Knowledge-Embedded Representation Learning for Fine-Grained Image Recognition

Tianshui Chen[1], **Liang Lin**[1,2*], **Riquan Chen**[1], **Yang Wu**[1] and **Xiaonan Luo**[3]
[1] Sun Yat-sen University, China
[2] SenseTime Research, China
[3] Guilin University of Electronic Technology, China
tianshuichen@gmail.com, linliang@ieee.org,
sysucrq@gmail.com, wuyoung567@gmail.com, luoxn@guet.edu.cn

Abstract

Humans can naturally understand an image in depth with the aid of rich knowledge accumulated from daily lives or professions. For example, to achieve fine-grained image recognition (e.g., categorizing hundreds of subordinate categories of birds) usually requires a comprehensive visual concept organization including category labels and part-level attributes. In this work, we investigate how to unify rich professional knowledge with deep neural network architectures and propose a Knowledge-Embedded Representation Learning (KERL) framework for handling the problem of fine-grained image recognition. Specifically, we organize the rich visual concepts in the form of knowledge graph and employ a Gated Graph Neural Network to propagate node message through the graph for generating the knowledge representation. By introducing a novel gated mechanism, our KERL framework incorporates this knowledge representation into the discriminative image feature learning, i.e., implicitly associating the specific attributes with the feature maps. Compared with existing methods of fine-grained image classification, our KERL framework has several appealing properties: i) The embedded high-level knowledge enhances the feature representation, thus facilitating distinguishing the subtle differences among subordinate categories. ii) Our framework can learn feature maps with a meaningful configuration that the highlighted regions finely accord with the nodes (specific attributes) of the knowledge graph. Extensive experiments on the widely used Caltech-UCSD bird dataset demonstrate the superiority of our KERL framework over existing state-of-the-art methods.

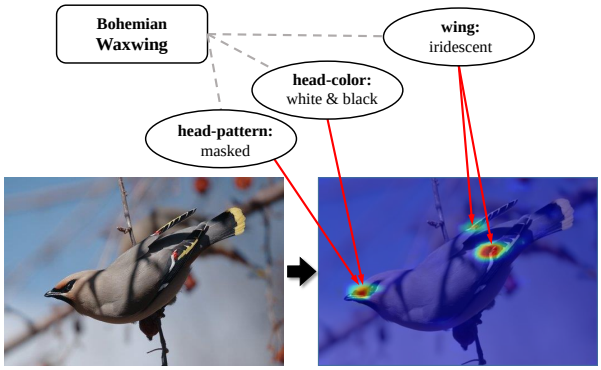

Figure 1: An example of how the professional knowledge aids fine-grained image understanding. Our proposed framework is capable of associating the specific attributes with the image feature representation.

1 Introduction

Humans perform object recognition task based on not only the object appearance but also the knowledge acquired from daily lives or professions. Usually, this knowledge refers to a comprehensive visual concept organization including category labels and their attributes. It is extremely beneficial to fine-grained image classification as attributes are always key to distinguish different subordinate categories. For example, we might know from a book that a bird of category "Bohemian Waxwing" has a masked head with color black and white and wings with iridescent feathers. With this knowledge, to recognize the category "Bohemian Waxwing" given a bird image, we might first recall the knowledge, attend to the corresponding parts to see whether it possesses these attributes, and then perform reasoning. Figure 1 illustrates an example of how the professional knowledge aids fine-grained image recognition.

Conventional approaches for fine-grained image classification usually neglect this knowledge and merely rely on low-level image cues for recognition. These approaches either

*Corresponding author is Liang Lin (Email: linliang@ieee.org). This work was supported by the National Natural Science Foundation of China under Grant 61622214, the Science and Technology Planning Project of Guangdong Province under Grant 2017B010116001, and Guangdong Natural Science Foundation Project for Research Teams under Grant 2017A030312006.

employ part-based models [Zhang *et al.*, 2014] or resort to visual attention networks [Liu *et al.*, 2016] to locate discriminative regions/parts to distinguish subtle differences among different subordinate categories. However, part-based models involve heavy annotations of object parts, preventing them from application to large-scale data, while visual attention networks can only locate the parts/regions roughly due to the lack of supervision or guidance. Recently, [He and Peng, 2017a] utilize natural language descriptions to help search the informative regions and combine with vision stream for final prediction. This method also integrates high-level information, but it directly models image-language pairs and requires detailed language descriptions for each image (e.g., ten sentences for each image in [He and Peng, 2017a]). Different from these methods, we organize knowledge about categories and part-based attributes in the form of knowledge graph and formulate a Knowledge-Embedded Representation Learning (KERL) framework to incorporate the knowledge graph into image feature learning to promote fine-grained image recognition.

