applied sciences

MDPI

Article

Globally Guided Confidence Enhancement Network for Image-Text Matching

Xin Dai, Gulanbaier Tuerhong * and Mairidan Wushouer

Xinjiang Multilingual Information Technology Laboratory, Xinjiang Multilingual Information Technology Research Center, College of Information Science and Engineering, Xinjiang University, Urumqi 830017, China; dx_@stu.xju.edu.cn (X.D.)

* Correspondence: gulam@xju.edu.cn

Abstract: Image-text matching is a crucial aspect of multi-modal intelligence. The main challenge in this area is accurately measuring the relevance between the image and text, using evidence obtained through matching. Previous studies either concentrated on obtaining a well-represented global feature to measure similarity directly or on investigating complex matching patterns at a local level before aggregating them, with little attention paid to combining them. We propose a Globally Guided Confidence Enhancement Network that combines both approaches by obtaining a good global representation to guide fine-grained local interactions. In this process, content that better matches the text from a global perspective is enhanced and represented with confidence scores. Extensive experiments demonstrate that the approach we have employed achieves superior performance on Flickr30K and MSCOCO datasets.

Keywords: image-text matching; global guidance; multimodal interaction

check for updates

Citation: Dai, X.; Tuerhong, G.; Wushouer, M. Globally Guided Confidence Enhancement Network for Image-Text Matching. *Appl. Sci.* **2023**, *13*, 5658. https://doi.org/10.3390/app13095658

Academic Editor: Andrea Prati

Received: 31 March 2023
Revised: 25 April 2023
Accepted: 30 April 2023
Published: 4 May 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Vision and language are essential for humans to comprehend the actual world, and extensive efforts have been made to bridge the two modalities. Image-text matching is one of the field's fundamental topics. Measuring the visual semantic similarity of images and text, and establishing the connection between vision and text in this way facilitates various visual-language tasks, such as Neural Machine Translation [1], Visual Question Answering [2], Text-To-Image Generation [3] and Image Captioning [4]. Despite the tremendous progress in recent times, image-text matching continues to pose significant challenges.

Images and text often describe the same objects or scenes in vastly different ways, resulting in significant semantic differences that must be reconciled for effective image-text matching. These challenges have motivated ongoing research in developing innovative algorithms and techniques to improve the accuracy of this critical task. As shown in Figure 1, existing research can be broadly divided into global alignment-based approaches and local inference-based approaches based on the granularity of semantic alignment. The global alignment-based approaches [5–9] encode the image and text independently and subsequently align them within a shared semantic space. The goal is to obtain a well-represented global feature of the image (text), on which cross-modal similarity is directly measured. Many strategies have been devised to enhance the discriminative ability of joint embeddings, such as semantic concept learning [3,8], region relationship reasoning [10], and multi-view hashing learning [11]. Although the feature representations in the learned common semantic space are better able to characterize high-level semantic concepts of multimodal data, they are insufficient for more specific retrieval requirements.

To capture fine-grained information for matching, attention mechanisms are extensively employed, leading to better matching performance. The local inference-based approaches aggregate the fine-grained local semantic alignment between image regions and words to derive the overall similarity. Motivated by the success of bottom-up attention [4],

authors of ref. [12] proposed a stacked cross-attention network that exploits attention mechanisms to obtain a region-based textual attention representation and a word-based image attention representation, exploring all possible semantic alignments between image regions and textual words. These networks achieved excellent performance and inspired many subsequent studies [13–19]. These investigated intricate semantic interactions between modalities to identify possible visual and textual alignments. Some studies Refs. [20–22] have proposed the use of graph convolutional neural networks or attentional inference to explore fine-grained alignment from multiple perspectives such as objects, relations and attributes. Even though significant progress has been made by considering various fine-grained interaction patterns, not all image-text matching pairs contribute equally to the overall match, and it would be unreasonable to simply aggregate them to obtain the final similarity.

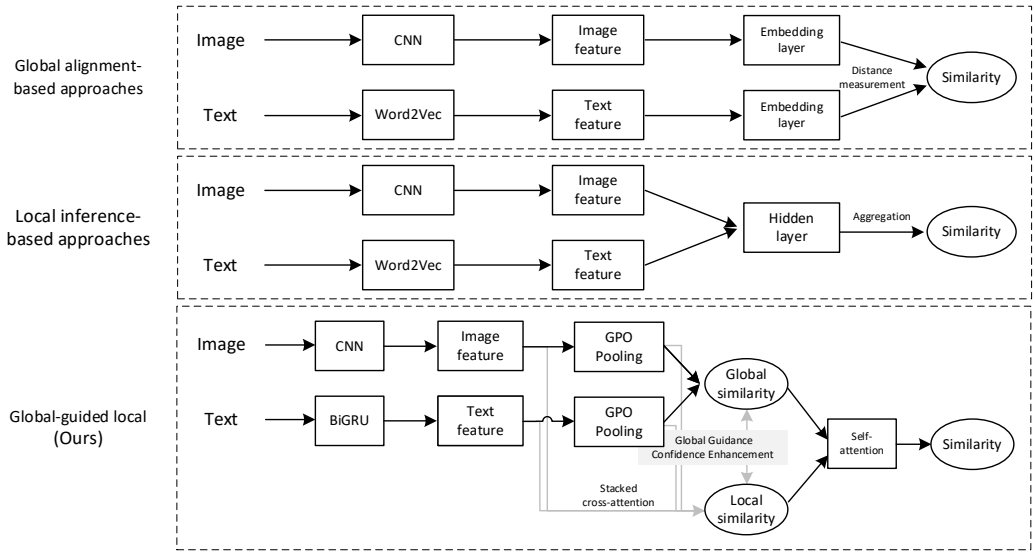


Figure 1. The existing two types of methods and our method that combines them.

Intuitively, well-represented global features better represent high-level semantic concepts of data, but they lack fine-grained information. Local inference-based approaches capture more fine-grained interactions and exhibit better matching performance, but their interaction processes are highly complex and they aggregate all possible image-text matching pairs with equal importance which is not reasonable. At the same time, we argue that the role of a well-represented global feature in local interaction is underestimated. Specifically, it can prevent the model from losing the sense of global semantics during complex fine-grained local interactions, help the model deepen its understanding of the content and guide the model to focus on meaningful local matching pairs.

To address these problems, we propose a globally guided confidence enhancement network (GGCEN), which enhances meaningful alignment in local similarity measures from the perspective of global semantics and can address both the big picture and details in a coordinated and effective way. The holistic characteristic of global features is used to guide fine-grained local interactions, to obtain more accurate matches. Specifically, a text (image) representation fused with an image (text) context and pooled global features work together to generate matching pairs' confidence. In this way, the relationship between image objects and text descriptions can be better captured, reducing the impact of locally similar but globally inconsistent matching relationships on the final similarity measure. Our approach consists of three parts: (1) Feature representation: We derive the representation of image regions and text words, then we use the GPO pooling strategy [23] to obtain the global representations of image and text. Subsequently, we add a self-attention (SA) mechanism to enhance these global representations. (2) Feature alignment: We infer the confidence level of the region-word matching pairs by evaluating how much the semantic similarity

between the text global representation and its corresponding image (text) contributes to the overall semantic similarity. Thus, this feature alignment approach serves to bolster the impact of meaningful region-word matching pairs from a global perspective. (3) Cross-modal similarity measuring: We weigh the local semantic similarity of each region (word) with the inferred confidence, and then use multiple self-attention mechanisms to capture the relationship between the weighted region (word) similarity and the global similarity to generate a more accurate image-text similarity.

