Neural Computing and Applications (2022) 34:21387–21401
https://doi.org/10.1007/s00521-022-07617-3

ORIGINAL ARTICLE

GA-SRN: graph attention based text-image semantic reasoning network for fine-grained image classification and retrieval

Wenhao Li[1] · Hongqing Zhu[1] · Suyi Yang[2] · Pengyu Wang[1] · Han Zhang[1]

Received: 22 December 2021 / Accepted: 4 July 2022 / Published online: 27 July 2022

Abstract
In this paper, a new fine-grained image classification (FGIC) network with feature relationship enhancement of multiple stages is established. After the engaging of scene text in FGIC and retrieval, basic architecture of local, global, text feature encoders and classifier have been approved. This method retains these portions and expands them into a five-module architecture. In specific, positional encoding is incorporated to both local and textual feature encoders such that complementary information carried could engage in feature representation. In local and textual feature encoders, intra-modal semantic relation reasoning is introduced for FGIC by a proposed General Feature Relation Enhancement (GFRE) module. GFRE is a feature reasoning module applicable to any two inputs of same modality or distinct modalities. GFRE adopts Graph Attention which represents and infers relationships among graph data. Moreover, latest multi-modal reasoning module is improved by a proposed Multi-Head Multi-Modal Joint Semantic Reasoning module consisted of cross-modal GFREs by multi-head fusion. Experimental results on multiple datasets verify the effectiveness of the proposed algorithm.

Keywords Graph attention · Image classification · Positional information · Scene text · Fine-grained · Cross-model fusion

1 Introduction

Fine-Grained Image Classification (FGIC) refers to task that distinguishes images belonging to multiple sub-categories within a basic-level category. Compared with conventional image classification problem, FGIC is more challenging due to the small inter-class similarity and large intra-class variance. Recent researches use Convolution Neural Networks (CNNs) to learn global and local features [1], and combine multi-level feature to locate and encode the distinguishable areas for FGIC [2]. However, traditional CNNs often focus on most salient regions while neglecting other inconspicuous but distinguishable parts. Also, they treat various features in isolation manner but ignore the relationships between features. To alleviate these limitations, attention-based approaches are introduced recently such that other relatively insignificant but distinguishable parts are also noticed [3]. Although visual information has been extensively exploited, differences that are not obvious may not be accurately distinguished by a typical classification model.

Scene text with additional cues in natural images carries rich semantic information that may be highly relevant to object [4]. Recently, localizing and recognizing text in image has been well explored in many fields [5]. A representative work proposed by Movshovitz et al. [6] recognizes text instances and extracts text information for classifying street store images. However, solely using text information to conduct classification would be extremely challenging

✉ Hongqing Zhu
hqzhu@ecust.edu.cn

Wenhao Li
y30190695@mail.ecust.edu.cn

Suyi Yang
suyi.yang@kcl.ac.uk

Pengyu Wang
y10180292@mail.ecust.edu.cn

Han Zhang
y30190703@mail.ecust.edu.cn

[1] School of Information Science and Engineering, East China University of Science and Technology, MeiLong Road No.130, Shanghai 200237, China

[2] Department of Mathematics, Natural, Mathematical and Engineering Sciences, King's College London, Strand, London WC2R 2LS, England, UK

especially under blurred text instance circumstance. Later studies combine visual and textual information, but intrinsic relation exploration between two modalities becomes an open question. For example, Bai et al. [7] combined visual and textual information to train classifier. However, these methods simply concatenate visual and textual features without analyzing inter/intra-modal semantic relations. Also, relations between visual and textual information that might bring potential improvement to classification accuracy aren't fully exploited. Mafla et al. [8] used classical object detector and text detector to obtain visual and textual information with a Graph Convolution Network (GCN)-based embryonic cross-modal reasoning module. However, intra-modal relations in textual or visual modality haven't been explored.

In this paper, a novel Graph Attention (GAT)-based text-image Semantic Reasoning Network (GA-SRN) is established for FGIC. Considering that the position of the detected object also provides potential information, the position features of each image are obtained by Faster R-CNN. At the beginning, positional encoding is incorporated to both textual and visual semantic encoders in which inner-modal reasoning is also fulfilled. Besides, cross-modal semantic relation enhancement is further improved. In specific, a plug-and-play General Feature Relation Enhancement (GFRE) module widely applicable to semantic relation reasoning of inner-modal features or cross-modal features is proposed. It can exploit the relations between any two branches of features independent of their sources of either visual, textual or positional. Fig. 1 illustrates the proposed semantic relation enhancement model. Although simple concatenation of visual and textual features from classical CNN and Optical Character Recognition (OCR) fulfills basic functionality as well [7], potential improvement could be explored by employing self-attention mechanism. In GFRE, advanced GAT [9] is adopted and is firstly introduced to FGIC network inspired by the GCN that designed specifically for processing graph data. By GFRE module, discriminative feature representation for visually similar categories could be enhanced. In the proposed GFRE for intra-modal reasoning, we can obtain position graphs containing indicative class information which graph generated from images belonging to similar classes should somehow see some similarities. In general, the position graph aids text and local encoding in inner-modal reasoning and ultimately engaged in cross-modal reasoning which sometimes has a big effect. For cross-modal semantic relationship reasoning, a Multi-Head Multi-Modal Joint Semantic Reasoning (M3JSR) module is proposed which conducts multi-modal feature relationship enhancement by a GFRE and a multi-head fusion. It is used to reason and fuse features from different modalities to generate discriminative features that can be used for classification. Experiments on publicly available datasets demonstrate the effectiveness of the proposed framework.

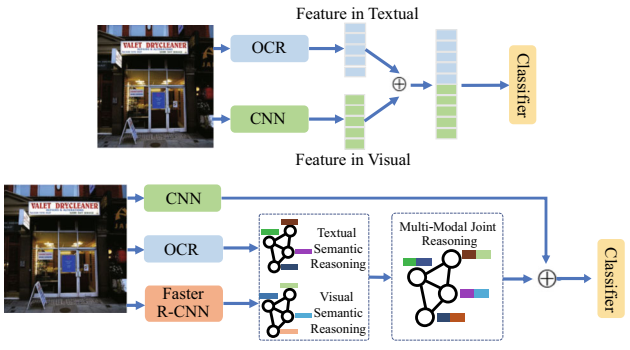

Fig. 1 Traditional mode of scene text based FGIC (top) in comparison to the proposed method with semantic reasoning modules (bottom). Instead of concatenation between features, relationship enhancement is operated both within and across modalities

The main contributions of this paper are summarized as follows:

- Intra-modal semantic relationship enhancement is firstly introduced to FGIC.
- Position information is plugged into textual and visual encoders for feature representation and intra-modal relationship reasoning.
- An easily applicable GFRE module independent of modalities with efficient GAT is proposed. It can reason both intra-modal positional relations and inter-modal relations between any two branches of features independent of their sources.
- The M3JSR module brings improvement to latest multi-modal reasoning module. This module fuses different modality features and generates more discriminative features for classification by cross-modal GFREs with multi-head fusion strategy.

This paper is organized as follows. Section 2 presents the related works about FGIC and scene text recognition. The methodology is introduced in Sect. 3. Experimental results and comparison are illustrated in Sect. 4. Discussion and conclusion are drawn in Sect. 5.

2 Related works

2.1 Fine-grained image classification

Recently, fine-grained image analysis with CNN has received extensive attention in computer vision communities [10]. FGIC requires algorithms to gain discriminative visual regions and classify objects through detailed regional features. Sun et al. [11] introduced a Searching Discriminative Regions (SDR) and Learning Discriminative Regions (LDR) based method using attention mechanism to search for high-

response regions in images and take them as clues to locate local discriminative regions. Zhang et al. [12] designed Intra-class Part Swapping (InPS) which avoids inter-class mixing, and thus alleviates label noise in the mixing process for FGIC. Since these studies solely classified based on visual features, later studies raised the idea of combining textual and visual features to improve classification. For example, He et al. [13] introduced a two-stream model Combining Vision and Language (CVL) for learning latent semantic representations. As vision stream and language stream are complementary classification accuracy can be achieved. Karaoglu et al. [14] utilized pre-trained GoogLeNet to extract text instances in scene images and build a word directory along with visual feature extraction. Bai et al. [7] used Textboxes as text instance detector and gained visual features with CNN-based network. Recent work by Mafla et al. [4] used pyramidal histogram of characters with Fisher Vector to obtain text instances which resulted in improving OCR accuracy. In general, fundamental structure for text detection in FGIC has just been verified effective and further interaction between features hasn't been explored.

