Hindawi
Complexity
Volume 2021, Article ID 8041029, 9 pages
https://doi.org/10.1155/2021/8041029

Research Article

Knowledge Graph Representation Fusion Framework for Fine-Grained Object Recognition in Smart Cities

Yang He[1], **Ling Tian**[1,2], **Lizong Zhang**[1,2] **and Xi Zeng**[3]

[1]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan Province 611731, China*
[2]*Trusted Cloud Computing and Big Data Key Laboratory of Sichuan Province, Chengdu 610000, China*
[3]*China Electronics Technology Cyber Security Co. Ltd, Chengdu, China*

Correspondence should be addressed to Ling Tian; lingtian@uestc.edu.cn

Received 21 April 2021; Revised 25 June 2021; Accepted 4 July 2021; Published 14 July 2021

Academic Editor: Zhihan Lv

Autonomous object detection powered by cutting-edge artificial intelligent techniques has been an essential component for sustaining complex smart city systems. Fine-grained image classification focuses on recognizing subcategories of specific levels of images. As a result of the high similarity between images in the same category and the high dissimilarity in the same subcategories, it has always been a challenging problem in computer vision. Traditional approaches usually rely on exploring only the visual information in images. Therefore, this paper proposes a novel Knowledge Graph Representation Fusion (KGRF) framework to introduce prior knowledge into fine-grained image classification task. Specifically, the Graph Attention Network (GAT) is employed to learn the knowledge representation from the constructed knowledge graph modeling the categories-subcategories and subcategories-attributes associations. By introducing the Multimodal Compact Bilinear (MCB) module, the framework can fully integrate the knowledge representation and visual features for learning the high-level image features. Extensive experiments on the Caltech-UCSD Birds-200-2011 dataset verify the superiority of our proposed framework over several existing state-of-the-art methods.

1. Introduction

In recent years, with the development of artificial intelligence and Internet of Things technology [1–5], the concept and construction of smart city are also constantly breaking through. As an important part of smart city field [6, 7], object recognition based on computer vision has attracted much attention. Specifically, fine-grained image classification has been widely used in vehicle type recognition [8–10], goods recognition [11], content-based image retrieval [12], and other smart city applications [13–17]. In these applications, recognizing fine-grained images is still challenging, due to the high similarity between images in the same categories and the high dissimilarity in the same subcategories caused by different poses, behaviors, and so on as shown in Figure 1.

Note. In this paper, we use "category" to refer to the abstract concept of object type. For example, the category of a bird refers to its family or genus, such as "Albatross," "*Turdus*." "Subcategory" refers to the concept of fine-grained object type. For example, the subcategory of a bird is the species, such as "Sooty Albatross," "Rusty Blackbird."

Traditional approaches of fine-grained image classification usually rely on the low level visual cues to capture features for recognition. These methods mainly involve part-based models, and visual attention networks [18–20] first locate regions/parts of the object and capture the visual features on the detected locations to learn the ability to distinguish the nuances between different subcategories. However, these models require heavy annotations of object parts and are more difficult to collect than image labels. Visual attention networks [21, 22] try to learn discriminative representations with attention mechanisms. However, these works only focus on capturing visual features with a lot of labeled images.