The primary contributions of our study can be briefly summarized as follows:

- We propose a Globally Guided Confidence Enhancement Network (GGCEN) that aims to enhance the confidence of meaningful local matching pairs by leveraging well-presented global features to guide fine-grained local interactions.
- We integrate a global feature aggregation module and update those obtained global features with self-attention to improve the global representation of images (text).
- We present a matching confidence enhancement method, which entails inferring the confidence levels of local region-word pairs by evaluating their contributions to the global similarity score. In this evaluation, local pairs that exert a greater influence on the overall similarity score are deemed more meaningful and thus are endowed with higher weights.
- The experimental results show that our model achieves better performance on the public benchmarks Flickr30K and MSCOCO compared to other published mainstream models, while also exhibiting twice the convergence speed.

2. Related Work

In the domain of image-text matching, aligning visual semantics across diverse modalities and measuring cross-modal similarity represent challenging tasks that have received considerable research attention. This is due to the inherent complexity of image-text matching, which involves dealing with multiple modalities and their corresponding semantics. Previous studies Refs. [5–11,23–25] belong to the global alignment-based approach and concentrate primarily on investigating methods of feature fusion or employing the underlying scene semantics to learn more discriminative representations. Specifically, authors of Ref. [5] utilized adversarial learning to construct generative adversarial tasks to learn a common representation. Extracting a representation from the given image or text, the model used the generative networks to obtain a more specific cross-modal target representation. However, the quality of the synthesized images was dependent on the quality of the pre-trained image generator model, which may introduce bias or errors. The cross-modal gap between vision and language was narrowed in Ref. [7] by creating a joint embedding space using Recurrent Residual Fusion (RRF) blocks. This approach results in a more discriminative embedding space that can accommodate both visual and textual representations. Aggregated image scene graphs were leveraged in Ref. [8] to extract frequently co-occurring concepts as scene commonsense knowledge and develop a Scene Concept Graph (SCG). Specifically, SCG detects semantic concepts from image regions and utilizes triplets of scene graphs collected from the Visual Genome dataset as commonsense information, and uses a gating mechanism to selectively expand the associated information of detected concepts. A simple and interpretable reasoning model was proposed in [10] to generate visual representations that capture key objects and semantic concepts in a scene. The approach first establishes connections between image regions and uses a graph convolutional network for reasoning, to generate region features with semantic relationships. Then, using gate and memory mechanisms, the feature enhanced with these relationships undergoes global semantic reasoning to select discriminative information and iteratively generate the representation for the entire scene. The Generalized Pooling Operator (GPO) is a technique introduced in Ref. [23] that automatically adapt the best pooling strategy for various data and feature backbones. The proposed GPO offered a new way to explore the optimal pooling strategy automatically and efficiently, saving the human effort of manual tuning and speeding up the research progress in image-text matching and beyond. A novel loss function has also

been introduced in Ref. [25] that incorporates hard negative samples, pairs that are visually similar but semantically dissimilar, to improve the quality of visual-semantic embeddings. Image processing and analysis was involved in Refs. [26,27] with a focus on identifying features or patterns through various techniques, to improve image quality, text analysis and data categorization. The method, based on global alignment, aimed to enable images and text to learn information from each other's modalities, leading to cross-modal representations with advanced semantics through the design of better distance loss functions, adversarial learning or other task objectives. However, this method overlooked the internal connections and detailed interactions between different modalities.

In an effort to establish a more precise relationship between vision and language, previous research [2,6] has attempted to learn correspondences between regions and words at a fine-grained level. The Dual Attentive Network (DAN) has been proposed by Ref. [2], in which both visual and textual attention mechanisms are utilized to capture subtle interactions between image regions and words. Through multiple stages, the DAN focuses on specific regions within the image and specific words within the text, collecting necessary information from both modalities and estimating their similarity by attending to shared semantics between the image and text. However, due to the high degree of coupling involved in alignment aggregation, these approaches have only been able to achieve limited alignment between the two modalities. Notably, authors of ref. [12] proposed a stacked cross-attention mechanism that leverages image regions and words as contexts to explore fine-grained semantic alignment, by aggregation of resulting local alignments through average pooling or Log-Sum-Exp pooling. This has achieved impressive performance, providing valuable insights into the research of image-text matching. Inspired by this work [12], many researchers refs. [13–15,17–22,28] have sought to design plausible cross-modal alignment mechanisms based on the local inference framework for satisfying visual semantic interactions to facilitate similarity measurement. One of these works, Ref. [13], proposed an Iterative Matching with Recurrent Attention Memory (IMRAM) method, which captures the correspondence between images and text through multi-step alignment. During the alignment process, memory extraction units extract alignment knowledge from early steps to later interactions. The approach aims to achieve more precise and accurate region-word correspondence, which greatly improves the local similarity and significantly reduces the distance between images and texts for positive image-text pairs. A Dual-Path Recurrent Neural Network (DP-RNN) was introduced by the authors of Ref. [14] that symmetrically processes images and text using Recurrent Neural Networks (RNNs). Specifically, given an input image-text pair, the model's image region input sequence is re-ordered according to the position of the most relevant word in the text. The RNN is then used to extract advanced features from the re-ordered region inputs, including joint information of semantically related regions. The model then utilizes a multi-attention cross-matching mechanism to better match the correspondence between image regions and text words. In Ref. [15] a Relation-wise Dual Attention Network (RDAN) was proposed to encode both local correlations and global dependencies between regions and words. This approach not only allows for the exploration of fine-grained local similarities but also enables the exploration of potential visual-semantic relationships, providing rich supplemental information for inferring visual-semantic alignment and measuring image-text similarity. Recently, a learnable vector-based similarity representation was proposed Ref. [20] to more effectively model cross-modal correlations, instead of using scalar cosine similarity to represent the alignment information, which is then used to enhance the interaction between global and local alignment via the utilization of graph convolutional networks. Based on graph convolutional networks [29], the authors of ref. [21] proposed a method of constructing multiple similarities and strengthening the alignment of regions and words. This construction method first divides the region features and word features into multiple blocks, and calculates the cosine similarity between each pair of blocks, making different blocks play different roles in graph transmission. However, the cosine similarity obtained by this block-wise calculation is not learnable, and the similarity representation

depends on the block encoding of region features and word features. The local inference-based approaches have made further explorations in optimizing single modality-enhanced interaction, optimizing multimodal feature fusion or directly improving attention units. However, authors of these studies try to find the nearest region (word) features to the current word (region), which tends to capture the so-called "associativity". Obviously this will inevitably increase the similarity of negative samples to the locality, making it difficult for positive and negative image-text pairs to produce large distances. Furthermore, these studies resulted in designs with very complex structures and introduced large amounts of additional information, thus, greatly increasing the complexity of time and space.