2.2 Scene text detection and recognition

Most methods for scene text recognition are divided into two stages, detection and recognition. Jaderberg et al. [5] obtained text region proposals with CNN-based network and used a classifier to classify text instances into words. Current studies tended to seek for better designed detector for text identification. For example, recent studies employed improved Faster R-CNN [15] as text detector to construct an end-to-end trainable scene text detection. Borisyuk et al. [16] proposed a Rosetta system with Faster R-CNN as text detector. He et al. [17] used a Long-Short Term Memory (LSTM) [18] to refine bounding boxes for recognition. Later, attention mechanism widely proved effective in many fields is also introduced to text recognition. Zhang et al. [19] applied an attention unit between CNN-based encoder and GRU [20]-based decoder to adapt location of character. In addition, other approaches such as PHOC [21] have also been used in querying text instances in natural scene images [22]. However, mainstream scene text detection adopted in FGIC only uses OCR to have the model complexity within control.

2.3 Graph attention networks

Graph Neural Network (GNN) [23] was recently proposed to process graph data directly by message passing between nodes. Gao et al. [24] for example used GNN to exploit relations of textual and visual instances. Li et al. [25] employed GCN to reason salient regions in images and text words for image-text matching. However, due to same weight to adjacent nodes in same neighborhood order, capture on spatial information correlation is restricted. GAT [9] that strengthen information of crucial nodes by updating attentional weights could better integrate correlation of features into model. For instance, Li et al. [26] used GAT to exploit relations of visual regions and bounding boxes for VQA. Wen et al. [27] proposed a dual semantic relation model based on GAT for text-image matching. Zeng et al. [28] generated a graph of all sub-sentences with a strong connection by using sentence-level GAT. Chen et al. [29] proposed a Hierarchical Graph Reasoning (HGR) model that employed attention-based graph reasoning to generate hierarchical textual embedding for fine-grained video-text matching. In this paper, we employ GAT to reason different concept information and enhance representation of image features.

2.4 Multi-modal fusion and relation

Since FGIC places high interaction to features between modalities, multi-modal feature fusion strategies such as Multi-modal Low-rank Bilinear Attention Network (MLB) [30], Block [31] and Visual-Semantic Aggregator [24] are also explored. Previously, Anderson et al. [32] proposed a bottom-up-attention for VQA that established relationship between regions and words. Kazemi et al. [33] concatenated image and text features to compute multiple attention distributions. Also, LSTM reasoning textual and visual semantic relations is proposed [34]. Later, GCN is discovered more suitable for exploiting semantic relations between textual and visual instances. Recently, GAT is applied to feature reasoning in text-image matching task [27]. To our best knowledge, it is the first attempt by this paper that GAT is used in the task of FGIC as well as exploiting relations between positional information and visual regions/text instances.

3 Methodology

As shown in Fig. 2, the proposed GA-SRN consists of five parts, Global Feature Encoder, Local Feature Encoder, Text Encoder, Multi-Head Multi-Modal Joint Semantic Reasoning Module and Classifier. The most similar combination of these modules was by latest FGIC method [8]. The proposed method follows the same structure while pattern of positional encoding is changed, inner-modal relationship exploitation is added, and cross-modal reasoning is enhanced. Specifically, retrieval image is delivered to a global feature encoder unit same as majority previous designs, along with the modified local feature encoder and text encoder with the proposed GFRE module.

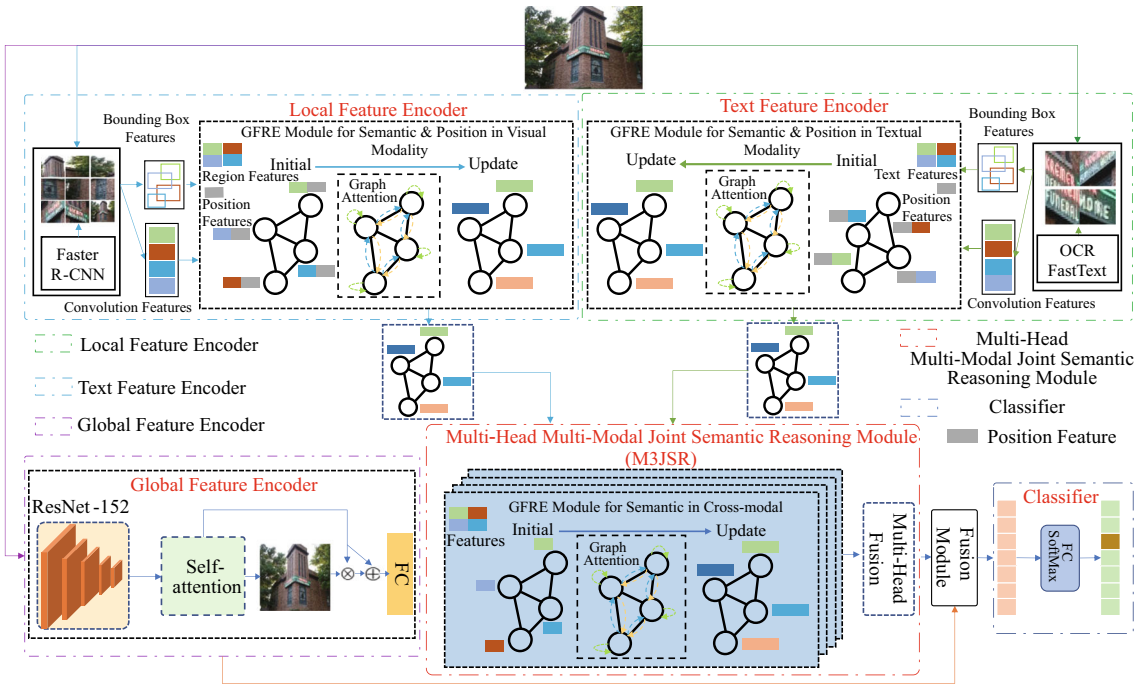

Fig. 2 Overall flowchart of the proposed model. Five modules are distinguished by dotted-line boxes in different colors. GFRE exploiting intra-modal or inter-modal feature relations is inserted. M3JSR strengthen relations of the relevant image regions and text words to generate discriminative features for classification

In this architecture, GFRE is applied to exploit semantic and positional relationship in local feature encoder and text feature encoder as well as semantic relationships between text and visual features in M3JSR. M3JSR is a new designed multi-modal reasoning module combining GFRE and multi-head fusion to generate more discriminative features for classifier. At the same time, M3JSR with GFRE adopts GAT which is regarded more effectively than GCN adopted earlier [8] in final cross-modal fusion.

3.1 Global feature encoder

We use ResNet-152 [16] pre-trained on ImageNet as global feature extractor. For example, after input image goes through ResNet-152, we get the original global image feature defined as V_G. Afterwards, a self-attention mechanism is used to obtain more discriminative features. By this attention mechanism, we obtain an attention mask $\text{attn}_{\text{mask}}$ which pays different attention to different regions. The attention weights are learned in an end-to-end way by utilizing convolution of 1×1 kernel with a Softmax function followed by. To obtain the complete global feature, we multiply $\text{attn}_{\text{mask}}$ with original global feature V_G and add it to the latter. The result is fed into a Fully Connected (FC) layer to get the final feature V_G^*:

$$V_G^* = \text{FC}(V_G + (V_G \times \text{attn}_{\text{mask}})). \tag{1}$$

3.2 General feature relationship enhancement module

Inspired by image-text matching study [27] that exploits regional relations with GAT, we designed GFRE module to reason relations of different features. Since GFRE can exploit the relationship between any two branches of features independent of their sources, input types I and II are used to denote these particular inputs. As shown in Fig. 3, input features V_S from type I are processed by Relation Unit I to obtain feature graph $G_S = (V_S, E_S)$, where E_S is an edge set denoted as the affinity matrix obtained by calculating affinity edge of each group of feature ${v_s}^i$ and ${v_s}^j$

$$E_S({v_s}^i, {v_s}^j) = ({v_s}^i)^T {v_s}^j, \tag{2}$$

where i and j represent the i-th and the j-th feature pair. In fact, edges of semantically relevant features of input feature I would have high affinity scores. Then, with a GAT module, the relation-enhanced feature of v_s^i can be defined as v_s^{i*}