To this end, our proposed KERL framework contains two crucial components: i) a Gated Graph Neural Network (GGNN) [Li *et al.*, 2015] that propagates node message through the graph to generate knowledge representation and ii) a novel gated mechanism that integrates this representation with image feature learning to learn attribute-aware features. Concretely, we first construct a large-scale knowledge graph that relates category labels with part-level attributes as shown in Figure 2. By initializing the graph node with information of a given image, our KERL framework might implicitly reason about the discriminative attributes for the image and associate these attributes with feature maps. In this way, our KERL framework can learn feature maps with a meaningful configuration that the highlighted regions finely associate with the relevant attributes in the graph. For example, the learned feature maps of samples from category "Bohemian Waxwing" always highlight the regions of head and wings, because these regions relate to attributes "head pattern: masked", "head color: white & black" and "wing: iridescent" that are key to distinguish this category from others. This characteristic also provides insight into why the framework improves performance.

The major contributions of this work are summarized to three-fold: 1) This work formulates a novel Knowledge-Embedded Representation Learning framework that incorporates high-level knowledge graph as extra guidance for image representation learning. To the best of our knowledge, this is the first work to investigate this point. 2) With the guidance of knowledge, our framework can learn attribute-aware feature maps with a meaningful and interpretable configuration that the highlighted regions are finely related to the relevant attributes in the graph, which can also explain performance improvement. 3) We conduct extensive experiments on the widely used Caltech-UCSD bird dataset [Wah *et al.*, 2011] and demonstrate the superiority of the proposed KERL framework over the leading fine-grained image classification methods.

2 Related Work

We review the related work in term of two research streams: fine-grained image classification and knowledge representation.

2.1 Fine-Grained Image Classification

With the advancement of deep learning [He *et al.*, 2016; Simonyan and Zisserman, 2014], most works rely on deep Convolutional Neural Networks (CNNs) to learn discriminative features for fine-grained image recognition, which exhibit a notable improvement compared with conventional hand-crafted features [He and Peng, 2017b; Lin *et al.*, 2015b]. To better capture subtle visual difference for fine-grained classification, bilinear models [Lin *et al.*, 2015b; Gao *et al.*, 2016; Kong and Fowlkes, 2017] is proposed to compute high-order representation that can better model local pairwise feature interactions by two independent subnetworks. Another common approach for distinguishing subtle visual difference among sub-ordinate categories is first locating discriminative regions and then learning appearance model conditioned on these regions [Zhang *et al.*, 2014; Huang *et al.*, 2016]. However, these methods involve in heavy annotations of object parts, and moreover, manually defined parts may not be optimal for the final recognition. Instead, He et al. [He and Peng, 2017a] adopt salient region localization techniques [Chen *et al.*, 2016; Zhou *et al.*, 2016] to automatically generate bounding box annotations of the discriminative regions. Recently, visual attention models [Mnih *et al.*, 2014; Wang *et al.*, 2017; Chen *et al.*, 2018; Liu *et al.*, 2018] have been intensively proposed to automatically search the informative regions, and some works also apply this technique to fine-grained recognition task [Liu *et al.*, 2016; Zheng *et al.*, 2017; Peng *et al.*, 2018]. [Liu *et al.*, 2016] introduce a reinforcement learning framework to adaptively glimpse local discriminative regions and propose a greedy reward strategy to train the framework with image-level annotations. [Fu *et al.*, 2017] further introduce a recurrent attention convolutional neural network to recursively learn the attentional regions at multiple scales and region-based feature representation. [Liu *et al.*, 2017] utilize part-level attribute to guide locating the attentional regions, which is related to ours. However, we organize the category-attribute relationships in the form of knowledge graph and implicitly reason discriminative attributes on the graph rather than using object-attribute pairs directly.

2.2 Knowledge Representation

Learning knowledge representation for visual reasoning increasingly receives attention as it benefits various tasks [Malisiewicz and Efros, 2009; Lao *et al.*, 2011; Zhu *et al.*, 2014; Lin *et al.*, 2017] in vision community. For instance, [Zhu *et al.*, 2014] learn a knowledge base using a Markov Logic Network and employ first-order probabilistic inference to reason the object affordances. These approaches usually involve in hand-crafted features and manually-defined propagation rules, preventing them from end-to-end training. Most recently, a series of efforts are dedicated to adapt neural networks to process graph-structured data [Duvenaud *et al.*,

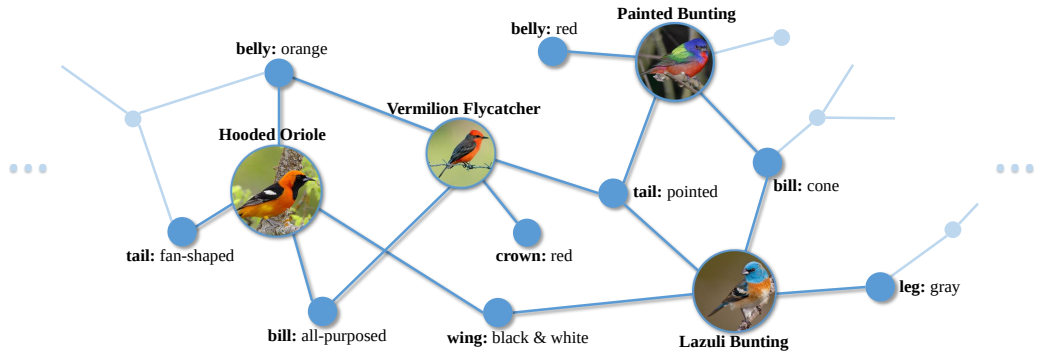

Figure 2: An example knowledge graph for modeling the category-attribute correlations on the Caltech-UCSD bird dataset.