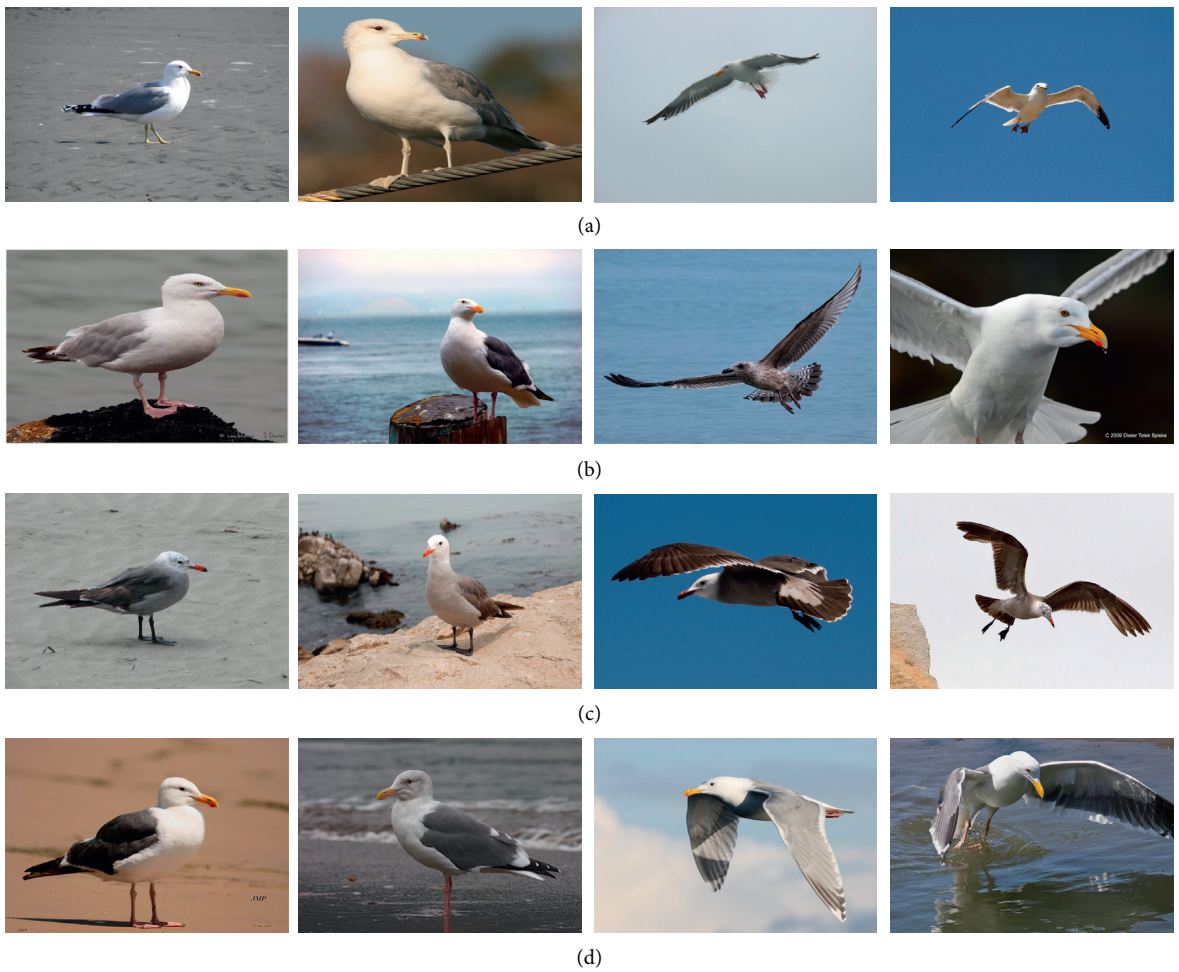

FIGURE 1: Samples from the Caltech-UCSD Birds-200-2011 dataset. The subcategories of birds in the four subfigures in a row are all the same, and the birds in all the subfigures belong to the same category: (a) California Gull; (b) Herring Gull; (c) Heermann's Gull; (d) Western Gull.

Different from the traditional approaches, human beings recognize objects in an image based on not only the visual information of objects, but also some prior knowledge acquired from their daily life experience. For example, we might know that yellow headed blackbird has yellow head and chest with black around its eyes. With this knowledge, when we see an image of "yellow headed blackbird," we can reason the classification correctly by combining visual information. Recently, Chen et al. [23] and Xu et al. [24] try to incorporate the prior knowledge as knowledge graph into fine-grained image classification. Although these methods achieve significant success, they still tend to consider the prior knowledge as the relations between subcategory labels and their attributes or introduce redundant text information as knowledge.

To take advantage of prior knowledge properly, this paper organizes prior knowledge about hierarchical relationships between categories-subcategories and subcategories-attributes relationships as knowledge graph, and designs a Knowledge Graph Representation Fusion (KGRF) framework for integrating knowledge representation and visual features for fine-grained image classification. The proposed model involves two key components: (1) the Graph Attention Network (GAT) [25] that aggregates information about nodes in the graph to learn the knowledge representation; (2) the Multimodal Compact Bilinear (MCB) [26] module that fuses knowledge representation with the captured visual features to learn the categories-subcategories and subcategories-attributes associations. Furthermore, the proposed method is validated on Caltech-UCSD Birds-200-2011 dataset [27] with 200 bird subcategories and 312 attributes. Compared with several baselines, the model shows superiority in fine-grained image classification task. In summary, the main contribution of this paper includes the following:

(1) We propose a novel KGRF framework for introducing knowledge graph in fine-grained image classification.

(2) Our model incorporates the MCB module for the first time to integrate the knowledge representation with visual features for fine-grained image classification.

(3) Extensive evaluation shows that our model outperforms several strong baselines in fine-grained image classification.

This paper is organized as follows. The related works are introduced in Section 2. The proposed model is described in Section 3. The model is validated by several experiments and compared with other methods in Section 4. The conclusion is presented in Section 5.

2. Related Work

During the past several years, there have been a number of researchers working on Convolutional Neural Networks (CNNs) [18, 28, 29] to capture the discriminative visual features for fine-grained image classification. Compared with the traditional handcrafted features [30–32], the method based on CNN shows a significant improvement. Bilinear CNN [33] uses two independent CNNs to learn the high-order representation which can capture interactions between subtle visual features, but the learned bilinear feature dimension is extremely high. In order to reduce the bilinear feature dimension, Gao et al. [34] proposed a compact model for approximating the high-dimensional feature with polynomial kernels. Kong et al. [35] proposed codecomposition to compress the Bilinear CNN model.

However, it is difficult for these approaches to capture the subtle visual features. Therefore, a series of studies [18, 19, 36] attempt to learn the part-based representation, which locates discriminative regions and captures the visual features. However, these methods rely on heavy manual part and bounding box annotations, making it difficult to apply them to the real world.

