Fine-Grained Image Classification via Combining Vision and Language

Xiangteng He and Yuxin Peng*, Peking University

北京大学 PEKING UNIVERSITY

Multimedia Information Processing Lab (MIPL)
Institute of Computer Science & Technology, Peking University

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

1. Introduction

Category	Vision	Language	Vision	Language
Heermann Gull		(1)A large bird with different shades of grey all over its body, white and black tail feathers, and a long sharp orange beak. (2)This bird is gray and black in color, with a orange beak. (3)This bird has black outer retices and white inner retires and an orange beak. ...		(1)This bird has a grey body, a white head with an orange bill and black wings, tarsus and feet. (2)A white bird with black wings and orange beak and eyes. (3)This white bird has a bright orange bill with a black tip. ...
Red Legged Kittiwake		(1)This bird has a white head, breast and belly with gray wings, red feet and thighs, and a red beak. (2)This is a white bird with gray wings, red webbed feet and a red beak. (3)Long bird with an orange beak and white feathers with grey colored wings. ...		(1)This white bird has grey wings and tail, with orange feet and tarsus and a pointed yellow bill. (2)This bird has a mellow yellow bill coloration and deep red feet (3)The bird has a yellow beak with a white head and orange web feet ...
Bohemian Waxwing		(1)This bird is light gray with a light orange patch on its under-tail covets, neck and crown, and a black malar stripe and nape. (2)This is a grey bird with a red and yellow tail and a red face. (3)This bird has wings that are gray and black and has a red crown ...		(1)This colorful bird has an orange crown, black eyebrows and secondaries with the primaries being rimmed in yellow. (2)This bird is grey with red and has a long, pointy beak. (3)The bird has a tan spiked crown and short bill. ...

➢ Motivation

- Not all the parts which obtained through the part detection models are beneficial and indispensable for classification.
- Fine-grained image classification requires more detailed visual descriptions which could not be provided by the part locations or attribute annotations.

➢ Contribution

- The ***vision stream*** learns deep representations from the original **visual information** via deep convolutional neural network.
- The ***language stream*** utilizes the natural language descriptions which could point out the discriminative parts or characteristics for each image, and provides a flexible and compact way of encoding the ***salient visual aspects*** for distinguishing sub-categories.
- Since the two streams are ***complementary***, combining the two streams can further achieves better classification accuracy.

2. Our CVL Approach

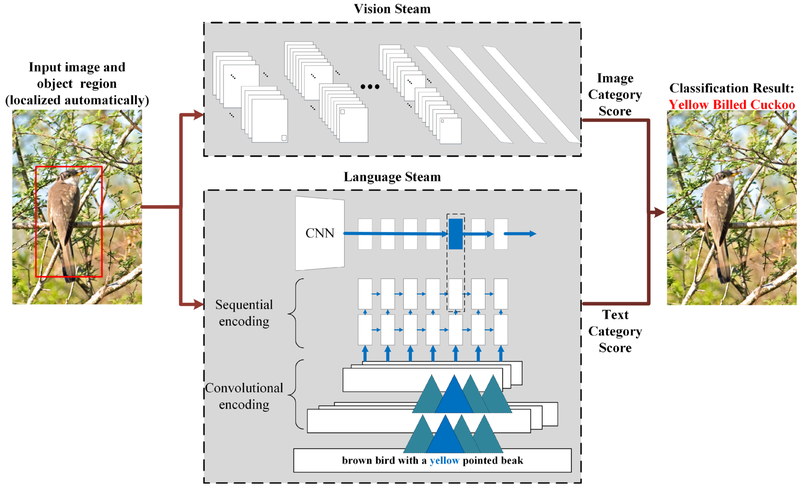
➢ Vision Stream

Given an image I, its object region b is generated via saliency extraction and co-segmentation, then the object region is clipped from the original image and saved as image I'. We take the original image I and its object image I' as the inputs of the CNN model to obtain the prediction, which is the result of the vision stream.

➢ Language Stream

We apply the deep structured joint embedding method, jointly embedding images and fine-grained visual descriptions. We use the inner product of features generated by deep neural encoders ($\theta(v)$ for images and $\varphi(t)$ for texts), and maximize the compatibility between a description and its matching image as well as minimize compatibility with images from other classes.

$$F(v,t) = \theta(v)^{\mathrm{T}}\varphi(t)$$



3. Experimental Results

Category	Image	Text Rank List(Top3)
Sooty Albatross		(1)This bird has **wings** that are **grey** and has a **black bill**. (2)This bird is **gray** in color, with a large curved beak. (3)This bird is white and brown in color, and has a **black beak**.
California Gull		(1)This bird has large feet, a short **yellow bill**, and a **black and white body**. (2)This bird has wings that are grey and has a **white belly** and **yellow bill**. (3)This bird has a **yellow beak** as well as a **white belly**.
Cerulean Warbler		(1)A little bird with a short, **grey bill**, **blue crown**, **nape**, **white breast**. (2)The bird has a **white** abdomen, black breast and **white** throat, **blue** specks. (3)This bird is **blue** and **white** in color with a black beak, and black eye rings.

➢ Comparisons with state-of-the-art methods

We compare with 12 state-of-the-art fine-grained image classification methods on CUB-200-2011 to verify the effectiveness of our CVL approach, which jointly integrates the vision and language streams to exploit the correlation between visual feature and nature language descriptions, and enhance their complementarity.

Method	Train Anno.		Test Anno.		Acc. (%)
	Bbox	Parts	Bbox	Parts	
Our CVL Approach					**85.55**
PD[Zhang et al, CVPR 2016]					84.54
Spatial Transformer[Jaderberg et al, NIPS 2015]					84.10
Bilinear-CNN [Lin et al, ICCV 2015]					84.10
NAC[Simon et al, ICCV 2015]					81.01
TL Atten[Xiao et al, CVPR 2015]					77.90
VGG-BGLm[Zhou et al, CVPR 2016]					75.90
PG Alignment[Krause et al, CVPR 2015]	√		√		82.80
Triplet-A (64)[Cui et al, CVPR 2016]	√		√		80.70
VGG-BGLm[Zhou et al, CVPR 2016]	√		√		80.40
SPDA-CNN[Zhang et al, CVPR 2016]	√	√	√		85.14
Part-based R-CNN[Zhang et al, ICCV 2014]	√	√	√	√	76.37
POOF[Berg et al, CVPR 2013]	√	√	√	√	73.30