2015; Niepert *et al.*, 2016]. For example, [Niepert *et al.*, 2016] sort the nodes in the graph based on the graph edges to regular sequence and directly feed the node sequence to a standard CNN for feature learning. These methods are tried on small, clean graphs such as molecular datasets [Duvenaud *et al.*, 2015] or are used to encode the contextual dependencies for vision tasks [Liang *et al.*, 2016].

GGNN [Li *et al.*, 2015] is a fully differentiable recurrent neural network architecture for graph-structured data, which recursively propagates node message to its neighbors to learn node-level features or graph-level representation. Several works have developed a series of graph neural network variants and successfully apply them to various tasks, such as 3DGNN for RGBD semantic segmentation [Qi *et al.*, 2017], model-based GNN for situation recognition [Li *et al.*, 2017], and GSNN for multi-label image recognition [Marino *et al.*, 2017]. Among these works, GSNN [Marino *et al.*, 2017] is mostly related to ours in the spirit of GGNN based knowledge graph encoding, but it simply concatenates image and knowledge features for image classification. In contrast, we develop a novel gated mechanism to embed the knowledge representation into image feature learning to enhances the feature representation. Besides, our learned feature maps exhibit insightful configurations that the highlighted regions finely accord with the semantic attributes in the graph, which also provide insight to explain performance improvement.

3 KERL Framework

In this section, we first briefly review the GGNN and present the construction of our knowledge graph that relates category labels with their part-level attributes. Then, we introduce our KERL framework in detail, which consists of a GGNN for knowledge representation learning and a gated mechanism to embed knowledge into discriminative image representation learning. An overall pipeline of the framework is illustrated in Figure 3.

3.1 Review of GGNN

We briefly introduce the GGNN [Li *et al.*, 2015] for completeness. GGNN is recurrent neural network architecture that can learn features for arbitrary graph-structured data by iteratively updating node features. For the propagation process, the input data is represented as a graph $\mathcal{G} = \{\mathbf{V}, \mathbf{A}\}$, in which $\mathbf{V}$ is the node set and $\mathbf{A}$ is the adjacency matrix that denotes the connections among nodes in the graph. For each node $v \in \mathbf{V}$, it has a hidden state $\mathbf{h}_v^t$ at time step t, and the hidden state at $t = 0$ is initialized by the input feature vector $\mathbf{x}_v$ that depends on the problem in hand. Thus, the basic recurrent process is formulated as

$$\begin{aligned}
\mathbf{h}_v^0 =& \mathbf{x}_v \\
\mathbf{a}_v^t =& \mathbf{A}_v^\top [\mathbf{h}_1^{t-1} \dots \mathbf{h}_{|\mathbf{V}|}^{t-1}]^\top + \mathbf{b} \\
\mathbf{z}_v^t =& \sigma(\mathbf{W}^z \mathbf{a}_v^t + \mathbf{U}^z \mathbf{h}_v^{t-1}) \\
\mathbf{r}_v^t =& \sigma(\mathbf{W}^r \mathbf{a}_v^t + \mathbf{U}^r \mathbf{h}_v^{t-1}) \\
\widetilde{\mathbf{h}_v^t} =& \tanh\left(\mathbf{W}\mathbf{a}_v^t + \mathbf{U}(\mathbf{r}_v^t \odot \mathbf{h}_v^{t-1})\right) \\
\mathbf{h}_v^t =& (1 - \mathbf{z}_v^t) \odot \mathbf{h}_v^{t-1} + \mathbf{z}_v^t \odot \widetilde{\mathbf{h}_v^t}
\end{aligned} \tag{1}$$

where $\mathbf{A}_v$ is a sub-matrix of $\mathbf{A}$ denoting the connections of node v with its neighbors. σ and $\tanh$ are the logistic sigmoid and hyperbolic tangent functions, respectively, and $\odot$ denotes the element-wise multiplication operation. The propagation process is repeated until a fixed iteration T, and we can obtain the final hidden states $\{\mathbf{h}_1^T, \mathbf{h}_2^T, \dots, \mathbf{h}_{|\mathbf{V}|}^T\}$. For notation simplification, we denote the computation process of equation (1) as $\mathbf{h}_v^t = \text{GGNN}(\mathbf{h}_1^{t-1}, \dots, \mathbf{h}_{|\mathbf{V}|}^{t-1}; \mathbf{A}_v)$.