In conclusion, most existing studies either concentrated on obtaining a well-represented global feature to measure similarity directly or on investigating complex matching patterns at a local level by simply aggregating them, with little attention paid to combining them. These two methods can mutually reinforce each other, leading to better results when they are used together. A well-represented global feature prevents the model from losing the sense of global semantics during complex fine-grained local interactions, helps the model deepen its understanding of the content and guides the model to focus on meaningful local matching pairs.

3. Methodology

Details on how to implement globally guided matching confidence enhancement are provided in this section. We focus on using better global representations to guide local similarity inference, specifically by enhancing the confidence of meaningful region-word matching pairs under a global perspective, to improve visual semantic similarity learning. As shown in Figure 2, we first describe how to encode images and text and the method to obtain their global representation. Then, we elaborate on how to improve the confidence of locally meaningful matching pairs from a global image-text perspective and elucidate their application in the ultimate inference of similarity. Finally, we present the objective function for training.

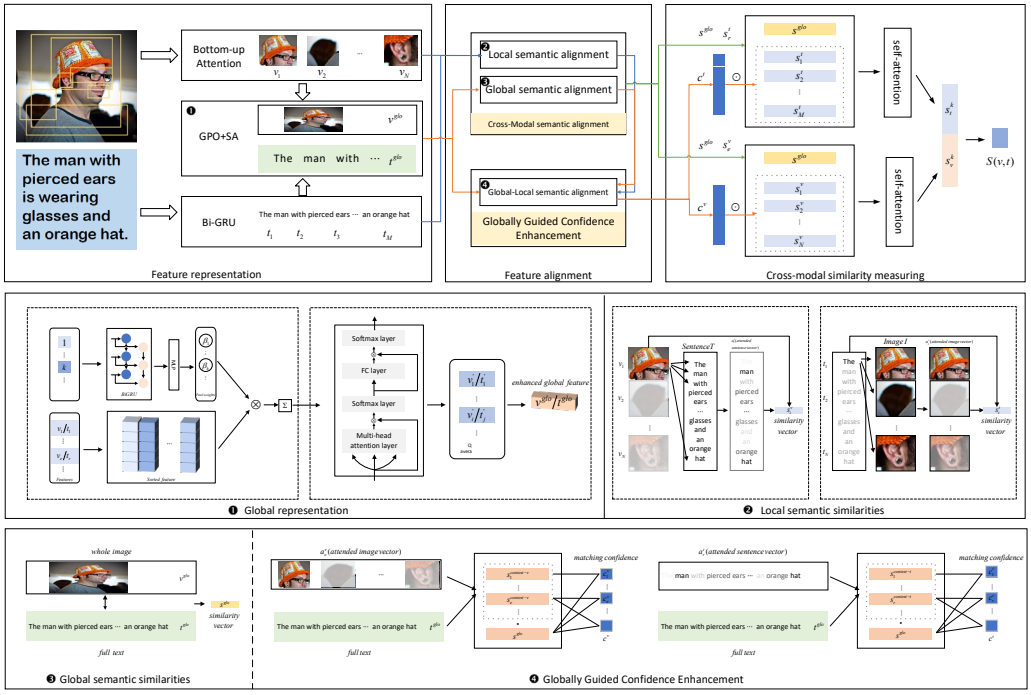


Figure 2. The figure depicts our proposed GGCEN. Illustration inspired by [22].

Our model's algorithm is shown in Algorithm 1. The core idea is to use better global representations to guide local fine-grained interaction operations, which is reflected in

the improvement of confidence levels that leads the model to focus on meaningful local matching pairs.

Algorithm 1 GGCEN for Image-Text Matching

Batch Input: f_{img}-original image region feature, f_{txt}-original word feature.
Batch Output: *sims*-image-text similarity.
/* get the updated feature representation of image regions and caption words
$f_{img_new} = \text{l2norm}(\text{FC}(f_{img}) + \text{MLP}(f_{img}))$
$f_{txt_new} = \text{l2norm}(\text{BiGRU}(f_{txt}))$
/* get global representation of image regions and caption words by GPO operation
$f_{img}^{glo} = \text{l2norm}(\text{GPO}(f_{img_new}))$ /* Equation (1) */
$f_{txt}^{glo} = \text{l2norm}(\text{GPO}(f_{txt_new}))$ /* Equation (1) */
/* update the global representation by GPO guided self-attention mechanism
$f_{img}^{glo} = \text{Self_attention}(f_{img}^{glo}, f_{img_new})$
$f_{txt}^{glo} = \text{Self_attention}(f_{txt}^{glo}, f_{txt_new})$
/* calculate the image-text similarity
$N_{img} = f_{img}.\text{size}(0), N_{txt} = f_{txt}.\text{size}(0)$ /* get the number of images and captions */
$sims = \text{torch.zeros}(N_{img}, N_{txt})$ /* initialize image − text similarity matrix */
for i to $f_{txt}.\text{size}(0)$ **do**/* iterate through each caption */
    $f_{txt_i} = f_{txt_new}[i]$ /* get words feature of ith caption */
    $f_{txt_i_\exp and} = f_{txt_i}.\text{repeat}(N_{img}, 1, 1)$ /* repeat N_{img} words feature */
    /* get the image-text interaction feature by SCAN attention
    $a_e^t = \text{SCAN}(f_{txt_i_\exp and}, f_{img_new})$ /* caption words attention feature, Equatqions (4) and (5) */
    $a_r^v = \text{SCAN}(f_{img_new}, f_{txt_i_\exp and})$ /* image regions attention feature, Equations(4) and (5) */
    /* get the local semantic alignment
    $f_{txt_i_\exp and}^{glo} = f_{txt}^{glo}[i].\text{repeat}(N_{img}, 1)$ /* repeat N_{img} caption global feature */
    $s_e^v = \text{l2norm}(\text{FC}(a_r^v - f_{img_new})^2)$ /* Equation (3) */
    $s_r^t = \text{l2norm}(\text{FC}(a_e^t - f_{txt_new})^2)$
    /* get the global semantic alignment
    $s^{glo} = \text{l2norm}(\text{FC}(f_{img}^{glo} - f_{txt_i_\exp and}^{glo})^2)$ /* Equation (6) */
    /* get the local confidence
    $s_r^{context-v} = \text{l2norm}(\text{FC}(a_r^v - f_{txt_i_\exp and}^{glo})^2)$ /* Equations (3) and (7) */
    $s_e^{context-t} = \text{l2norm}(\text{FC}(a_e^t - f_{txt_i_\exp and}^{glo})^2)$ /* Equation (8) */
    $c_e^v = \text{sigmoid}(\text{FC}(s_e^{context-t} \odot s^{glo}))$ /* Equation (10) */
    $c_r^t = \text{sigmoid}(\text{FC}(s_r^{context-v} \odot s^{glo}))$ /* Equation (11) */
    /* Cross-modal Similarity Measuring
    /* update local semantic alignment using the confidence
    $s_e^{v\prime} = c_e^v \odot s_e^v$ /* Equation (11) */
    $s_r^{t\prime} = c_r^t \odot s_r^t$ /* Equation (12) */
    /* catenate the global and local semantic alignment
    $S_v = [s^{glo}, s_e^{v\prime}],\ S_t = [s^{glo}, s_r^{t\prime}]$ /* Equation (13) */
    /* similarity graph reasoning by self_attention
    $S_v = \text{self_attention}(S_v), S_t = \text{self_attention}(S_t)$
    /* calculate the final score by catenating the image and caption similarity
    $S = \text{sigmoid}(\text{FC}([S_v, S_t]))$ /* Equation (14) */
    /* save the similarity between all images and ith caption
    $\text{sims}[:, i] = S$
end for

