

# Mining Discriminative Visual Features Based on Semantic Relations

**Abstract.** In this paper, we present an embedding-based framework for fine-grained image classification so that the semantic of background knowledge of images can be internally fused in image recognition. Specifically, we propose a semantic-fusion model which explores semantic embedding from both background knowledge (such as text, knowledge bases) and visual information. Moreover, we present a multi-level embedding model extract multiple semantic segmentations of background knowledge. Experimental results on a challenging benchmark CUB-200-2011 dataset verify that our approach outperforms 9 state-of-the-art methods with significant advances.

## 1 Introduction

The goal of fine-grained image classification is to recognize subcategories of objects, such as identifying the species of birds, under some basic-level categories. Different from general-level object classification, fine-grained image classification is challenging due to the large intra-class variance and small inter-class variance. Often, human beings recognize an object not only by its visual outline but also access their accumulated knowledge on the object.

In this work, we made full use of category attribute knowledge and deep convolution neural network to construct a fusion-based model Semantic Visual Representation Learning for fine-grained image classification. SVRL consists of a multi-level embedding fusion model and a visual feature extract model.

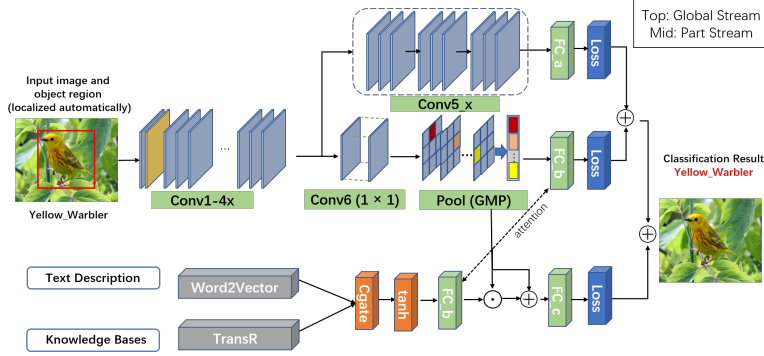
Our proposed SVRL has two distinct features: i) It is a novel weakly-supervised model for fine-grained image classification, which can automatically obtain the part region of image. ii) It can effectively integrate the visual information and relevant knowledge to improve the image classification.

## 2 Semantic Visual Representation Learning

The framework of SVRL is shown in Figure 1. Our method is based on a very simple intuition: knowledge conducting could find out the discriminative parts from other sub-categories. Therefore, we propose a multi-level fusion-based Semantic Visual Representation Learning model for learning latent semantic representations.

### 2.1 Discriminative Patch Detector

In this part, we adopt discriminative mid-level feature to classify images. Specifically, we set the  $1 \times 1$  convolutional filter as a small patch detector [4]. To



**Fig. 1.** Overview of our SVRL model. The structure of vision stream is ResNet50.

obtain a feature map of size  $C \times H \times W$ , the input image through a sequence of convolutional and pooling layers, each  $C \times 1 \times 1$  vector across channels at fixed spatial location represents a small patch at a corresponding location in the original image and the maximum value of the region can be found simply by picking the location in the entire feature map. In this way, we can pick out the discriminative region feature of the image.

## 2.2 Multi Embedding Fusion

From Figure 1, the knowledge stream consists of  $Cgate$  and visual fusion components. In our work, we use word2vector and TransR embedding method, note that, we can adaptively use  $N$  embedding methods not only two methods.

Given weight parameter  $w \in W$ , embedding space  $e \in E$ ,  $N$  is the number of embedding methods. The equation of  $Cgate$  as follow:  $Cgate = \frac{1}{N} \sum_1^N w_i e_i$ . where  $\sum_1^N w_i = 1$ . After we get the integrated feature space, we map semantic space into visual space by the same visual full connection  $FC\_b$  which is only trained by part stream visual vector. From here, we proposed an asynchronous learning, the semantic feature vector is trained every  $p$  epoch, but it does not update parameters of  $FC\_b$ . So the asynchronous method can not only keep semantic information but also learn better visual feature to fuse semantic space and visual space.

The equation of fusion is  $T = V + \alpha \times V \odot (\tanh(S))$ . The  $V$  is visual feature vector,  $S$  is semantic vector and  $T$  is fusion vector. Dot product is a fusion method which can intersect multiple information. The dimension of  $S$ ,  $V$ , and  $T$  are 200 we designed. The gate mechanism is consist of  $Cgate$ , tanh gate and the dot product of visual feature with semantic feature.