3.2 Knowledge Graph Construction

The knowledge graph refers to an organization of a repository of visual concepts including category labels and part-level attributes, with nodes representing the visual concepts and edges representing their correlations. The graph is constructed based on the attribute annotations of the training samples. An example knowledge graph for the Caltech-UCSD bird dataset [Wah *et al.*, 2011] is presented in Figure 2.

Visual concepts. A visual concept refers to either a category label or an attribute. The attribute is an intermediate semantic representation of objects, and usually, it is key to distinguish two subordinate categories. Given a dataset that covers C object categories and A attributes, the graph has a node set $\mathbf{V}$ with $C + A$ elements.

Correlation. The correlation between a category label and an attribute indicates whether this category possesses the corresponding attribute. However, for the fine-grained task, it

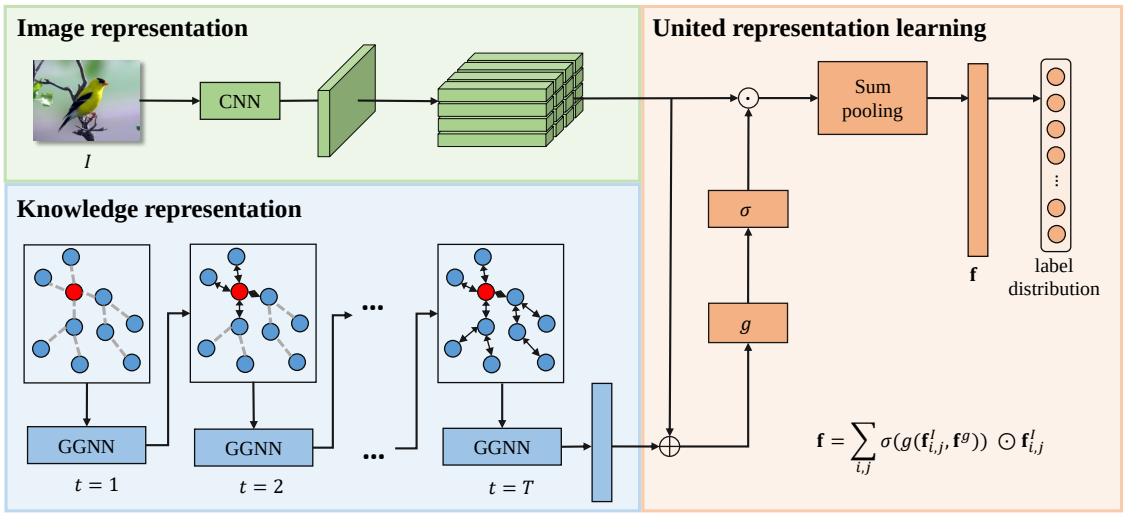

Figure 3: An overall pipeline of our proposed knowledge-embedded representation learning framework. The framework primally consists of a GGNN that takes the knowledge graph as input and propagates node information through the graph to learn knowledge representation, and a gated mechanism that embeds the representation into the image feature learning to learn attribute-aware features. All components of the framework can be trained in an end-to-end fashion.

is common that merely some instances of a category possess a specific attribute. For example, for a specific category, it is possible that one instance has a certain attribute, but another instance does not have. Thus, such category/attribute correlation is uncertain. Fortunately, we can assign an attribute/object instance pair with a score that denotes how likely this instance has the attribute. Then, we can sum up the scores of attribute/object instance pairs for all instances belonging to a specific category and obtain a score to denote the confidence that this category has the attribute. All the scores are linearly normalized to $[0, 1]$ to achieve a $C \times A$ matrix $\mathbf{S}$. Note that no connection exists between two object category nodes or between two attribute nodes; thus the complete adjacency matrix can be expressed as

$$\mathbf{A}_c = \begin{bmatrix} \mathbf{0}_{C\times C} & \mathbf{S} \\ \mathbf{0}_{A\times C} & \mathbf{0}_{A\times A} \end{bmatrix}, \tag{2}$$

where $\mathbf{0}_{W\times H}$ denotes a zero matrix of size $W \times H$. In this way, we can construct a knowledge graph $\mathcal{G} = \{\mathbf{V}, \mathbf{A}_c\}$.

3.3 Knowledge Representation Learning

After building the knowledge graph, we employ the GGNN to propagate node message through the graph and compute a feature vector for each node. All the feature vectors are then concatenated to generate the final representation for the knowledge graph.