3.1. Feature Representation

Visual Representations. For each input image, M region-level visual features are extracted according to [4] using the Faster R-CNN [30] and ResNet-101 [31] models pre-trained on the Visual Genomes dataset [32]. In particular, the Faster R-CNN model is used to detect salient regions in image I and encode a visual representation for each detected

image region. We convert them to d-dimensional vectors by adding a fully-connected layer. The image I can be denoted as $\left\{ v_e \middle| e = 1, 2, \ldots, M, v_e \in \mathbb{R}^d \right\}$, where M is the number of regions in image I.

Textual Representations. Given a text, we tokenize it into L words and input the word embeddings sequentially into a bidirectional GRU(Bi-GRU) [33] to improve the representation of the text using context semantics. After this, the representation of each word is produced by taking an average of the forward and backward hidden state at each time step, as $\left\{ t_r \middle| r = 1, 2, \ldots, L, t_r \in \mathbb{R}^d \right\}$.

Global Representation and Self-Attention Enhancement. Globally guided local alignment plays an important role in constructing effective contextual features that improve the retrieval performance of the model while reducing the training complexity. Our idea is to mine local alignment information using a globally guided model. Inspired by [23], we use GPO, a generalized pooling operator that automatically adapts to different data modalities and feature extractors, to find well-encoded global features, then we use self-attention to enhance them.

Specifically, the pooling coefficients $\theta = \{\theta_i\}_{i=1}^N$ are generated by GPO, and the global features f^{glo} are defined as a weighted sum of locally sorted features. Note that v^{glo} and t^{glo} are denoted as f^{glo} in Equation (1):

$$f^{glo} = \sum_{i=1}^{N} \theta_i \cdot \max_i(\{\varphi_n^i\}_{n=1}^K), \forall_i \tag{1}$$

The size of the coefficients θ is N. The ith maximum value among the N elements is denoted by θ_i. We enforce the constraint $\sum_{i=1}^{N} \theta_i = 1$ on the resulting weighted vector using Softmax. The positional encoding is used to enable the model to capture the sequential information of the input sequence. To achieve this, the positional encoding strategy in Transformers [34] is adopted, which vectorizes the positional index i and converts it to a dense vector p_i, $p_i^x = \begin{cases} \sin(w_y, i), & when \quad x = 2y \\ \cos(w_y, i), & when \quad x = 2y+1 \end{cases}, \forall i$, where $w_y = \frac{1}{10{,}000^{2y/d_3}}$ and the positional encoding has d_3 dimensions. Then, a sequence model is learned to generate the pooling weights, which consists of a bi-directional GRU (Bi-GRU) [33] and a multi-layer perceptron (MLP): $\{h_i\}_{i=1}^N = \text{BIGRU}(\{p_i\}_{i=1}^N)$, $\theta_i = \text{MLP}(h_i)$ that takes the location-encoded sequence $p = \{p_i\}_{i=1}^N$ as input and outputs the pooling coefficients $\theta = \{\theta_i\}_{i=1}^N$. At the same time, each dimension of the vector is sorted from the largest to smallest as φ and then multiplied by the corresponding coefficients for weighting to obtain the global features f^{glo}. In addition, we use a self-attention mechanism to update the global features to enhance the high-level semantic representation of the global features.

3.2. Feature Alignment

To capture a more fine-grained correspondence, inspired by [20], we use similarity vectors to capture fine-grained interactions between feature representations from different modalities.

The similarity function between vectors $v \in R^d$ and $u \in R^d$ is defined as:

$$s(v, u; W) = \frac{W|v - u|^2}{||W|v - u|^2||_2} \tag{2}$$

where $|\cdot|^2$ and $||\cdot||_2$ indicate element-wise square and L2-norm respectively, and $W \in \mathbb{R}^{m \times d}$ is a learnable parameter matrix to obtain the m-dimensional similarity vector.

Local Semantic Alignment. We compute the local semantic similarity representation s_e^v between the image region v_e and its semantically matched relevant words a_e^t with Equation (3):

$$s_e^v = s(v_e, a_e^t; W_M^v) \tag{3}$$

where $W_M^v \in \mathbb{R}^{m \times d}$ aims to learn the local similarity representation.

Following [12], we define a weighted combination of word representations to attend to words with respect to each image region. Specifically, this combination is referred to as the attended text vector a_e^t, and is defined with respect to the image region v_e:

$$a_e^t = \sum_{e=1}^{M} \alpha_{er} t_r, \tag{4}$$

where

$$\alpha_{er} = \frac{\exp^{(\lambda \hat{c}_{er})}}{\sum_{e=1}^{M} \exp^{(\lambda \hat{c}_{er})}}, \tag{5}$$

where $\hat{c}_{er} = \frac{[c_{er}]_+}{\sqrt{\sum_{r=1}^{L} [c_{er}]_+^2}}$ and c_{er} is the cosine similarity between region v_e and word t_r. Correspondingly, the establishment of semantic similarity s_r^t between a word t_r and its visually-matched context a_r^v within the image is facilitated by $s_r^t = s(t_r, a_r^v; W_L^t)$.

Global Semantic Alignment. We compute the global semantic similarity representation s^{glo} between the image global representation v^{glo} and the text global representation t^{glo} with Equation (6):

$$s^{glo} = s(v^{glo}, t^{glo}; W_G) \tag{6}$$

where $W_G \in \mathbb{R}^{m \times d}$ aims to learn the global similarity representation.

Globally Guided Confidence Enhancement. Salient image regions' visual semantics become fragmented when they are viewed separately, which results in regionally matched region-word pairs that might not be compatible with the overall image-text semantics. We use confidence to weigh how well each region (word) ascribes to the global image-text view, thereby boosting the confidence of meaningful local matching pairs. Specifically, we verify the semantic similarity between text global representation t^{glo} and its semantically matched relevant images (text) a_r^v (a_e^t) as:

$$s_r^{context-v} = s(a_r^v, t^{glo}; W_c^v) \tag{7}$$

and

$$s_e^{context-t} = s(a_e^t, t^{glo}; W_c^t) \tag{8}$$

where $W_c^v \in \mathbb{R}^{m \times d}$ and $W_c^t \in \mathbb{R}^{m \times d}$ are learnable parameter matrices.