Instead, visual attention networks [37–41] are proposed to automatically locate the informative regions without part and bounding box annotations by the self-attention mechanism, and show superiority in the fine-grained image classification task [21, 42–44]. Liu et al. [21] use a reinforcement learning framework to adaptively locate the discriminative regions. Zheng et al. [44] propose a multi-attention CNN to capture parts localization, and aggregate features from the located informative regions with the global image. Fu et al. [42] introduce a recurrent attention CNN that recursively locates the attentional regions at multiple scales and learns the representation of region-based feature.

Although these methods avoid the need for a lot of part and bounding box annotations, they can only capture the feature from regions roughly due to the lack of supervision. Therefore, some researchers try to introduce additional guidance to capture more semantic-related features to aid the fine-grained image classification task. For example, Liu et al. [45] incorporate part-level attribute with locating the discriminative regions. He and Peng [46] introduce detailed language descriptions to capture more discriminative parts and features. Chen et al. [23] utilize the knowledge graph to introduce subcategory-attribute relations for reasoning discriminative features. Xu et al. [24] use the visual-semantic embedding framework to introduce the text and knowledge base to learn the relations between subcategories and images.

Different from existing approaches, our method also introduces additional guidance for fine-grained image classification, but in the form of constructing knowledge graph involved in subcategories-attributes relationships and categories-subcategories hierarchical relationships. There have been several works to introduce prior knowledge into visual tasks. For example, Qi et al. [47] propose 3DGNN network for semantic segmentation. Marino et al. [48] use GSNN for multilabel image recognition.

3. Method

In this section, our constructed knowledge graph which contains associations between categories and part-attributes as well as hierarchy of categories has been first proposed. Then, we describe the KGRF framework which contains the knowledge representation module using GAT for modeling the constructed knowledge graph and the knowledge fusion module using MCB module for integrating knowledge representation into captured visual features. An overview of the framework is shown in Figure 2. The detailed descriptions of notations could be found in Table 1.

3.1. Knowledge Graph Construction. Essentially, a knowledge graph which consists of nodes and edges is a repository of semantic information about complex structures in our lives. In order to better utilize the knowledge graph in fine-grained image classification, we introduce a knowledge graph that includes categories-subcategories hierarchy relationships and subcategories-attributes relationships. The knowledge graph is constructed based on the subcategory labels, the part-attribute annotations of images, and the existing knowledge base DBpedia [49].

Nodes. A node in the constructed knowledge graph refers to a specific category, subcategory, or a part-attribute. The category is a coarse-grained type of an object in an image, such as "Albatross," "Blackbird." The subcategory refers to a specific type that needs to be identified in the fine-grained image classification task, such as "Sooty Albatross," "Rusty Blackbird." The part-attribute is a description of parts of an object such as color, shape, size. Suppose that there are C object categories, S object subcategories, and A part-attributes; the knowledge graph has $C + S + A$ nodes.

Edges. There are two main types of edges in the constructed knowledge graph. The edge between a category node and a subcategory node means that there is some kind of hierarchical semantic relationships, such as "Sooty Albatross is a kind of Albatross." The edge between a subcategory node and a part-attribute node indicates that the subcategory has the corresponding part-attribute, such as "Sooty Albatross has the corresponding part-attribute 'has shape: swallow like." It should be noted that there is no association between two category nodes, two subcategory nodes, or two part-attribute nodes in the constructed knowledge graph.

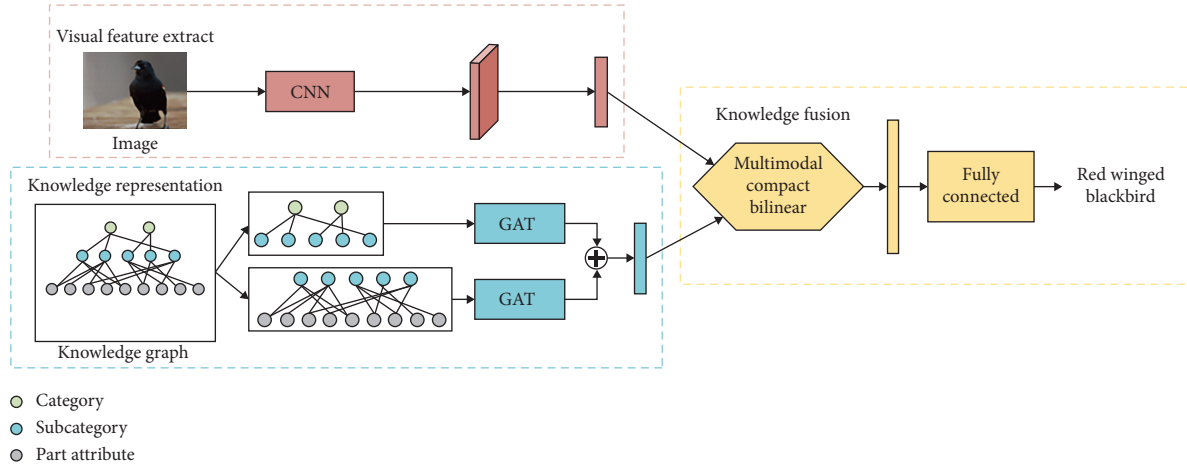

Figure 2: Overview of our KGFR framework. It consists of a visual feature extraction module, a knowledge expression module, and a knowledge fusion module. In visual feature extraction module, the CNN can extract the overall visual features in images. In the knowledge representation module, the constructed knowledge graph is split into two subgraphs, and the representations of the nodes are obtained by two GATs, respectively. Finally the knowledge representations of the same nodes in different GATs are concatenated to obtain the final knowledge representation. In the knowledge fusion module, the MCB module is introduced to fuse the knowledge representations with the extracted visual features, so as to enhance the fine-grained image classification.

