The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Audio Visual Attribute Discovery for Fine-Grained Object Recognition

Hua Zhang, Xiaochun Cao,* Rui Wang
State Key Laboratory of Information Security (SKLOIS), Institute of Information Engineering, CAS, Beijing, China
zhanghua@iie.ac.cn, caoxiaochun@iie.ac.cn*, wangrui@iie.ac.cn

Abstract

Current progresses on fine-grained recognition are mainly focus on learning the discriminative feature representation via introducing the visual supervisions e.g. part labels. However, it is time-consuming and needs the professional knowledge to obtain the accuracy annotations. Different from these existing methods based on the visual supervisions, in this paper, we introduce a novel feature named audio visual attributes via discovering the correlations between the visual and audio representations. Specifically, our unified framework is training with video-level category label, which consists of two important modules, the encoder module and the attribute discovery module, to encode the image and audio into vectors and learn the correlations between audio and images, respectively. On the encoder module, we present two types of feed forward convolutional neural network for the image and audio modalities. While an attention driven framework based on recurrent neural network is developed to generate the audio visual attribute representation. Thus, our proposed architecture can be implemented end-to-end in the step of inference. We exploit our models for the problem of fine-grained bird recognition on the CUB200-211 benchmark. The experimental results demonstrate that with the help of audio visual attribute, we achieve the superior or comparable performance to that of strongly supervised approaches on the bird recognition.

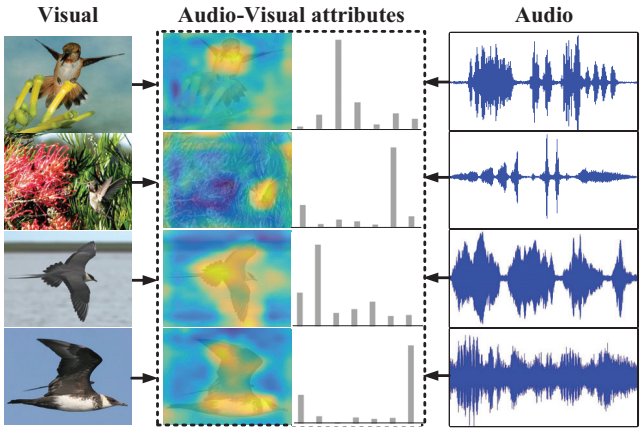

Figure 1: Visualization of audio visual attributes discovery model. Given the visual images (the first column) and the specified audios (the last column), we compute the correlations between these two modalities. In the second column, we present the visualization of training images via the audio visual module. And The third column shows the representation of audio visual attribute.

Introduction

Fine-grained recognition is an important research direction in the field of computer vision community, whose goal is to recognize the subcategories. Much progresses (Branson et al. 2014; Huang et al. 2016; Krause et al. 2015; Liu et al. 2012; Xie et al. 2013; Zhang et al. 2016a; 2014; 2016b) have been made with the development of deep convolutional neural network. The key challenge of fine-grained recognition is how to learn the discriminative feature to tell the subtle visual differences between the subcategories. In general, we can roughly divide the existing methods into two groups based on their distinct supervised label, which are the visual fine-grained annotations, e.g. the visual part label, (Branson et al. 2014; Huang et al. 2016; Krause et al. 2015; Liu et al. 2012; Xie et al. 2013; Zhang et al. 2016a; 2014; 2016b) and the text descriptions e.g. attributes (Berg and Belhumeur 2013; Elhoseiny, Saleh, and Elgammal 2013; Reed et al. 2016; Vedaldi et al. 2014; Zhang et al. 2013; He and Peng 2017).

Although there exists a great progress on fine-grained recognition assisted by the use of deep convolutional neural network, the performance is heavily dependent on the supervised labels to construct the discriminative feature representation, e.g. precise parts (Xie et al. 2013; Zhang et al. 2016a; 2014) or textual annotations (Reed et al. 2016; Vedaldi et al. 2014). However, the learned representation is still far from solving the fine-grained problems due to the limitations of supervised labels. For example, the variations of visual appearances would cause the region based representation ambiguous, and then weaken the discrimination of features. In addition, the textual descriptions are subjective which may be not consistent across persons. This would let the learned representation incorrect and limit the performance of fine-grained recognition.

