16], this paper compares both Training-from-scratch and Pre-trained then Fine-tuning baselines, due to the prominent advancement of the pre-train + fine-tune paradigm in both computer vision and natural language processing. This dataset comprises 240,000 image-text pairs for training and 30,000 pairs for both validation and testing from internet news articles with numerous picture-captions. In this paper, the maximum length of utterances is set to 500. Text longer than the maximum length is cut off. We use an open access tool\(^1\) to find the most consistent image since the vision input is only one picture in our proposed ICAF.

The evaluation metrics are calculated between the generated summary and the ground-truth by a public tool\(^2\).

- ROUGE [31], the standard metric to calculate the scores between the generated summary and reference sentences by using the recall and precision of n-gram overlaps.
- Relevance [32], we use embedding-based metrics to evaluate the summary relevance. In particular, Embedding Average, Embedding Extrema and Embedding Greedy are used to measure the effect of the generated summary.

\(^1\)https://github.com/zegami/image-similarity-clustering

\(^2\)https://github.com/rsennrich/Bleualign
The performance of this ICAF model and baselines are compared from three perspectives in this work. We demonstrate that the ICAF model is not only superior to de novo training approaches, but also advances pre-trained then fine-tuning paradigms.

**Comparisons with Training-from-scratch Methods.** These models are trained from scratch by randomly initialize the parameters, which can divided into two categories: 1) single-modal (text) baselines and 2) multimodal (text-image) baselines. Category 1 includes traditional methods such as S2S (RNN model) [34] and S2S+Attn (S2S with attention). PointerNet [35] learns the conditional probability of the output sequence with elements that are discrete tokens corresponding to the positions in the input sequence. Category 2 is directly related to our proposed ICAF. Doubly-Attn [36] uses various attention to fuse modalities and summarize the input ultimately. MSGMR [10] with the guidance of multimodal reference uses the losses from the summary generation and image selection. A location-aware approach for multimodal summarization (LAMS) based on Transformer [37] is used to illustrate that the performance of ICAF is not resulted from the transformer model. Some of our baselines are advised by the open source website – NLPedia³.

³http://explainaboard.nlpedia.ai/

For Training-from-scratch, the comparisons between ICAF and baselines are shown in Table I (bottom subtable). The results show the ICAF model outperforms all the baselines on ROUGE, Embedding-based metrics and Human evaluation with the only exception of Embedding Extrema. The results also show that adding recurrent alignment module and contrastive learning based auxiliary losses can improve the performance of the generative framework. Through the comparison of the top and bottom parts in this subtable, we can see that using multimodal data as input is better than pure text. Multimodal inputs can provide more references.
for summary generation. ICAF outperforming LAMS shows that the proposed framework can fuse multimodal references and generate better, not because of the superiority of the transformer model itself.

**Comparisons with Pre-train & Fine-tune Methods.** For the sake of fairness, this paper updates 5 epochs for the textual pre-trained models GPT-2, UniLM and multimodal pre-trained models VLBERT, UNITER. The former two have the ability of textual generation by designing model structures (GPT-2) or masked matrix (UniLM), while the latter two have the ability of cross-modal understanding by additional tasks such as masked classification (VLBERT) or matching (UNITER). OSCAR [14] is a more recent multimodal pre-training model that employed object tags to match with word tokens and salient regions. The same decoder as ICAF in OSCAR is trained for an extra 25 epochs.

The comparisons between ICAF and baselines of Pre-Trained+Fine-tuning are shown in the top of Table I. It can be illustrated that no matter whether compared with the single-stream or the multi-stream pre-trained models, the ICAF model always performs better in both metrics and human evaluation. This greatly proves the excellence of ICAF. Meanwhile, the baseline of text-only pre-trained model is better in summary fluency, while the multimodal pre-trained model is excellent in correlation.

**D. Ablation Analysis**

In this section, we first evaluate the rationality of the default hyperparameters settings and the chosen types of static factors in loss function, and then verify the effectiveness of the proposed recurrent alignment layer and cross-modal contrastive learning.