We initialize the node referring to category label i with a score s_i that represents the confidence of this category being presented in the given image, and the node referring to each attribute with zero vector. The score vector $\mathbf{s} = \{s_0, s_1, \ldots, s_{C-1}\}$ for all categories is estimated by a pre-trained classier that will be introduced in detail in section 4.1. Thus, the input feature for each node can be represented as

$$\mathbf{x}_v = \begin{cases} [s_i, \mathbf{0}_{n-1}] & \text{if node } v \text{ refers to category } i \\ [\mathbf{0}_n] & \text{if node } v \text{ refers to an attribute} \end{cases}, \tag{3}$$

where $\mathbf{0}_n$ is a zero vector with dimension n. As discussed above, messages of all nodes are propagated to each other during the propagation process. With the computational process of Equation 1, $\mathbf{A}_c$ are used to propagate message from a certain node to its neighbors, and we use matrix $\mathbf{A}_c^\top$ for reverse message propagation. Thus, the adjacency matrix is $\mathbf{A} = [\mathbf{A}_c \quad \mathbf{A}_c^\top]$.

For each node v, its hidden state is initialized using $\mathbf{x}_v$, and at timestep t, the hidden state $\mathbf{h}_v^t$ is updated using the propagation process as Equation (1), expressed as

$$\begin{aligned} \mathbf{h}_v^0 &= \mathbf{x}_v \\ \mathbf{h}_v^t &= \text{GGNN}(\mathbf{h}_1^{t-1}, \ldots, \mathbf{h}_{|\mathbf{V}|}^{t-1}; \mathbf{A}_v) \end{aligned}. \tag{4}$$

At each iteration, the hidden state of each node is determined by its history state and the messages sent by its neighbors. In this way, each node can aggregate information from its neighbors and simultaneously transfer its message to its neighbors. This process is shown in Figure 3. After T iterations, the message of each node has propagated through the graph, and we can get the final hidden state for all nodes in the graph, i.e., $\{\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_{|\mathbf{V}|}^T\}$. Similar to [Li *et al.*, 2015], the node-level feature is computed by

$$\mathbf{o}_v = o(\mathbf{h}_v^T, \mathbf{x}_v), v = 1, 2, \ldots, |\mathbf{V}|, \tag{5}$$

where o is an output network that is implemented by a fully-connected layer. Finally, these features are concatenated to produce the final knowledge representation $\mathbf{f}^g$.

3.4 United Representation Learning

In this part, we introduce the gated mechanism that embeds the knowledge representation to enhance image representation learning.

Image feature extraction. We start by introducing the image feature extraction. As compact bilinear model [Gao *et*

al., 2016] works well on fine-grained image classification, we straightforwardly apply this model to extract image features. Specifically, given an image, we utilize a fully convolutional network (FCN) to extract feature maps with a size of $W' \times H' \times d$, and a compact bilinear operator to produce feature maps $\mathbf{f}^I$. Note that we do not perform sum pooling like [Gao *et al.*, 2016] thus the size of $\mathbf{f}^I$ is $W' \times H' \times c$. For fair comparisons with existing works, we employ the convolutional layers of the VGG16-Net to implement the FCN and follow the default setting as [Gao *et al.*, 2016] to set c as 8192.

[Gao *et al.*, 2016] treats all features equally important and simply performs sum pooling to obtain c-dimensional features for prediction. In the context of fine-grained image classification, it is crucial to attend to the discriminative regions to capture the subtle difference between different subordinate categories. In addition, the knowledge representation encodes category-attribute correlation and it may capture the discriminative attributes. Thus, we embed this representation into image feature learning to learn feature corresponding to this attributes. Specifically, we introduce a gated mechanism that optionally allows the informative features through while suppressing the non-informative features under the guidance of knowledge, which can be formulated as

$$\mathbf{f} = \sum_{i,j} \sigma\left(g\left(\mathbf{f}^I_{i,j}, \mathbf{f}^g\right)\right) \odot \mathbf{f}^I_{i,j}, \tag{6}$$

where $\mathbf{f}^I_{i,j}$ is the feature vector at location (i, j). $\sigma\left(g\left(\mathbf{f}^I_{i,j}, \mathbf{f}^g\right)\right)$ acts as a gated mechanism that decides which location is more important. g is a neural network that takes the concatenation of $\mathbf{f}^I_{i,j}$ and $\mathbf{f}^g$ as input and outputs a c-dimensional real-value vector. It is implemented by two stacked fully connected layers in which the first one is 10752 (8192 + 512×5) to 4096 followed by the hyperbolic tangent function while the second one is 4096 to 8192. The feature vector $\mathbf{f}$ is then fed into a simple fully-connected layer to compute the score vector $\mathbf{s}$ for the given image.