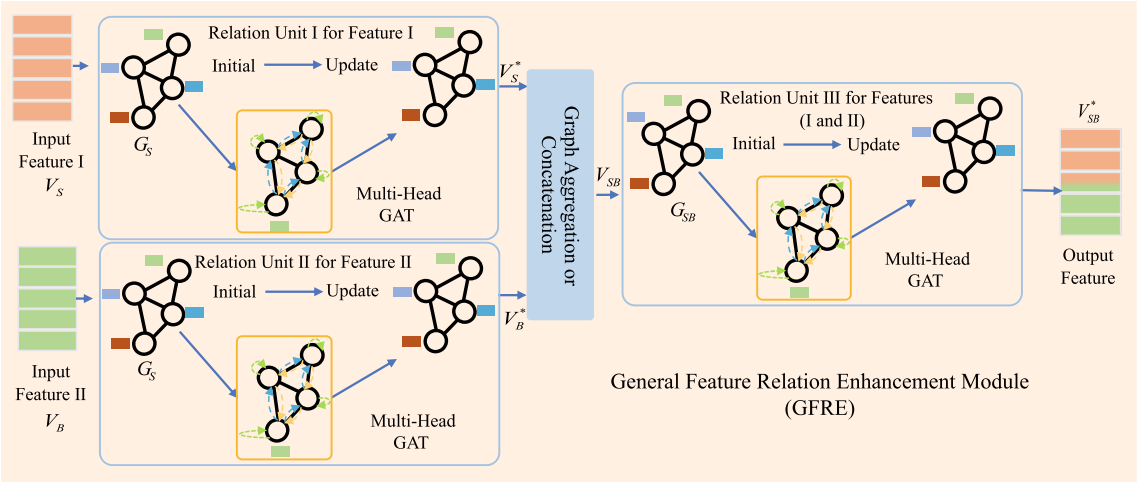

Fig. 3 Architecture of the proposed GFRE. It contains three relation units with GAT in which Relation Unit I and II are input by two sources of features (I and II) and their semantic relationship is exploited in Relation Unit III

$$v_s^{i*} = \mathrm{BN}\left(\mathrm{ReLU}\left(\sum_{j\in N_s^i} W_o \|_{h=1}^{H}\left(\mathrm{head}_s^1,\ldots,\mathrm{head}_s^h\right)\right)\right), \tag{3}$$

where $\|$ represents concatenation.

$$\mathrm{head}_s^h = \mathrm{Softmax}\left(W_{\mathrm{sq}}^h v_s^i \left(W_{\mathrm{sl}}^h v_s^j\right)^T/\sqrt{D_s}\right) W_{\mathrm{sv}}^h v_s^j. \tag{4}$$

where $W_{\mathrm{sq}}^h \in \mathbb{R}^{D_s\times d}$, $W_{\mathrm{sl}}^h \in \mathbb{R}^{D_s\times d}$, $W_{\mathrm{sv}}^h \in \mathbb{R}^{D_s\times d}$ and $W_o \in \mathbb{R}^{D_s\times d}$ are learnable parameters, D_s is vector dimension, N_s^i is the neighborhood of node i in graph G_S. Following parameter selection in [27], we employ $H = 8$ and $d = D_s/8$ in this model.

Similarly, Relation Unit II extracts input features V_B from type II to construct graph $G_B = (V_B, E_B)$, where E_B is the edge set defined by an affinity matrix calculating the affinity edge of each group of feature ${v_b}^i$ and ${v_b}^j$.

$$E_B\left({v_b}^i, {v_b}^j\right) = \left({v_b}^i\right)^T {v_b}^j. \tag{5}$$

In fact, edges of semantically relevant features in input feature II would gain high affinity score. Then, GAT is used to obtain relation-enhanced features v_b^{i*} by

$$v_b^{i*} = \mathrm{BN}\left(\mathrm{ReLU}\left(\sum_{j\in N_b^i} W_o \|_{h=1}^{H}\left(\mathrm{head}_b^1,\ldots,\mathrm{head}_b^h\right)\right)\right), \tag{6}$$

where

$$\mathrm{head}_b^h = \mathrm{Softmax}\left(W_{\mathrm{bq}}^h v_b^i \left(W_{\mathrm{bl}}^h v_b^j\right)^T/\sqrt{D_b}\right) W_{\mathrm{bv}}^h v_b^j, \tag{7}$$

where $W_{\mathrm{bq}}^h \in \mathbb{R}^{D_b\times d}$, $W_{\mathrm{bl}}^h \in \mathbb{R}^{D_b\times d}$, $W_{\mathrm{bv}}^h \in \mathbb{R}^{D_b\times d}$ and $W_o \in \mathbb{R}^{D_b\times d}$ are learnable parameters. N_b^i is the neighborhood of node i in graph G_B. Same as in Unit I, $H = 8$ and $d = D_b/8$ is taken. By Relation Unit I and II, intrinsic relations of features from each type could be found individually.

Finally, relation-enhanced feature $V_S^* = \{v_s^{1*}, v_s^{2*}, \ldots, v_s^{n*}\}$ from feature type I and $V_B^* = \{v_b^{1*}, v_b^{2*}, \ldots, v_b^{m*}\}$ from feature type II are concatenated to obtain feature $V_{\mathrm{SB}} = \{v_{\mathrm{sb}}^1, v_{\mathrm{sb}}^2, \ldots, v_{\mathrm{sb}}^n\}$, $v_{\mathrm{sb}}^i \in \mathbb{R}^{D_{\mathrm{sb}}}$, where $D_{\mathrm{sb}} = D_s + D_b$. Aggregation was proved to explore relationship between two features well [24], but we testify that directly concatenation between two features followed by another relationship unit seems to be more effective and thus is adopted in our GFRE module. After that, we construct graph $G_{\mathrm{SB}} = (V_{\mathrm{SB}}, E_{\mathrm{SB}})$, where V_{SB} contains crucial information from feature I and feature II as well as their relations. Specifically, we use feature v_{sb}^i to initial node i of G_{SB}, then we compute the affinity matrix E_{SB} of feature v_{sb}^i and v_{sb}^j to initial edge between nodes i and j in G_{SB}. The affinity matrix E_{SB} of edge set can be expressed as:

$$E_{\mathrm{SB}}\left({v_{\mathrm{sb}}}^i, {v_{\mathrm{sb}}}^j\right) = \left({v_{\mathrm{sb}}}^i\right)^T {v_{\mathrm{sb}}}^j. \tag{8}$$

where v_{sb}^i and v_{sb}^j are features which are obtained from V_S and V_B, respectively.

Another graph attention module (Relation Unit III) is used to process relation-enhanced fused graph G_{SB} and exploit relations between features from Relation Unit I and Unit II. The relation-enhanced feature v_{sb}^{i*} is obtained by

$$v_{\mathrm{sb}}^{i*} = \mathrm{BN}\left(\mathrm{ReLU}\left(\sum_{j\in N_{\mathrm{sb}}^i} W_o \|_{h=1}^{H}\left(\mathrm{head}_{\mathrm{sb}}^1,\ldots,\mathrm{head}_{\mathrm{sb}}^h\right)\right)\right), \tag{9}$$

where

$$\text{head}_{\text{sb}}^{h} = \text{Softmax}\left(W_{\text{sbq}}^{h} v_{\text{sb}}^{i} \left(W_{\text{sbl}}^{h} v_{\text{sb}}^{j}\right)^{T} / \sqrt{D_{\text{sb}}}\right) W_{\text{sbv}}^{h} v_{\text{sb}}^{j}, \tag{10}$$

where $W_{\text{sbq}}^{h} \in \mathbb{R}^{D_{\text{sb}} \times d}$, $W_{\text{sbl}}^{h} \in \mathbb{R}^{D_{\text{sb}} \times d}$, $W_{\text{sbv}}^{h} \in \mathbb{R}^{D_{\text{sb}} \times d}$ and $W_o \in \mathbb{R}^{D_{\text{sb}} \times d}$ are learnable parameters. N_{sb}^{i} is the neighborhood of node i in graph G_{SB}. $H = 8$ and $d = D_{\text{sb}}/8$ is employed in (10). The following pseudo code indicates the implementation details of GFRE module.