Table 1: Notations used in this paper.

Notations	Descriptions
$\mathbb{R}^F$	F-dimensional Euclidean space
v^I	Visual feature of image
s^g	Representation of the knowledge graph
$\mathbf{f}$	High-order features that combine knowledge and vision
Ψ	Count Sketch projection function
FFT	Fast Fourier transform
$\oplus$	Concatenation operation
$\otimes$	Outer product

3.2. Knowledge Representation. After constructing the knowledge graph, since there are two types of edges, we extract two subgraphs from the constructed knowledge graph and use the GAT to learn the feature vector for nodes, respectively. Then, we concatenate the features of the same node learned from the two subgraphs to obtain the final representation of each node in the constructed knowledge graph.

According to the types of edges, the extracted two subgraphs are categories-subcategories subgraph and subcategories-attributes subgraph, respectively. The categories-subcategories subgraph only contains the category nodes, the subcategory nodes, and the edges between two kinds of nodes. The subcategories-attributes subgraph only contains the subcategory nodes, the part-attribute nodes, and the edges between two kinds of nodes. For the two subgraphs, we input them into two separate GATs to learn the representation of knowledge after node information propagations.

Because the representation learning methods of the two subgraphs are completely consistent, the categories-subcategories subgraph is taken as an example. We first initialize nodes in the subgraph with the corresponding Word2Vec [50] features, which can reflect the linguistic contexts of concepts. Accordingly, the input to GAT can be expressed as follows:

$$\mathbf{h} = \left\{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_{C+S}\right\}, \quad \vec{h}_i \in \mathbb{R}^F, \tag{1}$$

where $C + S$ means the number of nodes in the categories-subcategories subgraph and F means the initial feature dimension number in each node. After the propagation of node information, we can get a new set of node representations:

$$\mathbf{h}' = \left\{\vec{h}'_1, \vec{h}'_2, \ldots, \vec{h}'_{C+S}\right\}, \quad \vec{h}'_i \in \mathbb{R}^{F'}. \tag{2}$$

GAT is a convolution style neural network that uses masked self-attention for aggregating information about neighbor nodes. In order to transform the initial input features into high-order knowledge features, it first uses a weight matrix $\mathbf{W} \in \mathbb{R}^{F' \times F}$ for each node for linear transformation. Then, it utilizes the self-attention to compute attention coefficients:

$$e_{ij} = a\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j\right), \tag{3}$$

which represents the importance of the features of node j to i. In order to facilitate the consideration of coefficients between different nodes, the softmax function is used to normalize all neighborhood nodes of j:

$$\alpha_{ij} = \text{soft}\max_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(e_{ik}\right)}, \tag{4}$$

where $\mathcal{N}_i$ is the first-order neighbors of node i in the subgraph. Since the attention mechanism a is a feedforward neural network in GAT, it can be represented by a weight

vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F'}$. Thus, attention coefficients can be computed with the LeakyReLU activation function:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}_i \oplus \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}_i \oplus \mathbf{W}\vec{h}_k\right]\right)\right)}, \quad (5)$$

where $\vec{\mathbf{a}}^T$ means transposition operation on $\vec{\mathbf{a}}$ and $\oplus$ means the concatenation operation. On this basis, the obtained attention coefficients are used to compute the linear combination of their corresponding features:

$$\vec{h}'_i = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}\vec{h}_j\right). \quad (6)$$

In order to stabilize the learning process of self-attention and get the accurate features of each node as output, we use the multihead attention GAT:

$$\vec{h}'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k\mathbf{W}^k\vec{h}_j\right), \quad (7)$$

where α_{ij}^k means the normalized attention coefficient computed by the k-th attention mechanism and $\mathbf{W}^k$ means the k-th input linear transformation's weight matrix.

Finally, we get different node features from the two GATs. In order to combine the categories-subcategories relationships and subcategories-attributes relationships, we concatenate the features learned by the same subcategory node in different GAT to get the final knowledge representation $\mathbf{s}^g$ for each subcategory node:

$$\mathbf{s}_i^g = \vec{h}'_{i1} \oplus \vec{h}'_{i2}, \quad (8)$$

where $\vec{h}'_{i1}$ is the knowledge feature learned in the categories-subcategories subgraph and $\vec{h}'_{i2}$ is the knowledge feature learned in the subcategories-attributes subgraph. Thus, the knowledge representation of the constructed knowledge graph can be expressed as follows:

$$\mathbf{s}^g = \{s_1^g, s_2^g, \ldots, s_S^g\}. \quad (9)$$

3.3. Knowledge Fusion. After the knowledge representation is obtained, the MCB module is introduced to fuse it with the extracted visual features, so as to enhance the fine-grained image classification.