To overcome above limitations of representation caused

*corresponding author

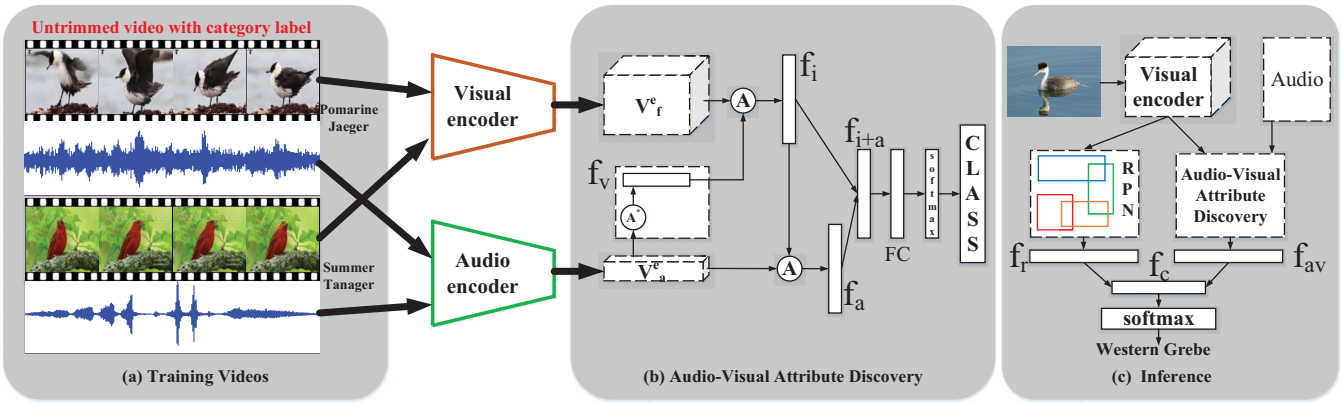

Figure 2: Overview flowchart of our proposed audio visual discovery and visual recognition. (a) Given the untrimmed videos with the category label, we firstly extract the visual and audio information which are fed into the feature encoder module to achieve the representations. (b) After that, we train the audio visual attribute discovery module based on the recurrent neural network. The loss function of this module is employed the softmax with the category label. (c) In the step of the inference, we extract the proposals with region proposal network (RPN), and then combine the visual features and the audio visual attribute to generate the image representation. Finally, a classifier is learned based on the representations to predict the image category.

by the supervised information, one directly way is to augment the supervised labels. But it is not an easy work to achieve these fine supervisions, which are very time-consuming and need the professional knowledge. Inspired by the fact that there exist an amount of untrimmed videos consisting of images and audio on the website, we could learn the audio visual attribute from these untrimmed videos with the weakly supervised category label. To that end, we firstly need to collect the videos from the web during the training, which is only labeled with the category labels. And our goal is to learn the audio and visual correlations from these untrimmed videos as shown in Fig. 1, which could generate the context representation for the object recognition and detection. As the novel feature representation learned from audio and images, we name it as audio visual attribute. However, the novel challenges is introduced to construct the audio visual attribute representation. The difficulties mainly concentrate on three aspects: Firstly, how to encode the audios from these untrimmed videos. Secondly, the way of discovering the correlations between the audio and the visual images. Last but not the least, how to efficiently infer the defined audio visual attribute from the images and train the discriminative classifiers. Further, all these three aspects should be considered at the same time to solve the challenges, .

In this paper, we address the challenges by constructing a novel deep neural network as shown in Fig.2, which is composed of two modules, feature encoder module and attribute discovery module, to extract the representations from the inputs and compute the correlation between images and audio. Specifically, the untrimmed videos are firstly segmented into several clips based on the uniform sampling as shown in Fig.2 (a). After that, the visual images are fed into the visual encoder to obtain the representation, while the audio is input into the audio encoder to achieve the consistency representation as presented in Fig.2 (b). Next, we use the extracted representations as the input to discover their correlations for generating the audio visual attribute as shown in Fig.2(c). To that end, an attention module based on the LSTM is employed for mining the aduio-visual attributes with the softmax loss function. To develop the end-to-end architecture, both the visual and audio encoders are implemented with the feed forward convolutional neural network. Finally, in the step of visual inference, we fed our proposed audio visual attribute into the classical object detection architecture faster rcnn (Ren et al. 2015a) as the context for the recognition. Experiments are conducted on the fine-grained benchmark CUB200-211. The experimental results demonstrate that with the help of audio visual attribute, we achieves the performance superior or comparable to that of strongly supervised approaches on the bird recognition.

Our contributions can be summarized into four folds: (i) We introduce a novel feature representation called **Audio Visual Attribute**, which is directly learned from the untrimmed videos with the weakly supervised category labels. (ii) A novel relevance mining approach is proposed to discover the correspondences between distinct modalities (the visual and the audio). (iii) Moreover, we propose an end-to-end framework in the step of feature generation and the category inference. (iv) Last but not the least, a novel video dataset is collected from the websites to extend the original CUB200-211 bird dataset.

Related work

Current state-of-the-art methods for fine-grained recognition are mostly based on deep neural network. In this section, we briefly present the advances on fine-grained recognition task with distinct supervisions, and then we give a summary on the progress of using the audio representation.