4 Experiments

4.1 Experiment Settings

Datasets. We evaluate our KERL framework and the competing methods on the Caltech-UCSD bird dataset [Wah *et al.*, 2011] that is the most widely used benchmark for fine-grained image classification. The dataset covers 200 species of birds, which contains 5,994 images for training and 5,794 for test. Except for the category label, each image is further annotated with 1 bounding box, 15 part key-points, and 312 attributes. As shown in Figure 4, the dataset is extremely challenging because birds from similar species may share very similar visual appearance while birds within the same species undergo drastic changes owing to complex variations in scales, viewpoints, occlusion, and background. In this work, we evaluate the methods in two settings: 1) "bird in image": the whole image is fed into the model at training and test stages, and 2) "bird in bbox": the image region at the bounding box is fed into the model at training and test stages.

Implementation details. For the GGNN, we utilize the compact bilinear model released by work [Gao *et al.*, 2016] to produce the scores to initialize the hidden states. For fair comparisons, the model is implemented with VGG16-Net and trained on the training part of the Caltech-UCSD bird dataset. The dimension of the hidden state is set to 10 and that of the output feature is set to 5. The iteration time T is set to 5. The KERL framework is jointly trained using the cross-entropy loss. All components of the framework are trained with SGD except GGNN that is trained with ADAM following [Marino *et al.*, 2017].

Figure 4: Samples from the Caltech-UCSD birds dataset. It is extremely difficult to categorize them due to large intra-class variance and small inter-class variance.

4.2 Comparison with State-of-the-Art Methods

In this subsection, we compare our KERL framework with 16 state-of-the-art methods, among which, some use merely image-level labels, and some also use bounding box/parts annotations; thus we also present this information for fair and direct comparisons. The methods are evaluated in both two settings, i.e., "bird in image" and "bird in bbox", and the results are reported in Table 1. For the "bird in bbox" setting, the previous well-performing methods are PN-CNN and SPDA-CNN that achieve the accuracies of 85.4% and 85.1% respectively, but they require strong supervision of ground truth part annotations. The accuracy of B-CNN is also up to 85.1%, but it relies on a very high-dimensional feature representation (250k dimensions). In contrast, the KERL framework requires no ground truth part annotations and utilize a much lower-dimensional feature representation (i.e., 8,192 dimensions), but it achieves an accuracy of 86.6% that outperforms all previous state-of-the-art methods. For the "bird in image" setting, most existing methods explicitly search discriminative regions and aggregate deep features of these regions for classification. For example, RA-CNN recurrently discovers image regions over three scales and achieves an accuracy of 85.3%. Besides, CVL combines detailed human-annotated text description for each image with the visual features to further improve the accuracy to 85.6%. Different from them, the KERL framework learns knowledge representation that encodes category-attribute correlations and incorporates this representation for feature learning. In this way, our method can learn more discriminative attribute-related

Methods	BA	PA	Acc. (%)
Part-RCNN [Zhang *et al.*, 2014]	√	√	76.4
DeepLAC [Lin *et al.*, 2015a]	√	√	80.3
SPDA-CNN [Zhang *et al.*, 2016a]	√	√	85.1
PN-CNN [Branson *et al.*, 2014]	√	√	85.4
PA-CNN [Krause *et al.*, 2015]	√		82.8
CB-CNN w/ bbox [Gao *et al.*, 2016]	√		84.6
FCAN w/ bbox [Liu *et al.*, 2016]	√		84.7
B-CNN w/ bbox [Lin *et al.*, 2015b]	√		85.1
AGAL w/ bbox [Liu *et al.*, 2017]	√		85.5
KERL w/ bbox	√		**86.6**
KERL w/ bbox & w/ HR	√		**86.8**
TLAN [Xiao *et al.*, 2015]			77.9
DVAN [Zhao *et al.*, 2017]			79.0
MG-CNN [Wang *et al.*, 2015]			81.7
B-CNN w/o bbox [Lin *et al.*, 2015b]			84.1
ST-CNN [Jaderberg *et al.*, 2015]			84.1
FCAN w/o bbox [Liu *et al.*, 2016]			84.3
PDFR [Zhang *et al.*, 2016b]			84.5
CB-CNN w/o bbox [Gao *et al.*, 2016]			85.0
RA-CNN [Fu *et al.*, 2017]			85.3
AGAL w/o bbox [Liu *et al.*, 2017]			85.4
CVL [He and Peng, 2017a]			85.6
KERL			**86.3**
KERL w/ HR			**87.0**

Table 1: Comparisons of our KERL framework with existing state of the arts on the Caltech-UCSD bird dataset. BA and PA denote bounding box annotations and part annotations, respectively, and HR denotes highlighted regions. √ indicates corresponding annotations are used during training or test.

features, leading to improvement in performance, i.e., 86.3% in accuracy. Note that AGAL also employs part-level attributes for fine-grained classification, but it achieves accuracies of 85.5% and 85.4% in two settings, respectively, much worse than ours. These comparisons well demonstrate the effectiveness of the KERL framework method over existing algorithms.