Visual Feature Extraction. Since there are several CNN models with good performance in fine-grained image classification, we directly choose the CB-CNN [34] to extract image visual features $\mathbf{v}^I$:

$$\mathbf{v}^I = \text{CNN}(\mathbf{I}). \quad (10)$$

Firstly, the input image is processed through a convolutional network to obtain feature maps with the size of $W' \times H' \times d$:

$$W_{\text{out}} = \frac{W_{\text{in}} - W_{\text{filter}} + 2P}{S} + 1,$$
$$H_{\text{out}} = \frac{H_{\text{in}} - H_{\text{filter}} + 2P}{S} + 1, \quad (11)$$

where W_{filter} and H_{filter} mean the width and height of the filter in CNN model, P is the padding, and S is the stride.

Then, the compact bilinear operation is performed to capture the final feature maps $\mathbf{v}^I$. Since we will fuse the knowledge representation with visual features, the final feature maps are not sum-pooled. For comparison with the existing works, we adopt VGG16-Net as the CNN model.

Most models only consider visual features in images but ignore a lot of implicit semantic correlation information. In the fine-grained image classification task, visual features alone make it difficult to capture subtle differences between subcategories. However, the obtained knowledge representation contains the categories-subcategories relations and subcategories-attributes relations, which might help capture important subtle features. Thus, we fuse the knowledge representation with visual features to learn better unified feature representation. Since the traditional method only performs simple concatenation operation on two different features, without considering the interaction between them, we introduce the MCB module to efficiently and expressively integrate knowledge representation and visual features. Traditional bilinear models take the outer product of two vectors, $c_1 \in \mathbb{R}^{d_1}, c_2 \in \mathbb{R}^{d_2}$, and learn a linear model $\mathbf{W}$ to allow all elements of two vectors to interact with each other:

$$\mathbf{y} = \mathbf{W}[c_1 \otimes c_2], \quad (12)$$

where $\otimes$ means the outer product and [] denotes linearizing the matrix in a vector. The direct calculation of the outer product leads to a large amount of memory consumption and high calculation time. Thus, MCB module uses the convolution of two count sketches to express the outer product of vectors:

$$\Psi(c_1 \otimes c_2) = \Psi(c_1, h, s)^* \Psi(c_2, h, s), \quad (13)$$

where $*$ means the convolution operator and Ψ is the Count Sketch projection function [51]. Additionally, according to the convolution theorem, convolution in the time domain is equivalent to element-wise product in the frequency domain. Therefore, the MCB module can be summarized in Figure 3 and described as Algorithm 1.

Suppose that $\mathbf{s}_i^g$ is the knowledge representation of a subcategory node i in the knowledge graph and $\mathbf{v}_j^I$ is the visual feature of an image j. The purpose of our MCB module is to get the comprehensive feature representation $\mathbf{f}_j$ for image j:

$$\mathbf{f}_j = \text{MCB}\left(\mathbf{s}_i^g, \mathbf{v}_j^i\right). \quad (14)$$

Firstly, the Tensor Sketching [52] is used to compress the dimensions of $\mathbf{s}_i^g$ and $\mathbf{v}_j^I$, respectively.

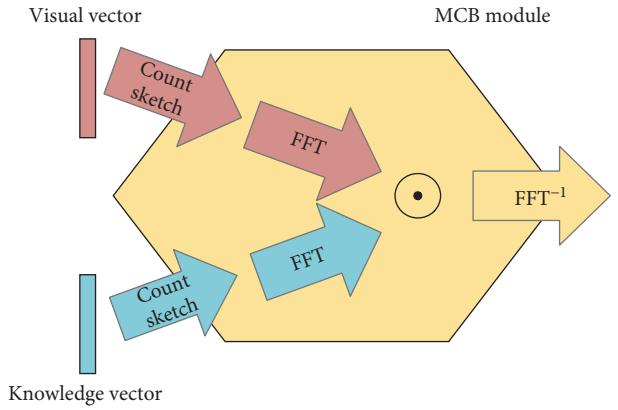

Figure 3: The structure of MCB module. The input of MCB module consists of the visual vector and knowledge vector. After using the Count Sketch projection function to reduce the dimension of them, the two kinds of vectors are fused in the Fourier space and finally output the high-order vector.

$$\begin{aligned} \mathbf{s}_i^{g'} &= \Psi\left(\mathbf{s}_i^{g}\right), \\ \mathbf{v}_j^{I'} &= \Psi\left(\mathbf{v}_j^{I}\right), \end{aligned} \tag{15}$$

where Ψ is the Count Sketch projection function and $\mathbf{s}_i^{g'}$ and $\mathbf{v}_j^{I}$ are the features after dimensionality reduction. Then, in the Fast Fourier Transform (FFT) space, the two features are fused by element-wise product.

$$\mathbf{f}_j = \mathrm{FFT}^{-1}\left(\mathrm{FFT}\left(\mathbf{s}_i^{g'}\right) \odot \mathrm{FFT}\left(\mathbf{v}_j^{I'}\right)\right), \tag{16}$$

where FFT means the Fast Fourier Transform, FFT^{-1} is the Inverse Fast Fourier Transform, and $\mathbf{f}_j$ is the high-order feature that combines knowledge representation with visual features. Finally, we feed the feature $\mathbf{f}_j$ into a full connection layer to classify the subcategory of the image j.