Then, we use the global semantic similarity s^{glo}. The normalized correlation between $s_r^{context-v}$ and $s_e^{context-t}$ for confidence perception with Equation (8) is defined as:

$$c_e^v = \sigma(LN(w_e^v(s^{glo} \odot s_e^{context-t}))) \tag{9}$$

and

$$c_r^t = \sigma(LN(w_r^t(s^{glo} \odot s_r^{context-v}))) \tag{10}$$

respectively, where $\left\{ c_e^v \middle| e = 1, 2, \ldots, M, c_e^v \in \mathbb{R}^d \right\}$ and $\left\{ c_r^t \middle| r = 1, 2, \ldots, L, c_r^t \in \mathbb{R}^d \right\}$. $w_r^t \in \mathbb{R}^{1 \times d}$ and $w_r^t \in \mathbb{R}^{1 \times d}$ are learnable parameter vectors, and $\odot$ indicates the element-wise product, σ represents the sigmoid function and LayerNorm represents the normalization of layers. Specifically, a text (image) representation fused with an image (text) context and global pooled features work together to generate matching pairs' confidence. The idea is to use a better global representation to guide local fine-grained interactions and guide the model to focus on meaningful local matching pairs.

3.3. Cross-Modal Similarity Measuring

To better capture the relationship between image objects and text descriptions, we reduced the impact of locally similar, but globally inconsistent matching relationships on the final similarity measure. In the initial stage, a multiplication operation is performed whereby each semantic similarity s_e^v that corresponds to a region request is multiplied by

its respective confidence weight c_e^v. This process obtains reinforced local similarities $s_e^{v\prime}$, as exemplified by the formulations:

$$s_e^{v\prime} = [c_1^v s_1^v, c_2^v s_2^v, \ldots, c_M^v s_M^v], e = 1, 2, \ldots, M \tag{11}$$

and

$$s_r^{t\prime} = [c_1^t s_1^t, c_2^t s_2^t, \ldots, c_L^t s_L^t], r = 1, 2, \ldots, L \tag{12}$$

Following this, we proceed to gather the overall semantic similarity s^{glo} and the enhanced semantic similarity $s_e^{v\prime}$ obtained through region-specific queries, yielding:

$$S_v = [s^{glo}, s_1^{v\prime}, s_2^{v\prime}, \ldots, s_M^{v\prime}] \tag{13}$$

In the meantime, we gather the global similarity s^{glo} and the enhanced word-specific semantic similarity $s_r^{t\prime}$ as: $S_t = [s^{glo}, s_1^{t\prime}, s_2^{t\prime}, \ldots, s_L^{t\prime}]$.

With the aim of obtaining more refined global alignments that are specific to each modality, multi-layer self-attention reasoning is performed separately on S_v and S_t, as:

$$S^{k+1} = \mathrm{RELU}(\mathrm{W}_s^k \cdot \mathrm{softmax}(\mathrm{W}_a^k S^k \cdot (\mathrm{W}_b^k S^k)^\top) \cdot S^k) \tag{14}$$

where $W_a^k \in \mathbb{R}^{m \times m}$ and $W_b^l \in \mathbb{R}^{P \times P}$ are parameter matrices that transform the attention query and key in the kth layer, while $W_s^k \in \mathbb{R}^{m \times m}$ is a parameter matrix that maps the attended features to the next $(k+1)$th layer. Note that S_v^k and S_t^k are represented as S^k in Equation (14).

To further quantify the cross-modal similarity between image and text, we combine the vision-enhanced global semantic similarity S_v^k with the language-enhanced global semantic similarity S_t^k in the final kth layer. This combined output is subsequently fed into a fully connected layer that is activated by the sigmoid function with Equation (15):

$$S(v, t) = \sigma(w_r([S_v^k : S_t^k])) \tag{15}$$

where $w_r \in \mathbb{R}^{1 \times 2d}$ represents the learnable parameters for mapping the concatenated similarity vector to a scalar similarity score.

3.4. Objective Function

Matching often uses ranking goals to clump matched image-text pairs and push unpaired ones apart in the common embedding space. In accordance with [25], bidirectional triplet loss is used for end-to-end training. For the sake of computational effectiveness, the hard negatives are contained within a minibatch:

$$\begin{aligned} \mathcal{L}(v, t) \quad &= [\alpha - S(v, t) + S(v, t^-)]_+ \\ &+ [\alpha - S(v, t) + S(v^-, t)]_+ \end{aligned} \tag{16}$$

where λ serves as the margin constraint, $[z]_+ = \max(z, 0)$. The cross-modal semantic similarity measurement, denoted by $S(\cdot)$, is defined by Equation (2). Specifically, for a positive pair (v, t), the hardest negatives in the training minibatch are $v^- = \mathrm{argmax}_{v^- \neq v} S(v^-, t)$ and $t^- = \mathrm{argmax}_{t^- \neq t} S(v, t^-)$.

4. Experiments

4.1. Datasets and Evaluation Metrics

Flickr30K [35] and MSCOCO [36] datasets were utilized to evaluate our method. Each of the 31,000 images in Flickr30K is accompanied by five captions. We utilized 29,000 images for training, 1000 images for validation, and 1000 images for testing, in accordance with the dataset splits in [12]. The MSCOCO dataset is split into three separate subsets, namely 123,287 training images, 5000 validation images, and 5000 testing images. We average over 5 folds of 1000 test images and tested on all 5000 test images. In adherence with established

information retrieval protocols, we conducted an evaluation of performance utilizing the R@K (Recall at K), which quantifies the proportion of queries that were accurately matched within the top K retrieved instances. Specifically, we assessed performance using three distinct R@K metrics, namely R@1, R@5, and R@10. A higher R@K value indicates better performance. We also calculated the overall matching performance by summing up all recall values as R@sum in both the image-to-text and text-to-image orientations.

4.2. Implementation Details

The detector from [4], which is the Faster R-CNN [30] with ResNet-101 [31], subsequently produces a feature vector with a dimension of 2048 for each proposed region. The dimensionality of the word embeddings was specified as 300, while the shared embedding space has a dimension of 1024, and the distance-based similarity vectors have a dimension of 256. The model employs a relevance-measuring self-attention mechanism with 3 layers and an initial learning rate of 0.0002 for optimization with the Adam optimizer [37]. In order to optimize the performance of the model, a training protocol was employed involving 40 epochs of training with the Flickr30K dataset and an additional 20 epochs of training with the MSCOCO dataset. Upon completion of this training regimen, the learning rate was attenuated by a factor of 10 to ensure stable convergence. Additionally, a margin parameter of 0.2 was empirically established within the triplet loss function to further enhance the discriminative power of the learned representations. It is worth noting that these parameter values were selected based on thorough experimentation and analysis to maximize the accuracy and robustness of the model. During testing, the finest validation snapshot is utilized. The source code will be published soon.