Fine-grained recognition

Based on the forms of supervision, existing method could be roughly divided into two groups: part annotations and tex-

tual descriptions. In (Zhang et al. 2014), Zhang *et al.* propose a framework named part R-CNN, which is inherited from (Girshick et al. 2016). The main idea is to construct the pose-normalization representation based on the training whole object and part detectors. Further, it uses the part annotations as the context supervised information. Lin *et al.* (Lin, RoyChowdhury, and Maji 2015) introduce a discriminative model by discovering the relationship from two object parts, and then generate the representation via a bilinear operator. The difference between our methods and (Zhang et al. 2014; Lin, RoyChowdhury, and Maji 2015) lies that our proposed feature is self-supervised learning without the need of the supervised labels.

Different from the part annotations, there exist an amount of approaches (Reed et al. 2016; Zhang et al. 2016a; Huang et al. 2016; Zhang et al. 2016b) focusing on learning the semantic features e.g., attributes. In (Reed et al. 2016), the authors introduce to learn the visual attributes with the sentence level descriptions, which has demonstrated its advantages on zero-shot learning. Huang et al. (Huang et al. 2016) introduce the part-stacked cnn architecture by explicitly explaining the salient visual differences between distinct categories. While Zhang et al. (Zhang et al. 2016a) propose to use two sub-networks: one for detection and one for recognition. The detection subnetwork is to detect the semantic part candidates, while the recognition subnetwork combine all the detections regions to develop the image representations for recognition. In (Zhang et al. 2016b), the authors propose to firstly pick the distinctive filters which refer to specific patterns, and then pool the deep filters scores via the weighted combination of Fisher Vector. The main advantage of this method is that it do not need any object or part annotations in the training and testing time.

Our proposed method is different from existing approaches in two aspects: One is the supervised information for learning the model. We propose to learn the audio visual attributes based on weakly supervised category label instead of supervised labels. Secondly, our proposed features is robust and easily achieved which can embed into any existing object detection framework.

Sound representation

With the development of deep neural network, the sound related applications has been extensively studies e.g., music recognition (Van den Oord, Dieleman, and Schrauwen 2013) and speech recognition (Hannun et al. 2014). Recent, some methods (Owens et al. 2016a; 2016b; Yusuf Aytar 2016) focus on constructing the visual related models to learn the sound representation. Owens el al. (Owens et al. 2016a) firstly propose to generative the audio information for the videos. They aim to generate the continuous audio for the object in the video. This demonstrates that there exist the correlations between the visual appearances and the audio feature. Furthermore, Owens et al. (Owens et al. 2016b) propose to extract the statistical characteristics of the audio information from the videos as the supervised label for image clustering. The experimental results demonstrate that the audio signal can be treated as the labels. While in (Yusuf Aytar 2016), SoundNet is developed which is the end-to-end framework to discover the corresponding between the audio and the visual appearances. The main differences between these methods and our proposed approach is that we aim to construct the feature representation based on the correlations between the visual and audio without much human efforts.

Method

In this section, we discuss our proposed method to encode the audio and visual image, discover the audio visual attribute and infer the category label with the help of extracted attributes. Firstly we introduce how to sample the training clips from the untrimmed videos. Then, the encoder module is described for extracting features from the distinct modalities. After that, we present the way of discovering the audio visual attribute. Finally, we are focusing on how to infer the category label in an end-to-end manner.

Clips sampling

We firstly collect the video from the websites[1] via treating the category label as the keyword. Each bird video usually contains the specific object with continuous and coherent motion patterns, which lasts for several seconds. However, the bird in untrimmed videos is not always producing the sound, in fact only part of the video clip includes the audio information. Thus, we need to select the video clips from the untrimmed videos, which contains the relatively consistent object patterns. Based on the response of audio, we segment the video into fragments with the equal length which is about 4 seconds. The steps of generating the clips are shown in Fig. 3.

Formally, a untrimmed video V is segmented into the clips set $C = \{c_i = \{f_b, f_e\}\}_{i=1}^{N}$, where N is the number of clips extract from the video, $\{f_b, f_e\}$ indicates the beginning and ending locations in the video, receptively. While the corresponding audio waveform for each clips c_i denoting as a_i. After the video clips extraction, we extract these frames from each clips. Considering that the frame rate of the video and the subtle motion within neighbor frames, each clips is uniform sampled and represented with several images. For the image in each clips c_i is represented as v_i^k where k denotes the k^{th} frame in the i^{th} clips.

Feature encoder module

As shown in Fig, 2, there exist two sorts of feature encoders, visual image encoder and audio encoder to extract the representations from the visual and audio inputs.