/* General Feature Relation Enhancement (GFRE) */
Input: Type I feature V_S and type II feature V_B
Output: Relation-enhanced feature V_{SB}^{*}
for each $x \in (S, B)$ **do**
  Initial Graph G_x from V_x;
  for each $i \in [1, node\_number]$ **do**
    for each $j \in N_x^i$ **do**
      for each h **do** $head_x^h$
      end for
      $Heads = W_o \|_{h=1}^{H} (head_x^1, \ldots, head_x^h)$;
    end for
    $all\_Heads = \sum_j Heads$;
    $v_x^{i*} = BN(ReLU(all\_Heads))$;
  end for
end for
$V_{SB} = [V_S^{*}; V_B^{*}]$. Initial Graph G_{SB} from V_{SB},
for each $i \in [1, node\_number]$ **do**
  for each $j \in N_{sb}^i$ **do**
    for each h **do** $head_{sb}^h$
    end for
    $Heads = W_o \|_{h=1}^{H} (head_{sb}^1, \ldots, head_{sb}^h)$;
  end for
  $all\_Heads = \sum_j Heads$;
  $v_{sb}^{i*} = BN(ReLU(all\_Heads))$;
end for

3.3 Local feature encoder

Inspired by works for VQA [26] and text-image matching task [27], our framework uses local feature encoder to detect salient image regions and encode these regions into local features. We employ Faster R-CNN pre-trained with Visual Genome [35] as the extractor to detect and encode salient image regions. We set the Intersection over Union (IoU) threshold at 0.7 and a confidence score threshold at 0.3, and select top n Region Of Interest (ROI) after sorting the predicted regions. Therefore, we obtain a set of salient region features $R = \{r_1, r_2, \ldots, r_n\}$, $r_i \in \mathbb{R}^D$ where $D = 2048$ and a set of bounding boxes of salient regions $B = \{b_1, b_2, \ldots, b_n\}$, $b_i \in \mathbb{R}^4$, r_i represents the i-th predicted region and $b_i = \{x_{i1}, y_{i1}, x_{i2}, y_{i2}\}$ denotes the bounding box of the i-th predicted region. Suggested by study [36] for image retrieval task, the position of salient regions would contributes to visual-text joint-embedding learning. Therefore, to enhance the representation of local visual embedding, we encode salient region position into positional features denoted as $V_B = \{v_b^1, v_b^2, \ldots, v_b^n\}$ which is transferred to 64 dimensions by an FC layer. To generate 1024 dimensional relation-enhanced features, we first employ an FC layer to transform 2048-dimensional embedding v_i obtained from original Faster R-CNN [15] into $V_R = \{v_r^1, v_r^2 \ldots v_r^n\}$, $v_r^i \in \mathbb{R}^{D_r}$, where D_r=960 in our model. Then, we feed them into GFRE and obtain the semantic positional enhanced local feature represented as $V_{\text{RB}}^{*} = \{v_{\text{rb}}^{1*}, \ldots v_{\text{rb}}^{n*}\}$, $v_{\text{rb}}^{i*} \in \mathbb{R}^{D_{\text{rb}}}$, where $D_{\text{rb}} = 1024$.

$$V_{\text{RB}}^{*} = \text{GFRE}(V_R, V_B). \tag{11}$$

3.4 Text feature encoder

To obtain text instance features, we employ the stable text detection model Google OCR[1]. By this model, we extract each word appeared represented as t_i and the corresponding bounding box of each word t_i denoted $b_w^i = \{x_{i1}, y_{i1}, x_{i2}, y_{i2}\}$. After recognizing the words, we use FastText [37] to generate word embedding $W = \{w_1, w_2, \ldots, w_m\}, w_i \in \mathbb{R}^D$, where $D = 1024$. To embed text positions, same method as obtaining positional embedding in local feature encoder is used and the text positional features obtained are denoted as $B_W = \{b_w^1, b_w^2, \ldots, b_w^m\}$, $b_w^i \in \mathbb{R}^{D_b}$ where D_b = 64. Then, we employ FC layer to transform 1024 dimensional w_i to 960 dimensional v_t^i. Hence, the set of word embedding can be described as $V_T = \{v_t^1, v_t^2, \ldots, v_t^m\}$, $v_t^i \in \mathbb{R}^{D_t}$, where D_t= 960. Then, we feed positional features B_w and regional features V_T into GFRE module and obtain the semantic positional enhanced textual feature represented as $V_{\text{TB}}^{*} = \{v_{\text{tb}}^{1*}, \ldots, v_{\text{tb}}^{n*}\}$, $v_{\text{tb}}^{i*} \in \mathbb{R}^{D_{\text{tb}}}$, where D_{tb} = 1024.

$$V_{\text{TB}}^{*} = \text{GFRE}(V_T, B_W). \tag{12}$$

3.5 Multi-head multi-modal joint semantic reasoning module

Having obtained the relation-enhanced features fused from local feature encoder and text feature encoder, respectively, another cross-modal reasoning module would have to be designed for joint relationship inference before final feature fusion. In this paper, M3JSR is proposed to exploit relations between text instances and salient regions as shown in Fig. 4. The module firstly applies GFRE to cross-modal semantic relation enhancement with multi-head fusion for joint feature representation. In fact, the M3JSR with multi-head fusion provides slightly more competitive results, while replacing M3JSR with only one GFRE could also

[1] https://cloud.google.com/vision/docs/ocr.

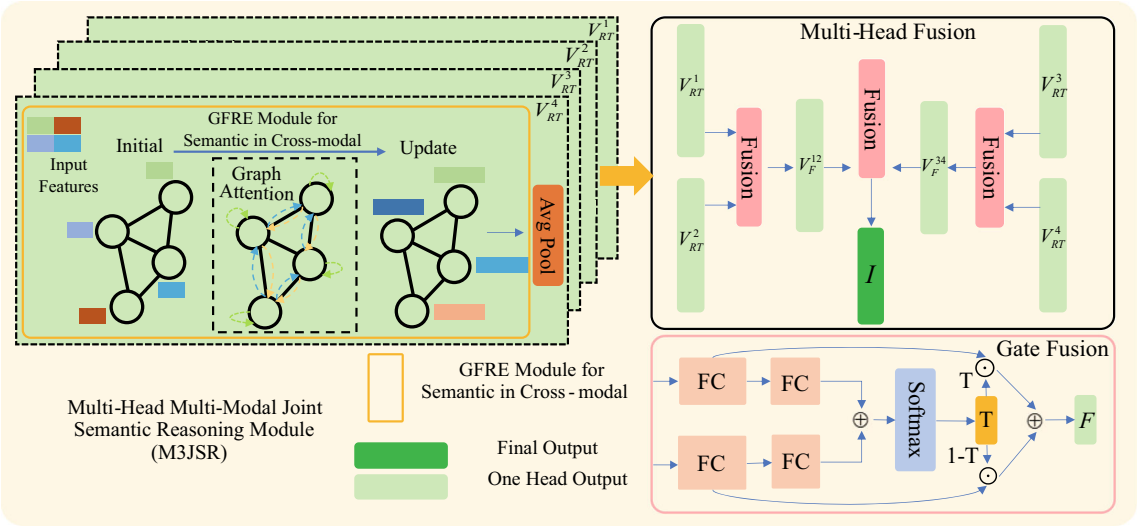

Fig. 4 Structure of the proposed M3JSR. It implements multi-head fusion in which heads refer to four cross-modal GFREs, respectively

realize basic functionality of relation-enhanced feature fusion if less computation is appreciated.

In specific, given the output of local feature encoder V^*_{RB} and the output of text encoder V^*_{TB}, they would be input to Relation Unit I and II, respectively, of another cross-modal GFRE model. Then, we concatenate or aggregate the relation-enhanced visual and textual features to generate features containing both visual and textual semantic information where their intrinsic relationship is exploited by the following Relation Unit III. After applying GFRE module, we obtain visual textual semantic relation-enhanced features which can be represented as V^*_{RT}. The whole process can be described as:

$$V^*_{\mathrm{RT}} = \mathrm{GFRE}(V^*_{\mathrm{RB}}, V^*_{\mathrm{TB}}). \tag{13}$$

Then, an average pooling layer is applied to obtain the final output V_{RT} of one head.

$$V_{\mathrm{RT}} = \mathrm{AvgPool}(V^*_{\mathrm{RT}}). \tag{14}$$

M3JSR contains four heads and thus four textual visual semantic relations-enhanced features are output. Having obtained a set of four output features $V_{\mathrm{RT}} = \{V^1_{\mathrm{RT}}, V^2_{\mathrm{RT}}, V^3_{\mathrm{RT}}, V^4_{\mathrm{RT}}\}$, we apply them with a specifically designed fusion module for M3JSR. This fusion module contains simpler fusion layer which is illustrated in the right bottom of Fig. 4. In our proposed M3JSR, we use multiple GFREs to reason the correspondences in textual and visual patterns to preserve the diversity of relations, but this approach may generate redundant information. Therefore, we should extract important information from the multi-heads generated features. As shown in Fig. 4, the proposed multi-head fusion uses multi-layers of Gate Fusion to refine redundant data. Compared to self-attention strategy, the proposed multi-head fusion method is more flexible as the number of heads is selected at its optimal such that the best number of Gate Fusion layers can be used to refine information. This fusion layer fuses vectors V^i_{RT} with V^j_{RT}, and generates output V^{ij}_F. We formulate the fusion layer as follow.

$$V^i_F = W_i V^i_{\mathrm{RT}}, \quad V^j_F = W_j V^j_{\mathrm{RT}}, \quad t = \delta\left(U_i V^i_F + U_j V^j_F\right) \tag{15}$$

$$V^{ij}_F = t\Theta V^i_F + (1-t)\Theta V^j_F, \tag{16}$$

where W_i, W_j, U_i, U_j are learnable parameters and δ represents the Sigmoid function. In order to fuse the four head output features from $V_{\mathrm{RT}} = \{V^1_{\mathrm{RT}}, V^2_{\mathrm{RT}}, V^3_{\mathrm{RT}}, V^4_{\mathrm{RT}}\}$ to generate terminal feature I, we formulate the fusion process as follow.

$$I = F_3\left(F_1\left(V^1_{\mathrm{RT}}, V^2_{\mathrm{RT}}\right), F_2\left(V^3_{\mathrm{RT}}, V^4_{\mathrm{RT}}\right)\right), \tag{17}$$

where F_1, F_2 and F_3 represent three simplified fusion layers. The proposed multi-head fusion scheme is relatively flexible, because the number of Gate Fusion layers can be determined by its performance of information refinement. Relevant experiments of heads are shown in Sect. 4.3.