Attention-based methods aggregate features of both image and located regions to promote fine-grained classification, and our results reported above merely use image features. Our KERL framework can learn feature maps that highlight the regions related to discriminative attributes, as discussed in section 4.4; thus, we also aggregate features of the highlighted regions to improve performance. Specifically, we sum up the feature values across channels to get a score at each location, and draw a region with a size of a 6×6 centered at each location. We adopt non-maximum suppression to exclude the seriously overlapped regions and select top three ones. Three corresponding regions with a size of 96×96 ($16\times$ mapping between the original image and feature map) in the image are cropped, resized to 224×224 and fed to the VGG16 net to extract feature, respectively. The features are concatenated and fed to a fully-connected layer to compute the score vector, which is further averaged with the results of KERL to achieve the final results. It boosts the accuracies to 86.8% and 87.0% in two settings respectively.

4.3 Contribution of Knowledge Embedding

Note that our KERL framework employs CB-CNN [Gao *et al.*, 2016] as the baseline. Here, we emphasize the comparison with this baseline method to demonstrate the significance of knowledge embedding knowledge. As shown in Table 2, the CB-CNN achieves accuracies of 84.6% and 85.0% in "bird in bbox" and "bird in image" settings. By embedding the knowledge representation, the KERL framework boosts the accuracies to 86.6% and 86.3%, improving those of the CB-CNN by 2.0% and 1.3%, respectively.

To further clarify the contribution of knowledge guided feature selection, we implement two more baseline methods: self-guided feature learning and feature concatenation.

Comparison with self-guided feature learning. To better verify the benefit of embedding knowledge for feature learning, we conduct an experiment that removes the GGNN and only feeds the image features to the gated neural network, with other components left unchanged. The comparison results are presented in Table 2. It merely exhibits minor improvement over the baseline CB-CNN as it does not incur additional information but only increasing the complexity of the model. As expected, it performs much worse than ours.

Comparison with feature concatenation. To validate the benefit of our knowledge embedding method, we further conduct an experiment that incorporates knowledge by simply concatenating the image and graph feature vectors, followed by a fully-connected layer for classification. As shown in Table 2, directly concatenating image and graph features can achieve accuracies of 85.4% and 85.5% in the two settings, which is slightly better than the original CB-CNN but still much worse than ours. This indicates our knowledge incorporation method can make better use of knowledge to facilitate fine-grained image classification.

Methods	"bird in bbox"	"bird in image"
CB-CNN	84.6	85.0
self-guided selection	84.8	85.3
concatenation	85.4	85.5
Ours	**86.6**	**86.3**

Table 2: Accuracy comparisons (in %) of our KERL framework, feature concatenation, self-guided feature selection and baseline CB-CNN model on the Caltech-UCSD bird dataset.

4.4 Representation Visualization

With knowledge embedding, our KERL framework can learn feature maps with an insightful configuration that the highlighted regions are always related to relevant attributes. Here, we visualize the feature maps before sum pooling to better evaluate this point in Figure 5. We sum up the feature values across channels at each location and normalize them to $[0, 1]$. At each row, we present the learned feature maps of several samples taken from a specific category and a sub-graph that shows the correlations of this category with its attributes. We find that the highlighted regions for samples of the same category refer to the same semantic parts, and these parts finely accord with the attributes that well distinguish this category from others. Taking the category of "Sayornis" as example, our KERL framework consistently highlights the regions of

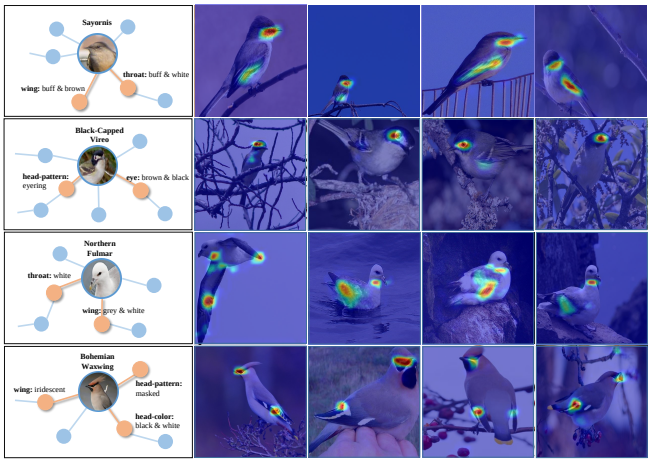

Figure 5: Visualization of the feature maps learnt by our KERL framework. At each row, we present some samples of a specific category and a sub-graph that denotes the correlations of this category with its attributes. The relevant attributes are highlighted in orange circles.

throats and wings for all samples, which correspond to two key attributes, i.e., "throat: buff & white" and "wing: buff & brown" (highlighted with orange circle in Figure 5). This suggests our KERL framework can learn attribute-aware features that can better capture subtle differences between different subordinate categories. Also, it can provide an explanation for the performance improvement of our framework.