Visual image encoder. After the visual images are extracted, we fed them into the visual encoder which is composed of feed forward convolutional neural network. Specifically, the image v_i^k of clip c_i is firstly resize into the 227×227. Then we employ the VGG-16(Simonyan and Zisserman 2014) as feature extractor whose the last convolution layer is treated as the visual representation $\mathbf{V}_i^e \in \mathbf{R}^{14\times14\times512}$. The reasons to choose the last convolution layer are that we aim to reduce the effect of the background

[1]These bird videos are mainly retrieved from www.youtube.com and www.arkive.org

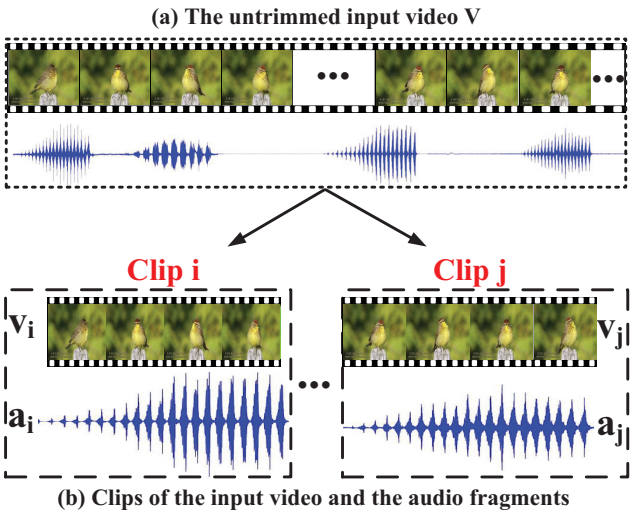

Figure 3: Flowchart of generating the video clips from untrimmed videos. (a) Given the untrimmed input video, we can observe that the audio is intermittent. (b) Then, based on the existing of the bird sound, we segment the original video into several clips which are composed of the consistency visual images and the corresponding audio fragment.

clutters and appearance variations. Furthermore, we also need to detect the discriminative regions in the image which are helpful for recognition.

Audio encoder. Since the audio waveform consists of the temporal sequences, the traditional approaches (Owens et al. 2016b) on encoding the audio are based on their statistical summary e.g. MFCC. However, it is quite unstable due to the inexact audio segmentations and the ambient noise. Instead, we propose to use the soundnet(Yusuf Aytar 2016) to encode the representation of each audio clip. Firstly, we fine-tune the soundnet with the category labels with the pre-trained model in (Yusuf Aytar 2016). After the model is learned, the last fully connected layer $\mathbf{V}_a^e \in \mathbf{R}^{512\times 1}$ is extracted to represent the audio clips.

Audio-visual attributes discovery

Our proposed audio visual attribute discovery module consisting of two components the correlation discovery and the classification. The correlation discovery is designed following the attention module based on the LSTM. When the visual representation $\mathbf{V}_i^e$ and the audio encoder $\mathbf{V}_a^e$ are generated, we transfer the visual representation to $\mathbf{V}_f \in \mathbf{R}^{196\times 512}$ and the audio to $\mathbf{V}_a \in \mathbf{R}^{196\times 512}$ via replication.

Attention mechanism. To discovery the correlation between the visual and the audio, we use the attention mechanism which is widely used in the field of feature selection. Different from these existing method, we introduce the attention mechanism to discover the correlation between the visual image and the audio. This is based on several observations: Firstly, we could extract different kinds of feature representations from the image while only a small proportion would be benefit for the classification, especially for the

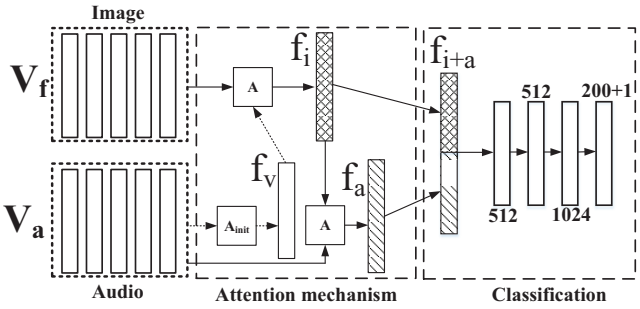

Figure 4: Overview of training the audio visual attribute discovery. Given the input features, we firstly compute the correlations between the visual and audio via the attention mechanism. Then, we concatenate the output to train the classifiers, which is composed of three hidden layers and one classification layer.

fine-grained recognition. Then, it is hard to achieve the standard audio template which would be random distribution. Thus we need to locate the discriminative parts of the audio waveform for the audio attribute discovery.