4. Experiment

4.1. Experiment Settings

Datasets. We evaluate the proposed KGRF framework on the Caltech-UCSD Birds-200-2011 [27] dataset, which is a widely used benchmark in fine-grained image classification. There are 200 subcategories of birds, 5,994 training images, and 5,794 test images in the dataset. In addition to the basic subcategory annotations, each image is further labeled with 1 bounding box, 15 part key-points, and 312 part-attributes. In this work, we use accuracy to evaluate the effectiveness of models on fine-grained image classification.

Knowledge Graph Details. Since the constructed knowledge graph contains categories-subcategories hierarchy relations and subcategories-attributes relations, there are three types of nodes, including category node, subcategory node, and part-attribute node. Specifically, the category node obtained from DBpedia refers to a type of coarse-grained bird species. The subcategory node represents the output in the fine-grained image classification task, according to the image label in the Caltech-UCSD Birds-200-2011 dataset. The part-attribute node means a type of attribute in a particular part of birds, according to the attribute annotation. In a word, there are 56 category nodes, 200 subcategory nodes, and 312 part-attribute nodes in the knowledge graph.

(1) input: $c_1 \in \mathbb{R}^{d_1}, c_2 \in \mathbb{R}^{d_2}$
(2) output: $\phi(c_1, c_2) \in \mathbb{R}^{d}$
(3) procedure MCB (c_1, c_2, d_1, d_2, d)
(4) for k in {1, 2} do
(5) if h_k, s_k not initialized then
(6) for i in $\{1, \ldots, n_k\}$ do
(7) sample $h_k[i]$ from $\{1, \ldots, d\}$
(8) sample $s_k[i]$ from $\{-1, 1\}$
(9) $c_k' = \psi(c_1, c_2, d_1, d_2)$
(10) $\phi = \mathrm{FFT}^{-1}(\mathrm{FFT}(\mathbf{c}_1') \odot \mathrm{FFT}(\mathbf{c}_2'))$
(11) **return** ϕ
(12) procedure $\psi(c, s, h, d)$
(13) $y = [0, \ldots, 0]$
(14) for i in $\{1, \ldots, d\}$ do
(15) $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$
(16) **return** y

Algorithm 1: Multimodal Compact Bilinear module.

4.2. Comparison with State-of-the-Art Methods. In this subsection, we compare the proposed model with several state-of-the-art methods, and the results are reported in Table 2. Some of the methods rely solely on the image labels, while others use the information of bounding box and part key-points. In the models that use bounding box annotations and part annotations, the part-based model PN-CNN [28] performs well with the accuracy of 85.4%, but this type of approach relies heavily on the guidance of bounding box annotations and part annotations. On the contrary, more existing methods do not rely on the bounding box annotations and part annotations. Instead, they attempt to search the distinguishing regions and capture the high-level visual features of these regions for classification. For example, Bilinear CNN [33] uses two separated CNNs to locate the key parts and capture the visual features of parts, and the accuracy achieved is 84.1%,

Table 2: Experimental results on Caltech-UCSD Bird dataset.

Method	Bounding box annotations	Part annotations	Accuracy (%)
Part-RCNN [18]	✓	✓	76.4
SPDA-CNN [36]	✓	✓	85.14
PN-CNN [28]	✓	✓	85.4
CB-CNN [34]	✓		84.6
FCAN [21]	✓		84.7
Bilinear CNN [33]			84.1
ST-CNN [43]			84.1
CVL [46]			85.55
A3M [53]			86.2
PC-DenseNet-161 [54]			86.87
KERL [23]			87.0
SEF [55]			87.3
Ensemble T-CNN [24]			87.3
TASN [56]			87.9
HSE [57]			88.1
GCL [58]			88.3
Ours			88.49

but it relies on very-high-dimensional representations of visual features. A3M [53] utilizes attribute-guided attention module to consider attribute information to select key features for different regions.

In addition, we compare the proposed model with the methods which introduce external information. Specifically, CVL [46] introduces the prior text descriptions to help locate the discriminative regions, and it achieves 85.55% accuracy. HSE [57] introduces four levels of category hierarchical semantic information to improve the accuracy to 88.1%. Further, there are also some methods trying to introduce the prior external knowledge, which are most related to our work. KERL [23] introduces a knowledge graph with a Gated Graph Neural Network (GGNN) [59], which models the correlations between categories and part-attributes. Ensemble T-CNN [24] introduces the text descriptions and knowledge base with visual-semantic embedding [60] at the same time. By contrast, our framework achieves very good results, especially with the two methods that also introduce the prior knowledge.