3.6 Classifier

Finally, outputs of global encoder and M3JSR module are concatenated to obtain a final 2048 dimensional vector $\Omega = [V_G{}^*, I]$ which is then input to a FC layer for classification. After applying a Softmax to the output of final FC layer, we obtain a probability distribution function indicating which the probability that input image belongs to each class. Overall, our model can be trained in end-to-end

manner and a cross-entropy loss function is used to optimize the model expressed as

$$J(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} y_i^n \log(p_i^n). \tag{18}$$

The pseudo code of the proposed GA-SRN can be summarized as follows.

/* The proposed GA-SRN algorithm */

Input: Set of salient region features R; bounding boxes of salient regions B; set of text embeddings W; features of textual boxes B_W; global feature V_G^*;

Output: Classification results

function $F(i,j)$
  $V_F^i = W_i V_{RT}^i$, $V_F^j = W_j V_{RT}^j$;
  $t = \delta(U_i V_F^i + U_j V_F^j)$, $V_F^{ij} = t\Theta V_F^i + (1-t)\Theta V_F^j$;
end function
Encode B into V_B; encode R into V_R; encode W into V_T;
$V_{RB}^* = GFRE(V_R, V_B)$; $V_{TB}^* = GFRE(V_T, B_W)$;
for each $i \in [1, K]$ **do**
  $V_{RT}^{*i} = GFRE(V_{RB}^*, V_{TB}^*)$; $V_{RT}^i = AvgPool(V_{RT}^{*i})$;
end for
$V_{RT} = \{V_{RT}^1, V_{RT}^2, V_{RT}^3, V_{RT}^4\}$
$I = F_3(F_1(V_{RT}^1, V_{RT}^2), F_2(V_{RT}^3, V_{RT}^4))$, $F_i, i \in [1,2,3]$ is function F;
$\Omega = [V_G^*, I]$
Input F into classifier;

4 Experimental results

4.1 Datasets

(1) *Con-Text dataset*. This dataset contains 24,255 images of 28 categories and is divided into three-folds to make up training and test sets. Although text is appreciated in visually similar scene for better certainty of FGIC, this dataset for fine-grained scene classification with many images containing no text instance is also engaged as a challenge to the method.

(2) *Drink-Bottle dataset*. It was firstly provided by Bai et al. [7] and is composed by 20 sub-categories of soft drink and alcoholic drink. This dataset contains 18,488 images and is divided into three-folds as well to construct the training and test sets. Since the dataset is designed for FGIC, a certain number of images contain text instances.

(3) *CUB-200-2011*. It contains 11,788 bird images spanning 200 sub-categories. The training dataset has 5994 images, and the test set has 5794 images. It provides text descriptions of the language modality for each bird image and part location label for different parts on body. In experiment, the image description label is used as text and the part location label is used as salient region. Half of the images in this dataset are used for training and the other half are used for testing.

4.2 Implementation details

To extract salient visual regions of the input image, we get the top $n = 36$ ROI following [15] and embed them along with the positional information into 1024 dimensional embedding. The recognized words are ranked by confidence score from highest to lowest and the top $m = 15$ predictions are reserved. Then, the reserved words are encoded into 300 dimensional embedding using the pre-trained FastText model. Faster R-CNN in local feature encoder module and OCR model in text encoder are used as the feature extractor. Text instances recognizer has been pre-trained and would not be updated during training stage. Our model is trained 20 epochs in total and is optimized by RAdam with batch size 64. At the training stage, initial learning rate is set as 0.001 which decays by scale of 0.1 on the 3th, 6th, 12th and 18th epochs. This network is implemented on PyTorch 1.5.1. We conduct all experiments on a server with AMD Ryzen 7 3700X CPU and NVIDIA GeForce RTX 2070s GPU.

4.3 Effect of GA-SRN on FGIC

(1) *Evaluation on multi-head numbers.* The proposed M3JSR reasoning relations between visual regions and textual words is composed of cross-modal GFREs with multi-head fusion. In this subsection, how the number of GFRE heads K in M3JSR module affects fusion performance is testified as shown in Fig. 5. It could be seen that classification accuracy presents sensitive to head numbers in M3JSR module. Classification mean Average Precision (mAP) is relatively similar when $K = 2$ on both datasets. However, setting $K = 4$ resulted in highest mAPs at 86.57 and 79.96 on Con-Text and Drink-Bottle datasets, respectively. In general, classification

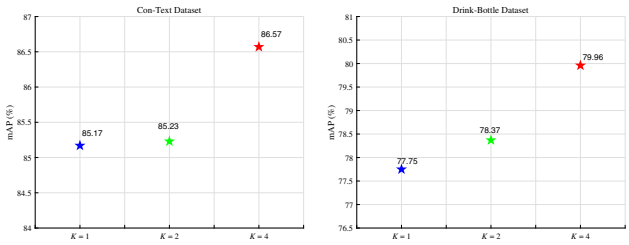

Fig. 5 Classification results mAP (%) by $K = 1, 2, 4$ in multi-head fusion on Con-Text and Drink-Bottle datasets. K is GFRE head numbers in M3JSR module

accuracy rises when number of heads increase since multi-head graph attention enables each head to focus on more local-textual relations. However, too large number of heads would lead to times of increase in algorithm complexity, so K is set to 4 in the proposed method.

(2) *Qualitative results.* Qualitative results of our model on individual images are shown in Fig. 6. In this figure, Ground truth and top-2 probability scores obtained by our model are shown below each image. It is noticeable that images could be classified to the correct descriptions in most cases with relatively high certainty. This might be because of the semantic relationship in visual features and textual features that our model discovers to enhance classification performance. It could also be observed that specific brand could be easily classified (images in Drink-Bottle) and images with scene text (such as "Theatre", "Barber" images) could also be well classified even if the scene texts aren't very related to retrieval results. This might be benefited from that text feature encoder comprehensively exploits semantic relation with positional feature. Classified examples such as "Tea House", "Cafe" and "Dinner" are visually similar, the proposed model can distinguish them with relatively high certainty. However, wrong classification cases still exist such as the last two samples listed for each dataset. It could be observed that images without textual words may suffer from higher probability of misclassification. In images with scene text, recognition errors may still occur when textual words in images are blurry or distorted.

Table 1 Classification results mAP (%) of GA-SRN and state-of-the-art methods on Con-Text and Drink-Bottle datasets. The method labeled with * is established in ensemble mode and the bold contents represent the best results

Methods	Con-text	Drink-bottle
Karaoglu et al. [38]	39.00	–
Karaoglu et al. [14]	77.30	–
Bai et al. [7]	78.90	–
Bai et al. [7]	79.60	72.80
Mafla et al. [4]	80.20	77.40
Mafla et al. [8]	85.81	79.87
Ours	**86.57**	**79.96**

(3) *State-of-the-art comparison.* In this section, quantitative performance of the proposed GA-SRN is compared with latest image classification methods Karaoglu et al. [38]; [14], Bai et al. [7], Mafla et al. [4] and Mafla et al. [8]. As shown in Table 1, classification results on Con-Text and Drink-Bottle datasets referring to original references are listed. Bai et al. [7] explores visual and textual relationship with attention mechanism. This architecture achieves a satisfied classification accuracy for both single classifier version (Bai et al. [7]) and the ensemble model with three classifiers (Bai et al. [7]). Although [4] is not trained in end-to-end manner, the efficient off-line estimation of Fisher Vector by training a Gaussian Mixture Model (GMM) resulted in noticeable increase in classification accuracy (80.20% and 77.40%). In addition, [8] brings noticeable improvement to previous methods which classification accuracy arrives at 85.81% and 79.87%. By comparison,