Our proposed attention model consists three steps. We firstly initialize the attention weighting with the audio features, and then we compute the attention features for the visual image. After that we use the visual attention features to compute the audio features as shown in Fig.4. The attention guidance is defined as $\mathcal{T}$ and the attention operation A is used to map the input features into the representation $f = A(V; \mathcal{T})$, where f is the output after using the attention model and V is the input features. In our method, we firstly initialize $\mathcal{T}$ to be ones, and then compute the attention map scores $f_v = A(V_a; \mathcal{T}_{init})$. After that, $\mathcal{T}$ is assigned to f_v. The process of compute the attention representation and the attention guidance $\mathcal{T}$ is defined as:

$$h = tanh(\mathbf{W}_a * \mathbf{V}_a^T + (\mathbf{W}_T * \mathcal{T})\mathbf{1}^T), \quad (1)$$

$$a^a = \frac{exp(w_{ha}^T h)}{\sum exp(w_{ha}^T h)}, \quad (2)$$

$$f_v = \sum a_i^a v_i^a, \quad (3)$$

where $\mathbf{W}_a \in R^{196\times 512}$ and $\mathbf{W}_T \in R^{196\times 512}$ is the weighting parameters. Moreover, $\mathbf{1}$ is the indicator matrix whose elements are all set to 1. a^a is the attention weighting of the input feature $\mathbf{V}_a$. f_v is the feature vector which encoded the attention weighting. For the visual image, we compute use the similarly steps:

$$h = tanh(\mathbf{W}_i * \mathbf{V}_i^T + (\mathbf{W}_T * \mathcal{T})\mathbf{1}^T), \quad (4)$$

$$a^i = \frac{exp(w_{hi}^T h)}{\sum exp(w_{hi}^T h)}, \quad (5)$$

$$f_i = \sum a_i^i v_i^i, \quad (6)$$

All the parameters have the same setting with the audio part, the only difference is that $\mathcal{T}$ is set to be f_a. After that we use

the iterations to compute the feature representations for the visual image f_i and the audio f_a.

Classification. We treat the audio attribute discovery as a classification problem. And a multi-layer perceptron is employed to compute the predictions. Specifically, we develop an architecture with three hidden layers, whose dimensions are set as 512, 512, and 1024, respectively.

$$h_1 = tanh(\mathbf{W}_{h_1} * f_{i+a}), \tag{7}$$
$$h_2 = tanh(\mathbf{W}_{h_2} * h_1), \tag{8}$$
$$h_3 = tanh(\mathbf{W}_{h_3} * h_2), \tag{9}$$
$$p = softmax(\mathbf{W}_c * h_3), \tag{10}$$

where h_1, h_2, and h_3 is the hidden representations, while $\mathbf{W}_{h_1}$,$\mathbf{W}_{h_2}$, and $\mathbf{W}_{h_3}$ are the weighting parameters to extract the hidden vectors. f_{i+a} is the concatenation of the attention feature f_i and f_a. p is the final predictions for each category and softmax means that we normalize the prediction scores with the softmax operation.

Training. After the descriptions of attribute discovery architecture in the previous subsection, we turn to discuss how to optimize the model parameters. we employ the standard back propagation method with with cross-entropy loss and L2 regularization. We set the number of category is 201. When the audio and the visual images are not in the same categories, the label is defined as 201. Moreover, the weighting parameter of the regularization is set to 0.05.

Inference

When the attributes model is training done, we extract the audio visual attribute from the test images. However, the audio visual attribute need two inputs: the visual image and the corresponding audio waveform. While in the test time, there is no audio information for the input images. To solve this problem, for each test image we configure it with the audio of all the categories. Thus the audio visual attribute for each test image is defined as the response scores with the audios whose dimension is set to 200 in all our experiments.

Considering the appearance variations, we propose to employ the framework of faster rcnn (Ren et al. 2015b) to improve the robustness of fine-grained classifiers. As shown in Fig. 2, our inference network is similar with faster rcnn which is firstly to extract the region proposals, and then learn the category classifiers. Differently we propose to embed the audio visual attribute into the classification network. Specifically, we concatenate the audio visual attribute representation to the last fully connected layer, and then the classifier is learned based on the extended representations.

Experiment

In this section, we show the experimental results and analysis of our proposed method. More specifically, the performance of our proposed is evaluated on the CUB-200-2011 bird dataset and the evaluation criterion is the mean average recognition accuracy.

Datasets

Image set Caltech-UCSD Birds dataset (CUB-200-2011) (Wah et al. 2011) is the widely used fine-grained classification benchmark. There exist 200 bird species with about 30 training images for each category. The total number images of this dataset is $11,788$. The supervised information contains the category label, the location of the bird in the image, the part locations of each bird, and the textual attribute descriptions. In our experiments, we only use the category label and the location of each bird in the image. Furthermore, we evaluate our proposed in two protocols as these existing methods (Zhang et al. 2016a; Lin, RoyChowdhury, and Maji 2015; Huang et al. 2016): One is that the location of bird does not provided in the training and test time, the other is that the location bounding box is given at the training and test time.