**Effect of similarity scale factor**, $\gamma$. In the recurrent alignment layer, the closer the cosine similarity is to 1, the higher the semantic correlation. Setting the default $\gamma$ as a negative value aims to filter more semantical information. As shown in Table II, a larger or smaller $\gamma$ value decreases the performance of ICAF. We can also see that when this threshold increases, the similarity between the selected image and target summary will decrease (the value of $M_{sim}$ becomes smaller). Experiments show that this is an extremely sensitive parameter for different situation or datasets.

**Effect of temperature parameter**, $\tau$. It is a hyperparameter for contrastive loss. Similar to [38], we also find that $\tau = 0.1$ is the most effective value for Mean pooling strategy.

**Effect of layers**, $K$. For ICAF, we gradually increase $K$ from 3 to 9 and evaluate them on the benchmark datasets. Due to space limitation, we only report the results on 3 and 9 in Table II. We can observe that for all variants, $K=9$ can achieve better but limited performance than $K=3$ or $K=6$. The model improvement in exchange for more model parameters is minimal in this analysis.

### Table II

| Methods | ROUGE | Relevance | $M_{sim}$ |
|---------|-------|-----------|-----------|
| ICAF    | 47.60 | 0.433     | 0.428     |
| Default $\gamma = -0.15$ | ICAF ($\gamma = -0.20$) | 47.31 | 0.427 | 0.433 |
|         | ICAF ($\gamma = -0.10$) | 47.51 | 0.430 | 0.391 |
|         | ICAF ($\gamma = -0.05$) | 46.18 | 0.418 | 0.347 |
|         | ICAF ($\gamma = 0$) | 45.72 | 0.414 | 0.302 |
| Default $\tau = 0.1$ | ICAF ($\tau = 0.01$) | 46.18 | 0.411 | - |
|         | ICAF ($\tau = 0.2$) | 40.66 | 0.378 | - |
| Default $K=6$ | ICAF ($K=3$) | 45.57 | 0.409 | - |
|         | ICAF ($K=9$) | 47.64 | 0.432 | - |
| Default $\beta_1 \Rightarrow$ Increase & $\beta_2 \Rightarrow$ Increase | ICAF ($\beta_1 \Rightarrow$ Random) | 46.44 | 0.421 | 0.419 |
|         | ICAF ($\beta_2 \Rightarrow$ Random) | 44.71 | 0.396 | 0.331 |
|         | ICAF ($\beta_1 \Rightarrow$ Random, $\beta_2 \Rightarrow$ Random) | 43.63 | 0.389 | 0.328 |
|         | ICAF ($\beta_1 \Rightarrow$ Decrease) | 46.07 | 0.414 | 0.397 |
|         | ICAF ($\beta_2 \Rightarrow$ Decrease) | 41.22 | 0.374 | 0.316 |
|         | ICAF ($\beta_1 \Rightarrow$ Decrease, $\beta_2 \Rightarrow$ Decrease) | 40.12 | 0.362 | 0.305 |

### Table III

| Methods | ROUGE | Relevance | $M_{sim}$ |
|---------|-------|-----------|-----------|
| ICAF    | 47.60 | 0.433     | 0.428     |
| MSGMR   | 45.64 | 0.423     | 0.264     |
| $-L^{B}_{T2I}$ | 39.27 | 0.352 | 0.302 |
| $-L^{B}_{T2T}$ | 45.48 | 0.411 | 0.387 |
| $-L^{B}_{T2I} \& L^{B}_{T2T}$ | 36.71 | 0.341 | 0.237 |

**Effect of choosen types of $\beta_1$ and $\beta_2$.** Because the cross-modal contrastive loss is accumulated with the iterations of RA layers, i.e., the loss of $k^{th}$ layer will be back propagated to the parameters of 1 to k layers. Therefore, using the $\beta$ of Increase type is conducing to the uniform optimization of model parameters intuitively. The experimental results also show that the performance degrades to a certain extent by replacing $\beta_1$ and $\beta_2$ with Random and Decrease.
Table IV

A Case Study. The references of original utterances and images are given and the corresponding summary generated by both ICAF and baselines are given as well.

| Original Textual Reference | Generated Summary |
|----------------------------|-------------------|
| 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 is 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. | **ICAF**: Li Na has been the most influential women’s tennis player following her 2011 French Open and Australian Open win. It’s not only about the first Asia’s singles grand slam champion its about the influence that she has in tennis. |