4.3. Contribution of Knowledge Representation. Since our KGRF framework is based on the CB-CNN [34] to extract visual features and fully integrates the obtained knowledge representation, we set up experiments to demonstrate the effectiveness of the knowledge representation. As shown in Table 3, CB-CNN relies heavily on visual features and only achieves an accuracy of 84.6%. Using this model as a benchmark, our KGRF framework uses the MCB module for introducing the knowledge representation to improve the accuracy to 88.49%, which is 3.89% higher than CB-CNN that only uses the visual features. This means that the introduction of prior knowledge into fine-grained image classification performs well. To further verify the contribution of our method of introducing prior knowledge, we use element-wise product and simple concatenation, respectively, to integrate knowledge representation with visual features for comparison with our framework. As shown in Table 3, the element-wise product and concatenation can achieve accuracies of 85.8% and 86.5%, respectively, which is a little better than the CB-CNN but still a lot worse than ours. This shows that the knowledge fusion method we adopted can make better use of the prior knowledge to promote fine-grained image classification.

Table 3: Accuracy comparisons with different knowledge fusion methods.

Method	Accuracy (%)
CB-CNN [34]	84.6
Element-wise product	85.8
Concatenation	86.5
Ours	**88.49**

5. Conclusion

In this paper, we propose a novel Knowledge Graph Representation Fusion (KGRF) framework to integrate knowledge representations and visual features for fine-grained image classification. In particular, the proposed framework includes the GAT to learn knowledge representations, and the MCB module to fuse the representations with the captured visual features from CNN for modeling the categories-subcategories and subcategories-attributes associations. Furthermore, the proposed framework is validated on a widely used dataset, Caltech-UCSD Birds-200-2011, and the experiment results show superiority in fine-grained image classification task, compared with several state-of-the-art methods. In the future, we could consider more reasonable and interpretable methods to introduce prior knowledge into relevant computer vision tasks.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work was partly supported by Sichuan Science and Technology Program (2019YFG0507 and 2020YFG0328), the National Natural Science Foundation of China (NSFC) (U19A2059), and Young Scientists Fund of the National Natural Science Foundation of China (61802050).

References

[1] Z. Cai and Z. He, "Trading private range counting over big IoT data," in *Proceedings of the IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 144–153, IEEE, Dallas, TX, USA, July 2019.

[2] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial IoTs," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.

[3] X. Zheng, Z. Cai, and Y. Li, "Data linkage in smart internet of things systems: a consideration from a privacy perspective," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 55–61, 2018.

[4] Q. Hou, M. Han, and Z. Cai, "Survey on data analysis in social media: a practical application aspect," *Big Data Mining and Analytics*, vol. 3, no. 4, pp. 259–279, 2020.

[5] J. Li, M. Siddula, X. Cheng, W. Cheng, Z. Tian, and Y. Li, "Approximate data aggregation in sensor equipped IoT networks," *Tsinghua Science and Technology*, vol. 25, no. 1, pp. 44–55, 2020.

[6] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, pp. 766–775, 2018.

[7] Z. Cai, X. Zheng, and J. Yu, "A differential-private framework for urban traffic flows estimation via taxi companies," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6492–6499, 2019.

[8] J. Špaňhel, J. Sochor, and A. Makarov, "Vehicle fine-grained recognition based on convolutional neural networks for real-world applications," in *Proceedings of 2018 14th Symposium on Neural Networks and Applications (NEUREL)*, pp. 1–5, IEEE, Belgrade, Serbia, November 2018.

[9] Z. Chen, C. Ying, C. Lin, S. Liu, and W. Li, "Multi-view vehicle type recognition with feedback-enhancement multi-branch CNNs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 2590–2599, 2017.

[10] B. Hu, J.-H. Lai, and C.-C. Guo, "Location-aware fine-grained vehicle type recognition using multi-task deep networks," *Neurocomputing*, vol. 243, pp. 60–68, 2017.

[11] Q. Hong, H. Zhang, P. Nie, and C. Zhang, "The recognition method of express logistics restricted goods based on deep convolution neural network," in *Proceedings of 2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, pp. 363–367, IEEE, Xiamen, China, May 2020.

[12] C. C. Hung, *A Study on a Content-Based Image Retrieval Technique for Chinese Paintings*, The Electronic Library, 2018.

[13] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, *Generative Adversarial Networks: A Survey towards Private and Secure Applications*, ACM Computing Surveys (CSUR), vol. 37, no. 4, ACM Computing Surveys (CSUR), New York, NY, USA, 2021.

[14] M. Sreenivasulu and M. Sridevi, "Comparative study of statistical features to detect the target event during disaster," *Big Data Mining and Analytics*, vol. 3, no. 2, pp. 121–130, 2020.

[15] J. Pang, Y. Huang, Z. Xie, Q. Han, and Z. Cai, "Realizing the heterogeneity: a self-organized federated learning framework for IoT," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3088–3098, 2021.

[16] J. Pang, Y. Huang, Z. Xie, J. Li, and Z. Cai, "Collaborative city digital twin for the COVID-19 pandemic: a federated learning solution," *Tsinghua Science and Technology*, vol. 26, no. 5, pp. 759–771, 2021.

[17] J. Chen, J. Li, and Y. Li, "Predicting human mobility via long short-term patterns," *Computer Modeling in Engineering & Sciences*, vol. 124, no. 3, pp. 847–864, 2020.