Video set We extend the CUB-200-2011 dataset via collecting videos from the websites. Specifically, we use the category label as the keyword to retrieve the bird videos. Although there exist an amount of videos, only a small parts are satisfying our conditions, which the audio information should exist. For each category, we collect about 3 different videos whose duration time is about 40 seconds. Then, we segment the videos into several fragments based on the duration of audio. The duration of the video fragment is about $4\sim8$ seconds and $4,850$ video fragments for 200 categories in total. The sources of these video fragments are mixed with avi, mkv and mp4. We format all the audio files in the same data type mp4.

Implementation Details

Our proposed architecture is implemented based on the open-source package torch7. More specifically, the input images are warped into a fixed size of 227×227. If the bounding box of the training samples provided, we firstly crop the images and then warp them to the fixed size. To train the feature encoder, we follow the fine-tuning training strategy. In all experiments, our networks are trained by stochastic gradient descent with 0.9 momentum. We initiate learning rate to be 0.0001 and decrease it by 0.1 after finishing about 20 epochs. The weight decay parameter is 0.0005. The visual image encoder is firstly fine-tuning with the pre-trained model on the ImageNet dataset, whose classifier layer is replaced with the 200 categories.

Results and Comparisons

Comparing with the baselines In this subsection, we verify the main parts of our unified framework. Specifically, the bird benchmark is divided into training, validation, and test part. The train and test samples are selected following (Wah et al. 2011). While for the validation set, we randomly choose 10% samples from the training set. We firstly check if the depth of the neural network would influence the recognition performance. To that end, the VGG-16 and the ResNet-50 are selected to the visual image encoder, and the rest part of our architecture is the same. In this comparison, we introduce two baselines: VGG-16 indicates that we fine-tune the deep model based on the original neural network. In the same way, ResNet-50 is also employed as the second baseline. The comparison results are listed in Table 1. The box in the visual annotation means that the location of object

Table 1: Comparisons of different settings of our unified framework on CUB-200-2011 dataset. Two kinds of convolutional neural network are selected as the basic.

Network	Features			Accuracy
	Train	Test	Audio	
VGG-16	Box	-	-	60.12%
	Box	Box	-	66.96%
	Box	-	+	65.69%
	Box	Box	+	69.82%
ResNet-50	Box	-	-	63.25%
	Box	Box	-	74.16%
	Box	-	+	69.80%
	Box	Box	+	75.38%

is used. We train the fine-grained classifiers by SVM based on concatenating the last fully connected layers of the visual features and the audio visual attribute.

From the experimental results, we can observe that with the help of the audio visual attribute the recognition accuracy is boosted about $\mathbf{5}\%$ from 60.12% to 65.69% based on VGG-16 and from 63.25% to 69.8% based on ResNet-50. When we use the location of object in the test, the accuracy is improved. This can demonstrate that our proposed audio visual attribute is capable of learning the location of object. Furthermore, the performance is raising with the depth of the neural network. The discriminative visual image encoder is benefit for our audio visual attribute discovery.

We believe that the performance could be explained in three folds: Firstly, our proposed model is introducing the audio as the context features. This could improve the discrimination of the image representation. Secondly, the audio visual attribute could be helpful for localizing the important visual regions, which would generate the robustness feature representation. Last but not the least, our proposed inference framework can integrate the audio and visual features into the unified image representation, which would further boost the performance of fine-grained recognition.

Comparing with the state-of-the-art In this subsection, we focus on comparing our proposed model with the state-of-the-arts approaches on fine-grained object recognition. As we use the faster rcnn as the basic inference structure, one of the important component is the region proposals generation. Specifically, we set the fixed threshold to choose the required training proposals. For each region proposal, we compute its IoU score with the groundtruth location and only save these proposals whose score is higher than 0.65. After that, each training images would generate about 400 proposals, which are resized into 227×227. In the test time, we select the proposals based on their objectness scores which is computed by RPN. Moreover, when the location of the object in the test image is given, the region proposals are chosen with the same policy as the training samples. Since there exist several proposals for each image, we collect all the predictions and use the max pooling operation to achieve the category prediction.

All the comparison results are presented in Table 2. To achieve the fair comparison, we set the visual encoder with two kinds of neural network, which are AlexNet and VGG-16. Our method obtain 83.6% accuracy without using the bounding box in the test. This is better than several baselines (Zhang et al. 2015; Lin, RoyChowdhury, and Maji 2015). Furthermore, comparing with these approaches (Lin, RoyChowdhury, and Maji 2015; Krause et al. 2015) with bounding box in the test, our method shows the advantages on the recognition accuracy. Even we also have a superiority on the accuracy comparing with (Simon and Rodner 2015), which uses the part annotation to learn the classifiers. This could further demonstrate the advantages of introducing the audio visual attribute to construct the image feature representation. However, our method is not improved much when we introduce the bounding box in the test time. This could be explained by the fact that the audio visual attribute is already considering these regions via attention maps. Comparing with (He and Peng 2017) which uses the text descriptions as the context, our approach obtains a litter improvements. This is because that the audio modality is more robust as the supervision, while the text descriptions are easily ambiguous caused by the visual variations. Furthermore, comparing with (Lam, Mahasseni, and Todorovic 2017), there exist a little degeneration on the performance. This could be explained in two folds: Firstly, with the body annotation the learned feature representation would be more robust to the pose variations and further improve the discrimination of deep features. Then, our proposed audio visual discovery module is biased due to the distribution of training images, which may lead to error on computing the attention map.