To clearly verify that it is the knowledge embedding that brings about such appealing characteristic, we further visualize the feature maps generated by the CB-CNN model in Figure 6. We visualize the samples the same with those of the first two categories in Figure 5 for direct comparison. It is observed that some highlighted regions lie in the background and some scatter over the whole body of the birds.

Figure 6: Visualization of the feature maps generated by the baseline CB-CNN model. The samples are the same with those of the first two categories in Figure 5 for direct comparison.

5 Conclusion

In this paper, we propose a novel Knowledge-Embedded Representation Learning (KERL) framework to incorporate knowledge graph as extra guidance for image feature learning. Specifically, the KERL framework consists of a GGNN to learn the graph representation, and a gated neural network to integrate this representation into image feature learning to learn attribute-aware features. Besides, our framework can learn feature maps with an insightful configuration that the highlighted regions are always related to the relevant attributes in the graph, and this can well explain the performance improvement of our KERL framework. Experiments and evaluations conducted on the Caltech-UCSD bird dataset well demonstrate the superiority of our KERL framework over existing state-of-the-art methods. It is an early attempt to embed high-level knowledge into the modern deep network to improve fine-grained image classification, and we hope it can provide a step towards the integration of knowledge and traditional computer vision frameworks.

References

[Branson *et al.*, 2014] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014.

[Chen *et al.*, 2016] Tianshui Chen, Liang Lin, Lingbo Liu, Xiaonan Luo, and Xuelong Li. Disc: Deep image saliency computing via progressive representation learning. *TNNLS*, 27(6):1135–1149, 2016.

[Chen *et al.*, 2018] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. In *AAAI*, pages 6730–6737, 2018.

[Duvenaud *et al.*, 2015] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, pages 2224–2232, 2015.

[Fu *et al.*, 2017] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, pages 4438–4446, 2017.

[Gao *et al.*, 2016] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR*, pages 317–326, 2016.

[He and Peng, 2017a] Xiangteng He and Yuxin Peng. Fine-graind image classification via combining vision and language. *CVPR*, pages 5994–6002, 2017.

[He and Peng, 2017b] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *AAAI*, pages 4075–4081, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huang *et al.*, 2016] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *CVPR*, pages 1173–1182, 2016.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.

[Kong and Fowlkes, 2017] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *CVPR*, pages 7025–7034, 2017.

[Krause *et al.*, 2015] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *CVPR*, pages 5546–5555, 2015.

[Lao *et al.*, 2011] Ni Lao, Tom Mitchell, and William W Cohen. Random walk inference and learning in a large scale knowledge base. In *EMNLP*, pages 529–539, 2011.

[Li *et al.*, 2015] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[Li *et al.*, 2017] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation recognition with graph neural networks. In *CVPR*, pages 4173–4182, 2017.

[Liang *et al.*, 2016] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *ECCV*, pages 125–143, 2016.

[Lin *et al.*, 2015a] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, pages 1666–1674, 2015.

[Lin *et al.*, 2015b] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, pages 1449–1457, 2015.

[Lin *et al.*, 2017] Liang Lin, Lili Huang, Tianshui Chen, Yukang Gan, and Hui Cheng. Knowledge-guided recurrent neural network learning for task-oriented action prediction. In *ICME*, pages 625–630, 2017.

[Liu *et al.*, 2016] Xiao Liu, Tian Xia, Jiang Wang, and Yuanqing Lin. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

[Liu *et al.*, 2017] Xiao Liu, Jiang Wang, Shilei Wen, Errui Ding, and Yuanqing Lin. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *AAAI*, pages 4190–4196, 2017.

[Liu *et al.*, 2018] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. In *IJCAI*, 2018.

[Malisiewicz and Efros, 2009] Tomasz Malisiewicz and Alyosha Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, pages 1222–1230, 2009.

[Marino *et al.*, 2017] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *CVPR*, pages 2673–2681, 2017.

[Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NIPS*, pages 2204–2212, 2014.

[Niepert *et al.*, 2016] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *ICML*, pages 2014–2023, 2016.

[Peng *et al.*, 2018] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention model for fine-grained image classification. *TIP*, 27(3):1487–1500, 2018.

[Qi *et al.*, 2017] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *CVPR*, pages 5199–5208, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wang *et al.*, 2015] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *ICCV*, pages 2399–2406, 2015.

[Wang *et al.*, 2017] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*, pages 464–472, 2017.

[Xiao *et al.*, 2015] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, pages 842–850, 2015.

[Zhang *et al.*, 2014] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849, 2014.

[Zhang *et al.*, 2016a] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *CVPR*, pages 1143–1152, 2016.

[Zhang *et al.*, 2016b] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR*, pages 1134–1142, 2016.

[Zhao *et al.*, 2017] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *TMM*, pages 1245–1256, 2017.

[Zheng *et al.*, 2017] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *CVPR*, pages 396–404, 2017.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[Zhu *et al.*, 2014] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, pages 408–424, 2014.