[18] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proceedings of European Conference on Computer Vision*, pp. 834–849, Springer, Zurich, Switzerland, September 2014.

[19] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1173–1182, Las Vegas, NV, USA, June 2016.

[20] Y. Liang, Z. Cai, J. Yu, Q. Han, and Y. Li, "Deep learning based inference of private information using embedded sensors in smart devices," *IEEE Network*, vol. 32, no. 4, pp. 8–14, 2018.

[21] X. Liu, T. Xia, J. Wang, and Y. Lin, "Fully convolutional attention localization networks," arXiv: 1603.06765, 2016.

[22] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, pp. 1487–1500, 2017.

[23] T. Chen, L. Lin, R. Chen, Y. Wu, and X. Luo, *Knowledge-embedded Representation Learning for Fine-Grained Image Recognition*, arXiv:1807.00505, 2018.

[24] H. Xu, G. Qi, J. Li, M. Wang, K. Xu, and H. Gao, *Fine-grained Image Classification by Visual-Semantic Embedding*, in *Proceedings of IJCAI*, pp. 1043–1049, Stockholm, Swedan, July 2018.

[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," arXiv: 1710.10903, 2017.

[26] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," arXiv: 1606.01847, 2016.

[27] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-Ucsd birds-200-2011 dataset," *Technical Report CNS-TR-2011-001*, California Institute of Technology, Pasadena, CA, USA, 2011.

[28] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," arXiv:1406.2952, 2014.

[29] R. Xin, J. Zhang, and Y. Shao, "Complex network classification with convolutional neural network," *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 447–457, 2020.

[30] L. Bourdev, S. Maji, and J. Malik, "Describing people: a poselet-based approach to attribute classification," in *Proceedings of 2011 International Conference on Computer Vision*, pp. 1543–1550, IEEE, Barcelona, Spain, November 2011.

[31] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis, "Birdlets: subordinate categorization using volumetric primitives and pose-normalized appearance," in *Proceedings of 2011 International Conference on Computer Vision*, pp. 161–168, IEEE, Barcelona, Spain, November 2011.

[32] N. Zhang, R. Farrell, and T. Darrell, "Pose pooling kernels for sub-category recognition," in *Proceedings of 2012 IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 3665–3672, IEEE, Providence, RI, USA, June 2012.

[33] T. Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1449–1457, Santiago, Chile, December 2015.

[34] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 317–326, Las Vegas, NV, USA, June 2016.

[35] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 365–374, Honolulu, HI, USA, July 2017.

[36] H. Zhang, T. Xu, M. Elhoseiny et al., "SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1143–1152, Las Vegas, NV, USA, June 2016.

[37] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[38] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, "Crowd counting using deep recurrent spatial-aware network," arXiv: 1807.00601, 2018.

[39] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," arXiv:1406.6247, 2014.

[40] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 464–472, Venice, Italy, October 2017.

[41] B. Zhao, P. Zhao, and P. Fan, "ePUF: a lightweight double identity verification in IoT," *Tsinghua Science and Technology*, vol. 25, no. 5, pp. 625–635, 2020.

[42] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4438–4446, Honolulu, HI, USA, July 2017.

[43] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," arXiv: 1506.02025, 2015.

[44] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5209–5217, Honolulu, HI, USA, July 2017.

[45] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin, "Localizing by describing: attribute-guided attention localization for fine-grained recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[46] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5994–6002, Honolulu, HI, USA, July 2017.

[47] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for RGBD semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5199–5208, Honolulu, HI, USA, July 2017.

[48] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: using knowledge graphs for image classification," arXiv:1612.04844, 2016.

[49] J. Lehmann, R. Isele, M. Jakob et al., "DBpedia-a large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[50] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," arXiv:1310.4546, 2013.

[51] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *Proceedings of International Colloquium on Automata, Languages, and Programming*, pp. 693–703, Springer, Malaga, Spain, July 2002.

[52] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 239–247, Chicago, IL, USA, August 2013.

[53] K. Han, J. Guo, C. Zhang, and M. Zhu, "Attribute-aware attention model for fine-grained representation learning," in *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 2040–2048, Seoul, Republic of Korea, October 2018.

[54] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 70–86, Munich, Germany, September 2018.

[55] W. Luo, H. Zhang, J. Li, and X.-S. Wei, "Learning semantically enhanced feature for fine-grained image classification," *IEEE Signal Processing Letters*, vol. 27, pp. 1545–1549, 2020.

[56] H. Zheng, J. Fu, Z. J. Zha, and J. Luo, "Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5012–5021, Long Beach, CF, USA, June 2019.

[57] T. Chen, W. Wu, Y. Gao, L. Dong, X. Luo, and L. Lin, "Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding," in *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 2023–2031, Seoul, Republic of Korea, October 2018.

[58] Z. Wang, S. Wang, H. Li, Z. Dou, and J. Li, "Graph-propagation based correlation learning for weakly supervised fine-grained image classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12289–12296, 2020.

[59] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," arXiv:1511.05493, 2015.

[60] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, Boston, MA, USA, June 2015.