Based on the experimental results and the comparisons, we could observe that there are three main factors determining the performance of the fine-grained recognition. Firstly, the visual variations of the object would be compressed by the fine body annotations, which is capable of aligning all the images. For example, when we use the body part annotations, the performance of all the related methods are improving. Then, we introduce the audio visual attribute as the context to improve the discrimination of feature representation which further demonstrate multi-source features would be helpful for fined-grained recognition. Last but not the least, how to design the neural network is still an open problem. From the results, we know that there exist the positive correlation between the depth of neural network and the recognition accuracy. However, we should consider the computational efficiency to select the right network.

Discussion

The main challenge of the fine-grained visual recognition is the appearance variations which is caused by distinct pose, view, scale and occlusion. From the failed examples, we observe that without the sufficient training samples, our audio visual attributes model would be easily degenerated. The advantages of our proposed feature are in two aspects: Firstly, it is easily achieved without the human intervention which is more objective than these textual descriptions. Then, the audio visual features is robust which can encode the global feature of the object. Although our method has achieved the expected performance on bird recognition, there still exist some problems to be solved. Firstly, how to train the au-

Table 2: Comparisons with the state-of-the-art methods on CUB-200-2011. **BB** refers to the location of the birds in the image, while **P** is the annotation of the semantic parts on the birds.

Net	Methods	Train	Test	Acc.(%)
Alex	Constellation (Simon and Rodner 2015)	n/a	n/a	68.5
	Weak FGCV (Zhang et al. 2015)	n/a	n/a	75.0
	Attention (Xiao et al. 2015)	n/a	n/a	69.7
	CNN Aug (Sharif Razavian et al. 2014)	BB	BB	61.8
	Alignment (Gavves et al. 2013)	BB	BB	67.0
	No parts (Krause et al. 2015)	BB	BB	74.5
	DPD + DeCAF (Donahue et al. 2014)	BB+P	BB	65.0
	Deep LAC (Lin et al. 2015)	BB+P	BB	**80.2**
	Multi-Proposal (Shih et al. 2015)	BB+P	BB	**80.3**
	Part R-CNN (Zhang et al. 2014)	BB+P	BB	76.4
	PS-CNN (Huang et al. 2016)	BB+P	BB	76.6
	Our method	BB	n/a	72.3
	Our method	BB	BB	**77.5**
VGG	PDFS (Zhang et al. 2016b)	n/a	n/a	84.5
	Weak FGCV (Zhang et al. 2015)	n/a	n/a	78.92
	Bilinear CNN (Lin, RoyChowdhury, and Maji 2015)	n/a	n/a	74.2
	No parts (Krause et al. 2015)	BB	BB	82.80
	Bilinear CNN (Lin, RoyChowdhury, and Maji 2015)	BB	BB	84.10
	Constellation (Simon and Rodner 2015)	BB+P	BB	81.01
	SPDA-CNN (Zhang et al. 2016a)	BB+P	BB	85.14
	HIHCA (Cai, Zuo, and Zhang 2017)	n/a	n/a	85.3
	CVL (He and Peng 2017)	n/a	n/a	85.5
	RA-CNN (Fu, Zheng, and Mei 2017)	n/a	n/a	85.3
	HSnet (Lam, Mahasseni, and Todorovic 2017)	BB+P	BB+P	**87.5**
	Faster-rcnn(Ren et al. 2015a)	BB	n/a	82.5
	Our method	BB	n/a	85.6
	Our method	BB	BB	**86.6**

dio encoder to get the discriminative and robust audio representation. In particular, the similarities between different categories need to embed into the representation. Then, we should design the end-to-end training neural network instead of with two stages. Last but the least, how to develop the hierarchical structure representation would be benefit for several related applications. Moreover, we could treat the audio information as the supervision to learn the features from the visual images.