## 3 Experiments

In our experiments, the VGGNet and ResNet model are pre-trained from ImageNet, and then fine-tuned on CUB dataset. We train our model using SGD with

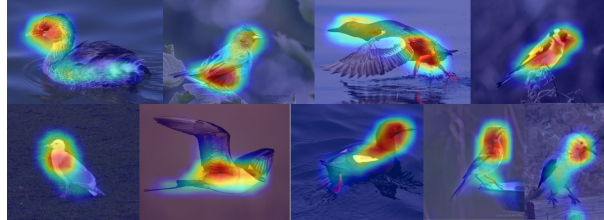
mini-batches 64 and learning rate is 0.0007. The hyperparameter weight of vision stream loss and knowledge stream loss are set 0.6, 0.3, 0.1. Two embedding weights are 0.3, 0.7 respectively.

**Table 1.** Comparisons with the state-of-the-art methods on CUB-200-2011 dataset.

Method	Train Annotation		Test Annotation		Accuracy
	Parts	BBox	Parts	BBox	
Part R-CNN	✓	✓	✓	✓	76.4
PA-CNN		✓		✓	82.8
SPDA-CNN	✓	✓		✓	85.1
AGAL-CNN[2]		✓		✓	85.5
DVAN					79.0
B-CNN					84.1
PDFS					84.5
CVL[1]					85.5
T-CNN[5]					86.2
<b>SVRL (ours)</b>					<b>87.1</b>

### 3.1 Classification Result and Comparison

Compared with 9 state-of-the-art fine-grained image classification methods, the result on CUB [3] of our SVRL are presented in Table 1. In our experiments, we did not use part annotations and BBox. We get 1.6% higher accuracy than the best part-based method AGAL which both use part annotations and BBox. Compared with T-CNN and CVL which do not use annotations and BBox, our method got 0.9%, 1.6% higher accuracy respectively. These works got good performance combined knowledge and vision, the difference between us is we fused multi-level embedding to get the knowledge representation and the mid-level vision patch region learns the discriminative feature. All in all, these experimental results clearly demonstrate the effectiveness of our SVRL approach, which fuses the vision and knowledge to exploit the correlation between visual feature and knowledge descriptions.



**Fig. 2.** The visualization of discriminative region in CUB-200-2011 dataset.

### 3.2 More Experiments and Visualization

To clearly demonstrate our SVRL model performance, we compare different variants of our SVRL approach. From Table 2, we can observe that combining vision and multi-level knowledge can achieve high accuracy than only one stream, which demonstrates that visual information with text description and knowledge are complementary in fine-grained image classification. Fig 2 is the visualization of discriminative region in CUB dataset.

**Table 2.** The result of different components and variants on CUB-200-2011.

Knowledge Components	Accuracy(%)	Vision Components	Accuracy(%)
Knowledge-W2V	82.2	Global-Stream Only	80.8
Knowledge-TransR	83.0	Part-Stream Only	81.9
Knowledge Stream-VGG	83.2	Vision Stream-VGG	85.2
Knowledge Stream-ResNet	83.6	Vision Stream-ResNet	85.9
<b>Our SVRL-VGG</b>	<b>86.5</b>	<b>Our SVRL-ResNet</b>	<b>87.1</b>

## 4 Conclusion

In this paper, we proposed a novel fine-grained image classification model SVRL as a way of efficiently leveraging external knowledge to improve fine-grained image classification. In particular, we observed that implicit structured and unstructured information can be well embedded into semantic embedding space. One important advantage of our approach was that our SVRL model could reinforce vision and knowledge representation, which can capture better discriminative feature for fine-grained classification. The experiment results verify that our proposed semantic-fused approach can really improve NN-based models utilizing only the visual information in fine-grained image classification. We believe that our proposal is helpful in fusing semantics internally when processing the cross media multi-information.

## References

1. He, X., Peng, Y.: Fine-grained image classification via combining vision and language. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7332–7340. IEEE (2017)
2. Liu, X., Wang, J., Wen, S., Ding, E., Lin, Y.: Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In: AAAI. pp. 4190–4196 (2017)
3. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
4. Wang, Y., Morariu, V.I., Davis, L.S.: Learning a discriminative filter bank within a cnn for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4148–4157 (2018)
5. Xu, H., Qi, G., Li, J., Wang, M., Xu, K., Gao, H.: Fine-grained image classification by visual-semantic embedding. In: IJCAI. pp. 1043–1049 (2018)