4.3. Comparisons with State-of-the-Art Methods

To evaluate the efficacy of our proposed GGCEN, we conducted a comprehensive comparative evaluation against the latest state-of-the-art approaches on the widely-used Flickr30K and MSCOCO datasets. To ensure a fair comparison, we employed the same feature extraction backbone for all the methods and thoroughly benchmarked our proposed method against the others. Specifically, Faster R-CNN was used for images and Bi-GRU was used for text to ensure a fair comparison. Through rigorous experimentation and analysis, we were able to demonstrate the superior performance of GGCEN over the existing state-of-the-art methods, thereby highlighting its potential as a powerful tool for solving challenging image-text matching tasks. Tables 1–3 present the comparison results for Flickr30K, MSCOCO 1K, and 5K, respectively. The experimental findings indicate that the proposed Globally Guided Confidence Enhancement Network (GGCEN) performs remarkably well on both the Flickr30K and MSCOCO datasets.

Table 1 shows the experimental results of the GGCEN model compared with existing advanced methods on the Flickr30k dataset, where the model outperforms the current baseline and achieves the best performance. Compared with the VSRN model [10], which uses a typical global reasoning mechanism for image-text matching, the GGCEN model emphasizes the inference of fine-grained region-word relationships guided by well-represented global features. Meaningful matching pairs in a global view are enhanced in confidence and play a greater role in the image-text matching process, resulting in more accurate similarity. The results demonstrate that compared to the current state-of-the-art method CMCAN, the proposed GGCEN model improves the R@1 results for text retrieval and image retrieval by 0.8% and 1.3%, respectively, due to the introduction of the GPO [23] pooling method for better global feature representation. We do not rely on complex prior knowledge, but instead, use a way to enhance the confidence of meaningful matching pairs in a global view to achieving better results. The SGRAF model [20] also calculates the correspondence between global and local features and performs graph reasoning and attention filtering on similarity. Compared with the SGRAF model, the proposed GGCEN model achieves a 2.6% and 2.4% improvement in R@1 results for text and image retrieval, respectively. The GGCEN model enhances the confidence of meaningful region-word matching pairs during

relationship reasoning, more accurately eliminating redundant information, and further enhancing matching performance. Furthermore, the model not only uses the relationship between fine-grained local alignment features but also introduces a global feature representation with advanced semantic information to focus on more abundant potential semantic information, and more effectively uses local and global features to mutually guide accurate matching. The above experimental results show that the proposed method fully and efficiently utilizes cross-modal interaction information and the relationships between them, more accurately aligning significant fine-grained features with global features, further reducing the distance between query samples and correct results, and achieving relatively significant improvements in R@1. More potential semantic information that can be aligned is explored in local reasoning, achieving good results in R@5 and R@10.

Table 1. Comparative analysis of Flickr30K. The best performances are in bold. The symbol (↑) used in the table signifies the degree of enhancement in model performance.

Method	Text Retrieval			Image Retrieval			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++$_{2017}$	52.9	-	87.2	39.6	-	79.5	-
SGM$_{2020}$	71.8	91.7	95.5	53.5	79.6	86.5	478.6
CAMP$_{2019}$	68.1	89.7	95.2	51.5	77.1	85.3	466.9
SCAN$_{2018}$	67.4	90.3	95.8	48.6	77.7	85.2	465.0
MMCA$_{2020}$	74.2	92.8	96.4	54.8	81.4	87.8	487.4
IMRAM$_{2020}$	74.1	93.0	96.6	53.9	79.4	87.2	484.2
VSRN$_{2019}$	71.3	90.6	96.0	54.7	81.8	88.2	482.6
GSMN$_{2020}$	76.4	94.3	97.3	57.4	82.3	89.0	496.8
SGRAF$_{2021}$	77.8	94.1	97.4	58.5	83.0	88.8	499.6
CMCAN$_{2022}$	79.6 *	94.0 *	97.4 *	59.6 *	84.1 *	89.2 *	503.9 *
GGCEN	80.4 (0.8↑)	96.0 (1.7↑)	97.9 (0.5↑)	60.9 (1.3↑)	84.6 (0.5↑)	90.1 (0.9↑)	509.9 (6.0↑)

* The values in this table were obtained from our own implementation.

Table 2. Comparative analysis of MSCOCO 1K test images. The symbol (↑) used in the table signifies the degree of enhancement in model performance.

Method	Text Retrieval			Image Retrieval			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++$_{2017}$	64.6	90.0	95.7	52.0	84.3	92.0	478.6
SGM$_{2020}$	73.4	93.8	97.8	57.5	87.3	94.3	504.1
CAMP$_{2019}$	72.3	94.8	98.3	58.5	87.9	95.0	506.8
SCAN$_{2018}$	72.7	94.8	98.4	58.8	88.4	94.8	507.9
MMCA$_{2020}$	74.8	95.6	97.7	61.6	89.8	95.2	514.7
IMRAM$_{2020}$	76.7	95.6	98.5	61.7	89.1	95.0	516.6
VSRN$_{2019}$	76.2	94.8	98.2	62.8	89.7	95.1	516.8
GSMN$_{2020}$	78.4	96.4	98.6	63.3	90.1	95.7	522.5
SGRAF$_{2021}$	79.6	96.2	98.5	63.2	90.7	96.1	524.3
CMCAN$_{2022}$	78.0 *	96.0 *	98.6 *	63.0 *	90.1 *	95.6 *	521.3 *
GGCEN	80.5 (0.9↑)	96.5 (0.1↑)	98.8 (0.2↑)	65.5 (2.2↑)	91.2 (0.5↑)	96.3 (0.2↑)	528.8 (4.5↑)

* The values in this table were obtained from our own implementation.

As shown in Table 2, compared to the current best methods, the GGCEN model further improves text and image retrieval performance on the MSCOCO 1K dataset. Compared to the SGRAF model [20], our method improves R@1 results by 0.9% and 2.3% in text and image retrieval, respectively. The SGRAF model takes spatial information into account for text features and significantly improves the effectiveness of text retrieval, especially in R@1. However, the GGCEN model focuses on obtaining crucial alignment information

between the two modalities. Compared with the IMRAM [13] model with an attention network, the GGCEN model introduces a confidence filtering mechanism during local alignment, effectively enhancing meaningful matching information during information exchange between modalities, and ensuring that the focus during global reasoning is on potential semantic relationships between significant aligned features. Additionally, the GGCEN model weighs the local semantic similarity of each region (word) with the inferred confidence and then uses multiple self-attention mechanisms to capture the relationship between the weighted region (word) similarity and the global similarity to generate a more accurate image-text similarity.

Table 3. Comparative analysis of MSCOCO 5K test images. The symbol (↑) used in the table signifies the degree of enhancement in model performance.

Method	Text Retrieval			Image Retrieval			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++$_{2017}$	41.3	-	81.2	30.3	-	72.4	-
SCAN$_{2018}$	50.4	82.2	90.0	38.6	69.3	80.4	410.9
IMRAM$_{2020}$	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SGRAF$_{2021}$	57.8	-	91.6	41.9	-	81.3	-
CMCAN$_{2022}$	57.8 *	83.9 *	91.4 *	41.4 *	70.9 *	81.2 *	426.6 *
GGCEN	60.0 (2.2↑)	86.1 (2.2↑)	92.5 (0.9↑)	43.8 (1.9↑)	72.8 (1.9↑)	83.1 (1.9↑)	438.3 (11.7↑)

* The values in this table were obtained from our own implementation.