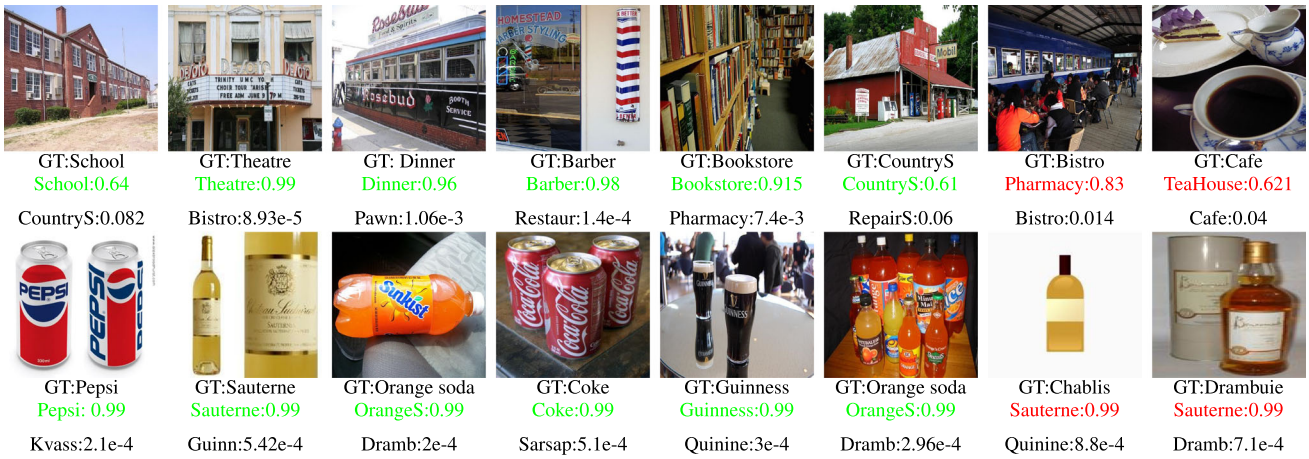

Fig. 6 Classification results mAP (%) on Con-Text (up) and Drink-Bottle (down). Ground truth and the categories with top-2 probabilities obtained by our model are shown. Text in red denotes incorrect predictions and text in green represents correct predictions

the proposed method with inner-modal and cross-modal reasoning with positional features and GAT inserted achieves slightly better results at 87.57% and 79.96% on Con-Text and Drink-Bottle datasets, respectively.

As a scene-based image classification method, we can also conduct it on fine-grained content-based image classification. Results in comparison with He et al. [13], Zhang et al. [12], Chang et al. [39], Luo et al. [40] and Sun et al. [11] on CUB-200-2011 are shown in Fig. 7. The dataset provides images with single object, image description labels and part location labels, but each image tends to be content-focused and there is no scene text. Methods target specifically at these types of queries mostly use object localization and exploit the relations between the sentence for descriptions and local regions. For comparison, we adjust our text encoder by directly applying the image description labels as text and part location labels as salient regions. We noticed that as content-based images by CUB-200-2011 provide slightly less perplexing scene, our method still stays at a good level of classification accuracy. An mAP of 88.72% is received, 1.16% higher than the third but mildly falls behind than the highest 88.9% by Sun [11]. This might be because of the specific focus on locating highly discriminative regions based on high-response region searching by Sun [11] which is extremely contributing to content-based image retrieval.

(4) *Images without scene text.* As an integrated method with sematic reasoning, we would still like to examine its generalization ability on images either contain a scene text (I + T) or not (I – T). As shown in Table 2, it could be seen that all methods on I + T have higher classification results, which verifies the idea that engaging text encoding and imposing semantic relation reasoning are useful. On the two datasets, our method receives higher mAP on both I + T and I – T. For I + T images classification, we achieve 87.52 and 83.53 which are both approximately 1% higher than Mafla et al. [8] leading at the second place. On I – T images, the text feature encoding probability wouldn't engage in semantic reasoning, but the essential structure of GFRE with positional reasoning on local features of purely scene graph still provides outstanding network performance (75.28 and 70.26).

(5) *Ablation studies.* In this section, GFRE and M3JSR is ablated on three datasets, and input features of different types are cut off. As shown in Table 3, classification results with (w/) or without (w/o) GFRE and M3JSR is compared to those with same

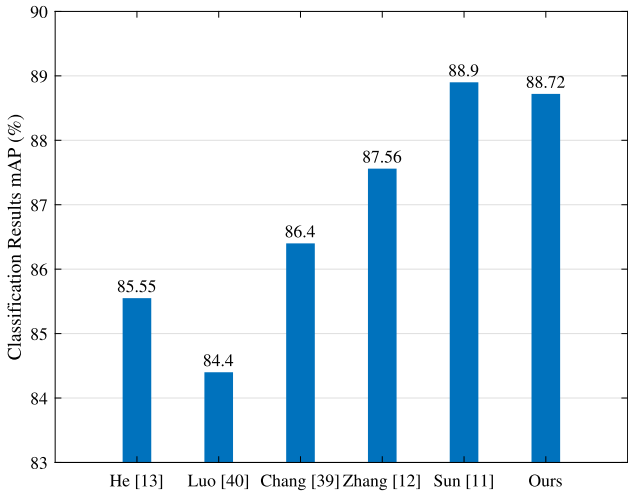

Fig. 7 Classification accuracy mAP (%) of GA-SRN and state-of-the-art fine-grained content-based image classification methods on CUB-200-2011 dataset

Table 2 Classification results mAP (%) on Con-Text and Drink-Bottle in four subsets which images contain scene text (I + T) and not contain scene text (I – T) are seperate

Methods	Con-text dataset		Drink-bottle dataset	
	I + T	I – T	I + T	I – T
Bai et al. [7]	78.92	71.63	71.61	62.25
Mafla et al. [4]	80.94	72.59	78.57	68.92
Mafla et al. [8]	86.76	74.31	82.75	69.16
Ours	87.52	75.28	83.52	70.26

Table 3 Classification results mAP (%) of ablation studies. V_G: global features, $V_G{}^*$: global features with self-attention, V_R: local features, V_T: textual features, V_B: positional features

Features	Con-text	Drink-bottle	CUB-200-2011
V_G	62.13	65.64	84.18
$V_G{}^*$	63.87	66.76	84.92
w/o GFRE and w/o M3JSR			
$V_G{}^* + V_R$	70.56	72.94	85.16
$V_G{}^* + V_R + V_T$	77.94	74.68	85.89
$V_G{}^* + V_R + V_T + V_B$	78.43	75.32	–
w/ GFRE and w/o M3JSR			
$V_G{}^* + V_R$	70.56	72.94	85.53
$V_G{}^* + V_R + V_T$	80.74	76.20	86.94
$V_G{}^* + V_R + V_T + V_B$	81.67	76.83	–
w/ GFRE and w/ M3JSR			
$V_G{}^* + V_R + V_T$	86.37	79.21	**88.72**
$V_G{}^* + V_R + V_T + V_B$	**86.57**	**79.96**	–

The bold contents represent the best results

features concatenated. Results for ablated models with different input combination: global features (V_G), global features with self-attention ($V_G{}^*$), local features (V_R), textual features (V_T), positional features (V_B) are also presented. Encoders w/o GFRE just concatenate two sources of input features (e.g. $V_G{}^* + V_R + V_T + V_B$) or directly deliver the only feature (e.g. $V_G{}^* + V_R$ and $V_G{}^* + V_R + V_T$) to downstream modules. It is noted when we conduct the ablation studies on CUB-200-2011, since each image in this dataset does not contain scene text, we only use its image description label as text, so we cannot perform ablation experiments on bounding box V_B, in Table 3, "–" refers to experiments unavailable. It could be observed that textual information from scene text (V_T) significantly improves classification results for all selections of GFRE and M3JSR. In addition, incorporating positional features to encoder provides slightly better results, such as 0.49% and 0.64% increase is achieved when ablating both GFRE and M3JSR. While all input features are selected, applying GFRE provides 3.24% and 1.51% better results on Con-Text and Drink-Bottles. Moreover, the integrated architecture in comparison with M3JSR ablated structure sees even further improvement from 81.67 to 86.57% on Con-Text and 76.83% to 79.96% on Drink-Bottles. The proposed GA-SRN module improves the accuracy by 1.05% and 1.78% on CUB-200-2011 when ablating GRRE and M3JSR, respectively. This shows that the proposed positional reasoning method with GAT would contribute more noticeably when images are relatively complex. These results indicate that the comprehensive combination of features engaged brings improvement to some extent. However, the relationship reasoning modules proposed are more determining in the overall level of classification results.