Conclusion

In this paper, we introduce a novel representation named audio visual attribute, which is discovered from the untrimmed videos. To obtain the audio visual attribute, a unified framework is developed, which consists of encoder module and attribute discovery module. On the encoder module, we use two types of deep neural network to encode the image and the audio into vector representations. Furthermore, the encoder neural network is executed with the feed forward convolutional neural network. While we implement the attribute discovery module via the attention model based on LSTM to detect the audio visual attributes. Thus, an end-to-end inference architecture is generated to test the images. Finally, we validate the effectiveness of our approach on the benchmark CUB200-211. From the experimental results, our proposed method achieves superior or comparable to that of strongly supervised approaches on the bird recognition, which demonstrates the effectiveness of introducing the audio visual attributes. In the future work, we would focus on how to fuse the multiple supervisions to learn the fine-grained object classifier and how to improve the discrimination of image representation via merging the multi-view features.

Acknowledgement

This work was supported by the National Key R&D Program of China(No.2016YFC0801002), National Natural Science Foundation of China (No.61602464, U1605252,61422213).

References

Berg, T., and Belhumeur, P. 2013. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 955–962.

Branson, S.; Van Horn, G.; Belongie, S.; and Perona, P. 2014. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*.

Cai, S.; Zuo, W.; and Zhang, L. 2017. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *ICCV*.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolu-

tional activation feature for generic visual recognition. In *ICML*, 647–655.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2584–2591.

Fu, J.; Zheng, H.; and Mei, T. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 4438–4446.

Gavves, E.; Fernando, B.; Snoek, C. G.; Smeulders, A. W.; and Tuytelaars, T. 2013. Fine-grained categorization by alignments. In *ICCV*, 1713–1720.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2016. Region-based convolutional networks for accurate object detection and segmentation. *TPAMI* 38(1):142–158.

Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

He, X., and Peng, Y. 2017. Fine-graind image classification via combining vision and language. In *CVPR*, 5994–6002.

Huang, S.; Xu, Z.; Tao, D.; and Zhang, Y. 2016. Part-stacked cnn for fine-grained visual categorization. In *CVPR*, 1173–1182.

Krause, J.; Jin, H.; Yang, J.; and Fei-Fei, L. 2015. Fine-grained recognition without part annotations. In *CVPR*, 5546–5555.

Lam, M.; Mahasseni, B.; and Todorovic, S. 2017. Fine-grained recognition as hsnet search for informative image parts. In *CVPR*, 2520–2529.

Lin, D.; Shen, X.; Lu, C.; and Jia, J. 2015. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, 1666–1674.

Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear cnn models for fine-grained visual recognition. In *CVPR*, 1449–1457.

Liu, J.; Kanazawa, A.; Jacobs, D.; and Belhumeur, P. 2012. Dog breed classification using part localization. In *ECCV*, 172–185.

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016a. Visually indicated sounds. In *CVPR*, 2405–2413.

Owens, A.; Wu, J.; McDermott, J. H.; Freeman, W. T.; and Torralba, A. 2016b. Ambient sound provides supervision for visual learning. In *ECCV*, 801–816.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 49–58.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015a. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015b. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 91–99.

Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR Workshops*, 806–813.

Shih, K. J.; Mallya, A.; Singh, S.; and Hoiem, D. 2015. Part localization using multi-proposal consensus for fine-grained categorization. *arXiv preprint arXiv:1507.06332*.

Simon, M., and Rodner, E. 2015. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *ICCV*, 1143–1151.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Van den Oord, A.; Dieleman, S.; and Schrauwen, B. 2013. Deep content-based music recommendation. In *NIPS*, 2643–2651.

Vedaldi, A.; Mahendran, S.; Tsogkas, S.; Maji, S.; Girshick, R.; Kannala, J.; Rahtu, E.; Kokkinos, I.; Blaschko, M. B.; Weiss, D.; et al. 2014. Understanding objects in detail with fine-grained attributes. In *CVPR*, 3622–3629.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; and Zhang, Z. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 842–850.

Xie, L.; Tian, Q.; Hong, R.; Yan, S.; and Zhang, B. 2013. Hierarchical part matching for fine-grained visual categorization. In *ICCV*, 1641–1648.

Yusuf Aytar, Carl Vondrick, A. T. 2016. Soundnet: Learning sound representations from unlabeled video. In *NIPS*.

Zhang, N.; Farrell, R.; Iandola, F.; and Darrell, T. 2013. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 729–736.

Zhang, N.; Donahue, J.; Girshick, R.; and Darrell, T. 2014. Part-based r-cnns for fine-grained category detection. In *ECCV*, 834–849.

Zhang, Y.; Wei, X.-s.; Wu, J.; Cai, J.; Lu, J.; Nguyen, V.-A.; and Do, M. N. 2015. Weakly supervised fine-grained image categorization. *IEEE TIP* 25(4):1713 – 1725.

Zhang, H.; Xu, T.; Elhoseiny, M.; Huang, X.; Zhang, S.; Elgammal, A.; and Metaxas, D. 2016a. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *CVPR*, 1143–1152.

Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; and Tian, Q. 2016b. Picking deep filter responses for fine-grained image recognition. In *CVPR*, 1134–1142.