According to Table 3, the MS COCO 5K dataset test results demonstrate that GGCEN improves the R@1 results of text and image retrieval by 2.2% and 2.4%, respectively, compared with the CMCAN model [22], which considers local fine-grained alignment and explores the relatedness of local features with text as a bridge. However, the proposed GGCEN model fully utilizes the interdependence between all local features and effectively guides local feature matching from a global feature perspective. The model introduces global information as guidance for understanding the potential semantic relationship between image and text, which is reflected in the increased confidence of meaningful matching pairs, displaying its effectiveness and robustness.

4.4. Ablation Study

To demonstrate the efficacy of integrating matching confidence enhancement with global guidance in cross-modal similarity measures, we conducted an assessment of performance on Flickr30K without such enhancements, as outlined in Table 4. In both image and text retrieval, the cross-modal similarity measure that incorporates confidence enhancement and global feature (plus self-attention) outperforms the approach that lacks these enhancements across all metrics. Notably, the inclusion of global guidance leads to improvements in R@1, R@5, and R@10 by 4.0%, 2.9%, and 1.3%, respectively, for text retrieval, and by 4.9%, 3.0%, and 3.4%, respectively, for image retrieval. This indicates that guiding region-word pair matching from a global perspective based on local fine-grained alignment can effectively improve matching performance. This is because the well-represented global features prevent the model from losing global semantics in complex fine-grained local interactions, helping the model deepen its understanding of content and guiding the model to focus on meaningful local matching pairs. From Table 4, it can be seen that after including confidence enhancement, the R@1, R@5, and R@10 text retrieval results show improvements of 4.3%, 2.2%, and 0.8% respectively, while the image retrieval results indicate improvements of 4.0%, 1.3%, and 1.3%. These findings suggest that the confidence scores, which are inferred from the degree of semantic similarity between globally represented text and related images (text) that match semantically, effectively represent the contribution of meaningful region-word matching pairs to the overall similarity in the global context. One possible reason for this may be the incorporation of global information in the region-word relationship reasoning process.

Based on these confidence-enhanced features, more accurate matching can be achieved in the process of cross-modal global similarity reasoning. Additionally, the inference is accomplished by averaging the similarity scores of two trained models, which outperform the similarity measurement without confidence enhancement and global feature (plus self-attention) by 19.6% and 14.4% on the overall performance metric R@sum, respectively.

Table 4. Ablation on Flickr30K. Inference is accomplished by averaging the similarity scores of two trained models.

Method	Text Retrieval			Image Retrieval			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
without global feature (plus SA)	76.4	93.1	96.6	56.0	81.6	86.7	490.3
without confidence enhancement	76.1	93.8	97.1	56.9	82.4	88.8	495.1
GGCEN	80.4	96.0	97.9	60.9	84.6	90.1	509.9

4.5. Qualitative Analysis

In order to evaluate the effectiveness of GGCEN, we have presented a graphical representation of the outcomes of confidence enhancement under global guidance in Figure 3, where the red boxes represent not only the key scenes in the image-text matching but also the image regions that are highly correlated with the text from a global perspective, and the blue boxes represent the weakly correlated image regions.

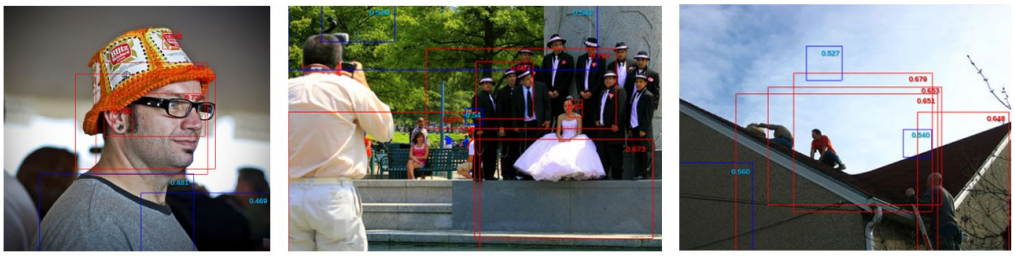


Figure 3. Visualization the result of confidence enhancement under global guidance. The red boxes represent not only represent the key scenes in the image-text matching but also denote the image regions that are highly correlated with the text from a global perspective, and the blue boxes represent the weakly correlated image regions.

Qualitative Results of the Image-to-Text Retrieval. The qualitative results of text retrieval according to image queries on Flickr30K are depicted in Figure 4. For each image query, the top-5 retrieved captions are presented. The ranking is determined by GGCEN's predicted similarity scores. According to these results, our GGCEN is able to derive the correct results from the top-ranked text. The model outputs some understandable inconsistencies, such as those in Figure 4a (5). The visualization indicates that GGCEN can deeply understand the content of images and text, focusing on meaningful region-word matching pairs.

Qualitative Results of the Text-to-Image Retrieval. The qualitative results of image retrieval according to text queries on Flickr30K are depicted in Figure 5. Specifically, for each text description, the top-3 retrieved images are presented. Accurate matches are highlighted in green, while inaccurate matches are denoted in red. We observe that our model successfully retrieves the ground-truth image within the top-3 results. It is worth noting that other results are also understandable, which included entities of similar objects or semantic concepts as described in the captions. Moreover, even for images with highly similar local features, our model is capable of distinguishing between them and accurately retrieving the correct image at the top-1 rank.

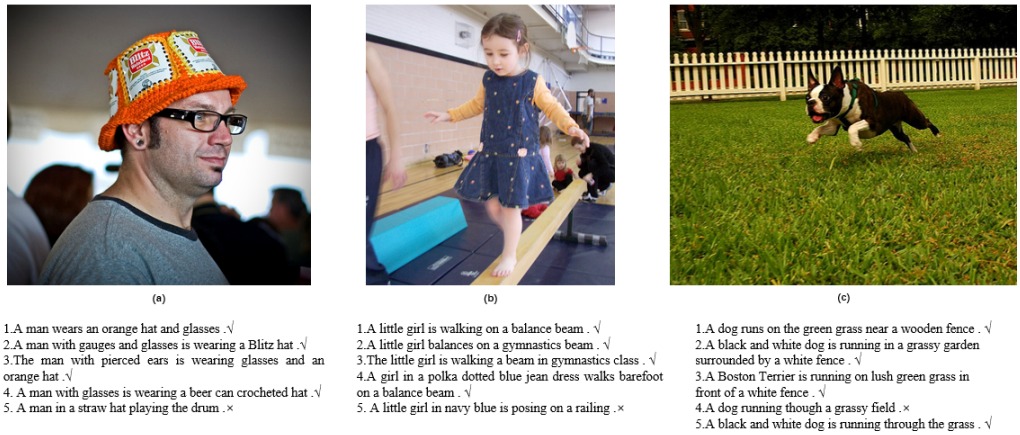


Figure 4. Qualitative results of the image-to-text retrieval on Flickr30K. For each image query, we show the top-5 ranked caption. The symbol $\surd$ represents the same with ground-truth captions, while $\times$ represents the mismatches.