In Fig. 8, ablation study for GFRE of each encoder and M3JSR is conducted on three datasets to verify the effectiveness of semantic reasoning. It could be observed that applying GFRE in text encoder is slightly more effective than in local feature encoder since classification results on Con-Text falls to 86.19% and 85.90% for w/o visual GFRE and w/o textual GFRE, respectively. The self-attention mechanism in feature encoder also resulted in mild decline. However, M3JSR is still more crucial as mAP decreases to 81.67% and 86.94% on Con-Text and CUB-200-2011, respectively, more significant than any other ablation models of individual GFRE. In general, the relationship reasoning module GFRE is contributing to both visual and textual feature encoding, whereas applying cross-modal reasoning with M3JSR achieves overall optimized results.

Furthermore, we explore different structures of Graph Fusion Module in GFRE and Feature Fusion Module that fuses M3JSR module output with global feature encoder output $V_G{}^*$ as shown in Table 4. In specific, comparison Graph Fusion Modules are feature concatenation and the Graph Aggregation Module proposed by [24]. For Feature Fusion Module, performance of fusion methods MLB [30], Block [31] and feature concatenation are compared. It could be observed that concatenation architecture in Graph Fusion Module and Feature Fusion Module get the best results. This might be because that direct concatenation preserves more information than aggregation which is contributing to Relation Unit III.

(6) *Comparison on model complexity.* To evaluate the computational cost of this method, parameters, FLOPs and training time are compared with other models as shown in Table 5. With the inserted positional encoding branch in local and text feature encoding, the GFRE and M3JSR modules, this method has relatively higher model complexity (248.4 M parameters and 15.3 G FLOP). However,

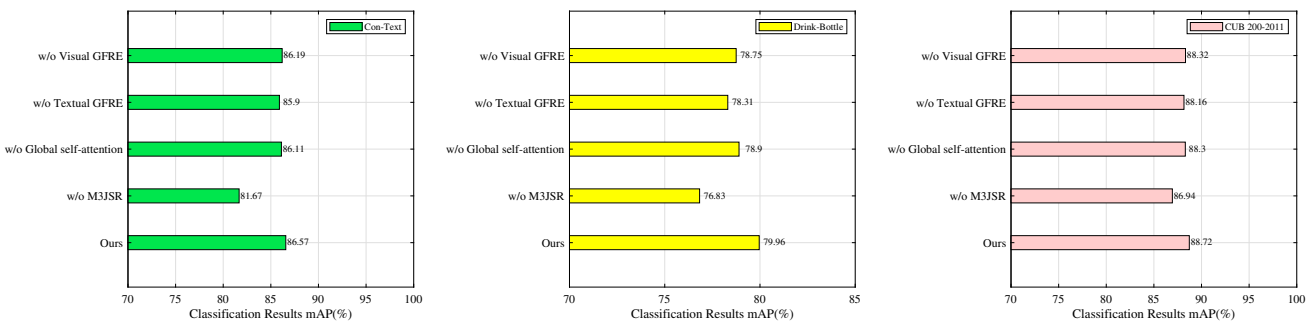

Fig. 8 Classification results mAP (%) of ablation studies on individual GFRE and M3JSR

Table 4 Classification results mAP (%) by using different strategies of concatenation in Graph Fusion Model and Feature Fusion Module of the proposed architecture

Graph fusion modules	Feature fusion modules	Con-text	Drink-bottle	CUB-200-2011
Aggregation [24]	MLB [30]	84.05	79.16	87.83
	Block [31]	84.49	78.78	87.62
	Concatenation	85.72	79.22	88.18
Concatenation	MLB [30]	84.92	78.78	87.85
	Block [31]	85.23	79.10	88.26
	Concatenation	**86.57**	**79.96**	**88.72**

The bold contents represent the best results

Table 5 Evaluation on parameters, FLOP and training time on both datasets

Methods	Param.	FLOP	Training time (min)	
	(M)	(G)	Con-text	Drink-bottle
Karaoglu et al. [14]	121.5	7.5	93	67
Bai et al. [7]	138.4	8.4	80	49
Mafla et al. [4]	164.8	11.7	72	54
Mafla et al. [8]	405.3	18.8	125	78
Ours	248.4	15.3	83	52

model complexity is still comparably lower than latest method by Mafla et al. [8] where the number of parameters and FLOP are 405.3 M and 18.8 G, respectively. This is because [8] contains both classical CNN and multi-layer GCN in which dimension of inner features is 2048. In comparison, the proposed model uses 1024 dimensional inner features and only uses one layer GAT suppressing parameter and FLOP to some extent. Bai et al. [7] achieves a good trade-off between computation and performance where classic GoogLeNet is used for image feature extraction. Although slightly larger storage space is needed, training time of the proposed method is relatively shorter among all image methods.

(7) *Limitations and discussions.* In fact, text encoding is relatively challenging compared to visual. Structural improvement of the proposed method that incorporates positional encoding branch of two encoders couldn't bring sufficient improvement to text encoder since basic classification error occurs even in encoding phase by FastText. Downstream GFRE for intra-modality reasoning couldn't make any correction to false embedding but only make further improvement to appropriate ones by positional information engaged. In this case, false textual information is also passed on to cross-modal reasoning throughout the classification. However, this seems to be a general problem for state-of-the-art methods with scene text encoder. Since it occurs at the beginning of the stage, future works that could alleviate this problem might experience a noticeable improvement on images with scene text.

4.4 Effect of GA-SRN on fine-grained image retrieval

We also evaluate the retrieval performance of this model on query image. In practice, the vector before classification representing image features are taken and Cosine similarity matrix is used to retrieve the semantically nearest images. Given an image from the dataset, the model would pick out images belong to the same class as input in the form of a ranked list. As shown in Fig. 9, retrieval results of the proposed method on two datasets are compared with other approaches who also testify retrieval. Due to extra difficulty of recognizing text instances appeared in images of Drink-Bottle, results for all methods on Drink-Bottle are generally lower than on Con-Text. Retrieval mAP of the proposed method arrives at 76.48% and 67.62% on Con-Text and Drink-Bottle datasets which surpass previous best model by 0.98% and 2.23%, respectively. Benefited from the semantic reasoning modules, the proposed network can obtain more discriminative features which eventually improve mAP in both classification and retrieval.

Figure 9 shows the retrieval performance comparison of our GA-SRN with other content-based image classification methods Sun et al. [11], Teh et al. [41] and Zeng et al. [42] on CUB-200-2011 dataset. The proposed method achieves relatively higher results than current methods, which could be explained as our better ability to discriminate subtle differences in recognizing categories. At the same time, the comparative experiments show that Sun et al. [11] shows comparatively the best results among all. This might be because of its specific focus on local discriminative region location and feature representation based on high-response