Figure 5. Qualitative results of the text-to-image retrieval on MSCOCO dataset. We present a ranking of the top-3 retrieved images arranged from left to right for each textual query. The accurate matches are indicated by green boxes, while inaccurate matches are identified with red boxes.

5. Conclusions

In this paper, we propose a globally guided confidence enhancement network (GGCEN) for image-text matching, which enhances the confidence level of meaningful local matching pairs from a global perspective to obtain more accurate image-text matching results. Furthermore, the introduction of GPO pre-training to obtain global features and a self-attention mechanism further strengthen the results. A better global feature is used to guide fine-grained local interactions; thus, we not only improve the performance on image-text retrieval tasks but we double the convergence speed. Extensive experiments conducted on MSCOCO and Flikr30K datasets show that the proposed model surpasses the current best methods for image-text matching. Compared to the attention-based approach of simply aggregating from pairwise similarities between regions and words, we show that a well-represented global feature and the introduction of a global perspective in the local matching process to enhance the meaningful local matching pairs can further improve the performance of image-text matching. However, our method relies heavily on the accuracy of region detection and text description. Inaccurate region detection or ambiguous text description can lead to unreliable matching, which may affect the overall performance of our method. To address these limitations, future work could focus on developing more robust and accurate region detection and text description methods. Finally, further investigation could be conducted to explore better ways of combining the confidence of matched region-word pairs with local semantic similarities to achieve more accurate relevance measurement in image-text matching.

Author Contributions: Conceptualization, X.D. and G.T.; methodology, X.D.; software, X.D.; validation, X.D.; formal analysis, X.D.; investigation, X.D. and G.T.; resources, X.D. and G.T.; data curation, X.D.; writing—original draft preparation, X.D.; writing—review and editing, X.D., G.T. and M.W.; visualization, X.D.; supervision, G.T. and M.W.; project administration, G.T. and M.W.; funding acquisition, G.T. and M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Science Foundation of Autonomous. Region under Grant 2021D01C118 and in part by the Autonomous Region High-Level Innovative Talent Project under Grant 042419006.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets involved in this study are available in publicly accessible repositories.

Acknowledgments: The authors are sincerely grateful to the editor and anonymous reviewers for their insightful and constructive comments, which helped us improve this work greatly.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationship that could have appeared to influence the work reported in this paper.

References

1. Toyama, J.; Misono, M.; Suzuki, M.; Nakayama, K.; Matsuo, Y. Neural Machine Translation with Latent Semantic of Image and Text. Available online: https://arxiv.53yu.com/abs/1611.08459v1 (accessed on 14 February 2023).
2. Nam, H.; Ha, J.-W.; Kim, J. Dual Attention Networks for Multimodal Reasoning and Matching. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2156–2164.
3. Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1316–1324.
4. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
5. Gu, J.; Cai, J.; Joty, S.; Niu, L.; Wang, G. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7181–7189.
6. Huang, Y.; Wang, W.; Wang, L. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7254–7262.
7. Liu, Y.; Guo, Y.; Bakker, E.M.; Lew, M.S. Learning a Recurrent Residual Fusion Network for Multimodal Matching. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4127–4136.
8. Shi, B.; Ji, L.; Lu, P.; Niu, Z.; Duan, N. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Macao, China, 10–16 August 2019; pp. 5182–5189.
9. Wang, L.; Li, Y.; Lazebnik, S. Learning Deep Structure-Preserving Image-Text Embeddings. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5005–5013.
10. Li, K.; Zhang, Y.; Li, K.; Li, Y.; Fu, Y. Visual Semantic Reasoning for Image-Text Matching. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE Computer Soc.: Los Alamitos, CA, USA, 2019; pp. 4653–4661.
11. Zhu, L.; Lu, X.; Cheng, Z.; Li, J.; Zhang, H. Deep Collaborative Multi-View Hashing for Large-Scale Image Search. *IEEE Trans. Image Process.* **2020**, *29*, 4643–4655. [CrossRef] [PubMed]
12. Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked Cross Attention for Image-Text Matching. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11208, pp. 212–228; ISBN 978-3-030-01224-3.
13. Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; Han, J. IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12652–12660.
14. Chen, T.; Luo, J. Expressing Objects Just Like Words: Recurrent Visual Embedding for Image-Text Matching. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 10583–10590. [CrossRef]

15. Hu, Z.; Luo, Y.; Lin, J.; Yan, Y.; Chen, J. Multi-Level Visual-Semantic Alignments with Relation-Wise Dual Attention Network for Image and Text Matching. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Macao, China, 10–16 August 2019; pp. 789–795.
16. Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; Shao, J. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5763–5772.
17. Wehrmann, J.; Kolling, C.; Barros, R. Adaptive Cross-Modal Embeddings for Image-Text Alignment. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12313–12320. [CrossRef]
18. Wu, Y.; Wang, S.; Song, G.; Huang, Q. Learning Fragment Self-Attention Embeddings for Image-Text Matching. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2088–2096.
19. Zhang, C.; Yang, Z.; He, X.; Deng, L. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 478–493. [CrossRef]
20. Diao, H.; Zhang, Y.; Ma, L.; Lu, H. Similarity Reasoning and Filtration for Image-Text Matching. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 1218–1226. [CrossRef]
21. Liu, C.; Mao, Z.; Zhang, T.; Xie, H.; Wang, B.; Zhang, Y. Graph Structured Network for Image-Text Matching. *arXiv* **2020**, arXiv:2004.00277.
22. Zhang, H.; Mao, Z.; Zhang, K.; Zhang, Y. Show Your Faith: Cross-Modal Confidence-Aware Network for Image-Text Matching. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 3262–3270. [CrossRef]
23. Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; Wang, C. Learning the Best Pooling Strategy for Visual Semantic Embedding. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15784–15793.
24. Cui, H.; Zhu, L.; Li, J.; Yang, Y.; Nie, L. Scalable Deep Hashing for Large-Scale Social Image Retrieval. *IEEE Trans. Image Process.* **2020**, *29*, 1271–1284. [CrossRef] [PubMed]
25. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. *arXiv* **2018**, arXiv:1707.05612.
26. Karanwal, S. Analysis on Image Enhancement Techniques. *IJEM* **2023**, *13*, 9–21. [CrossRef]
27. Sultan, T.; Jahan, N.; Basak, R.; Jony, M.S.A.; Nabil, R.H. Machine Learning in Cyberbullying Detection from Social-Media Image or Screenshot with Optical Character Recognition. *Int. J. Intell. Syst. Appl.* **2023**, *15*, 1–13. [CrossRef]
28. Wang, Y.; Yang, H.; Qian, X.; Ma, L.; Lu, J.; Li, B.; Fan, X. Position Focused Attention Network for Image-Text Matching. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Macao, China, 10–16 August 2019; pp. 3792–3798.
29. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2017**, arXiv:1609.02907.
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
33. Schuster, M.; Paliwal, K. Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
35. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
36. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* **2015**, arXiv:1405.0312.
37. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.