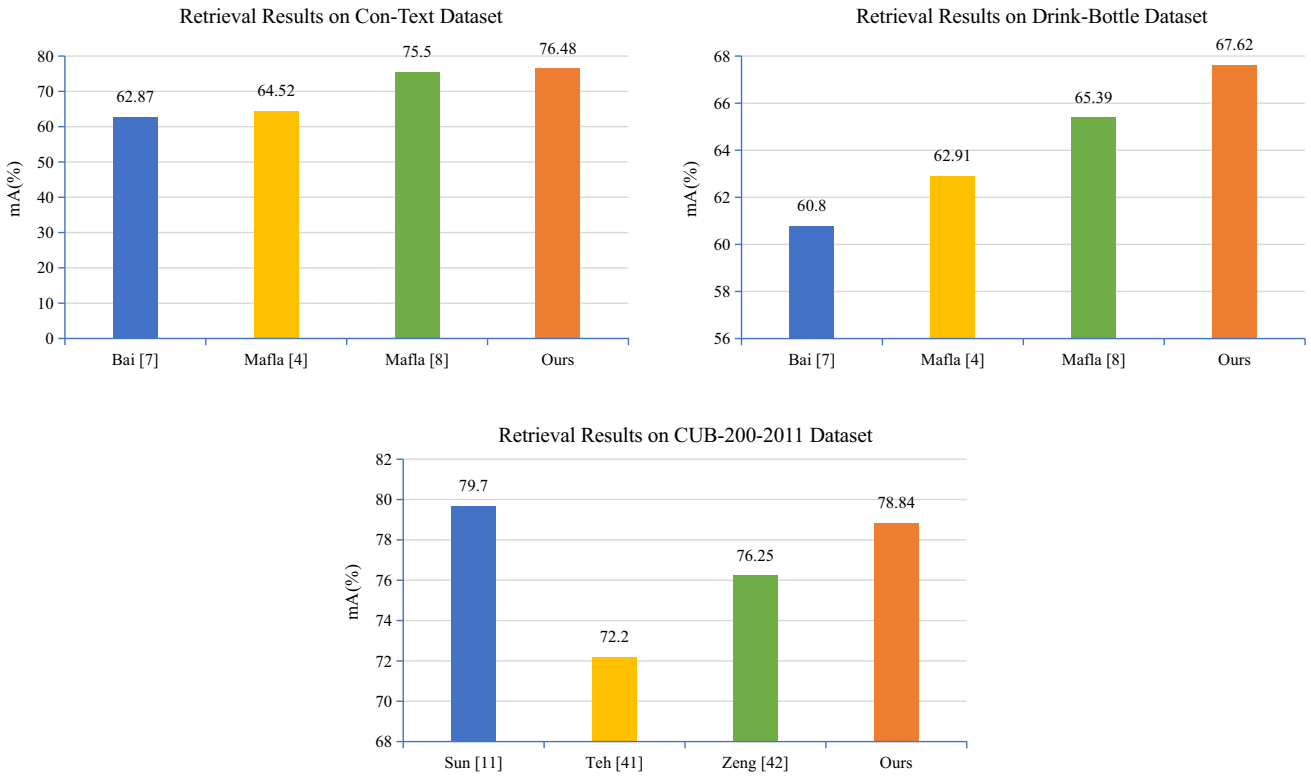

Fig. 9 Retrieval results mAP (%) of GA-SRN in comparison with Bai et al. [7], Mafla et al. [4], Mafla et al. [8] on Con-Text and Drink-Bottle datasets, and the results of comparison with Sun et al. [11], Teh et al. [41], Zeng et al. [42] on CUB-200-2011 dataset

region searching which is extremely contributing to content-based image retrieval.

5 Conclusions

In this paper, an end-to-end feature-relationship enhancement concerning network for FGIC is proposed. In specific, semantic relationship reasoning with positional features is firstly realized for both local and textual encoding by the proposed GFRE with efficient GAT. GFRE could provide feature representation and relation exploitation between any two input features. M3JSR is established by cross-modal GFREs with multi-head fusion for re-representing visual and text encoder outputs. This module enhances previous multi-modal reasoning with GCN to some extent. Ablation studies verify the effectiveness of generating more discriminatively represented features by the proposed semantic reasoning modules and rearranged positional encoding. Experimental results show that the proposed method achieves slightly more surpassing results on Con-Text, while mistakenly embedded text instances initially on certain Drink-Bottle images may hinder downstream structural advance from taking effect. Future studies might seek for more competent text embedding network in practice to deal with basic errors caused by recognition on varying scene text. Also, improvement could be further made on GAT-based semantic relationship enhancement structure.

Acknowledgements This work was supported by the National Nature Science Foundation of China under Grant 61872143.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Wenyong W, Yongcheng C, Guangshun L, Chuntao J, Song D (2020) A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition. Neural Comput Appl 32:14613–14622
2. Chang D, Ding Y, Xie J, Bhunia AK, Li X, Ma Z, Wu M, Guo J, Song YZ (2020) The devil is in the channels: mutual-channel loss for fine-grained image classification. IEEE Trans Image Process 29:4683–4695
3. Huang Z, Duan X, Zhao B, Lü J, Zhang B (2021) Interpretable attention guided network for fine-grained visual classification. arXiv preprint arXiv:2103.04701
4. Mafla A, Dey S, Biten AF, Gomez L, Karatzas D (2020) Fine-grained image classification and retrieval by combining visual and locally pooled textual features. In: Proceedings of the IEEE

winter conference on applications of computer vision, pp 2950–2959
5. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A (2016) Reading text in the wild with convolutional neural networks. Int J Comput Vis 116(1):1–20
6. Movshovitz-Attias Y, Yu Q, Stumpe MC, Shet V, Arnoud S, Yatziv L (2015) Ontological supervision for fine grained classification of street view storefronts. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1693–1702
7. Bai X, Yang M, Lyu P, Xu Y, Luo J (2018) Integrating scene text and visual appearance for fine-grained image classification. IEEE Access 6:66322–66335
8. Mafla A, Dey S, Biten AF, Gomez L, Karatzas D (2021) Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4023–4033
9. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv:1710.10903
10. Wei XS, Song YZ, Mac Aodha O, Wu J, Peng Y, Tang J, Yang J, Belongie S (2021) Fine-grained image analysis with deep learning: a survey. IEEE Trans Pattern Anal Mach Intell
11. Sun K, Zhu J (2022) Searching and learning discriminative regions for fine-grained image retrieval and classification. IEICE Trans Inf Syst 105(1):141–149
12. Zhang L, Huang S, Liu W (2021) Intra-class part swapping for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3209–3218
13. He X, Peng Y (2017) Fine-graind image classification via combining vision and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7332–7340
14. Karaoglu S, Tao R, Gevers T, Smeulders AW (2016) Words matter: scene text for image classification and retrieval. IEEE Trans Multimed 19(5):1063–1076
15. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497
16. Borisyuk F, Gordo A, Sivakumar V (2018) Rosetta: large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 71–79
17. He T, Tian Z, Huang W, Shen C, Qiao Y, Sun C (2018) An end-to-end textspotter with explicit alignment and attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5020–5029
18. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
19. Zhang Y, Nie S, Liu W, Xu X, Zhang D, Shen HT (2019) Sequence-to-sequence domain adaptation network for robust text image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2740–2749
20. Dey R, Salem FM (2017) Gate-variants of gated recurrent unit GRU neural networks. In: 2017 IEEE 60th international midwest symposium on circuits and systems, pp 1597–1600
21. Almazán J, Gordo A, Fornés A, Valveny E (2014) Word spotting and recognition with embedded attributes. IEEE Trans Pattern Anal Mach Intell 36(12):2552–2566
22. Gómez L, Mafla A, Rusinol M, Karatzas D (2018) Single shot scene text retrieval. In: Proceedings of the European conference on computer vision, pp 700–715
23. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2008) The graph neural network model. IEEE Trans Neural Netw 20(1):61–80
24. Gao D, Li K, Wang R, Shan S, Chen X (2020) Multi-modal graph neural network for joint reasoning on vision and scene text. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp 12746–12756
25. Li K, Zhang Y, Li K, Li Y, Fu Y (2019a) Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4654–4662
26. Li L, Gan Z, Cheng Y, Liu J (2019b) Relation-aware graph attention network for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10313–10322
27. Wen K, Gu X, Cheng Q (2020) Learning dual semantic relations with graph attention for image-text matching. IEEE Trans Circuits Syst Video Technol
28. Zeng J, Liu T, Jia W, Zhou J (2021) Fine-grained question-answer sentiment classification with hierarchical graph attention network. Neurocomputing 457:214–224
29. Chen S, Zhao Y, Jin Q, Wu Q (2020) Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 10638–10647
30. Kim JH, On KW, Lim W, Kim J, Ha JW, Zhang BT (2016) Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325
31. Ben-Younes H, Cadene R, Thome N, Cord M (2019) Block: bilinear superdiagonal fusion for visual question answering and visual relationship detection. Proc AAAI Conf Artif Intell 33:8102–8109
32. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6077–6086
33. Kazemi V, Elqursh A (2017) Show, ask, attend, and answer: a strong baseline for visual question answering. arXiv preprint arXiv:1704.03162
34. Zellers R, Bisk Y, Farhadi A, Choi Y (2019) From recognition to cognition: visual commonsense reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6720–6731
35. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123(1):32–73
36. Wang Y, Yang H, Qian X, Ma L, Lu J, Li B, Fan X (2019) Position focused attention network for image-text matching. arXiv preprint arXiv:1907.09748
37. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
38. Karaoglu S, van Gemert JC, Gevers T (2013) Con-text: Text detection using background connectivity for fine-grained object classification. In: Proceedings of the ACM international conference on multimedia, pp 757–760
39. Chang D, Ding Y, Xie J, Bhunia AK, Li X, Ma Z, Wu M, Guo J, Song YZ (2020) The devil is in the channels: mutual-channel loss for fine-grained image classification. IEEE Trans Image Process 29:4683–4695

40. Luo W, Zhang H, Li J, Wei XS (2020) Learning semantically enhanced feature for fine-grained image classification. IEEE Signal Process Lett 27:1545–1549
41. Teh EW, DeVries T, Taylor GW (2020) Proxynca++: revisiting and revitalizing proxy neighborhood component analysis. In: European conference on computer vision, pp 448–464
42. Zeng Z, Wang J, Chen B, Dai T, Xia ST (2021) Pyramid hybrid pooling quantization for efficient fine-grained image retrieval. arXiv preprint arXiv:2109.05206