Visual-textual Attention Driven Fine-grained Representation Learning

Xiangteng He and Yuxin Peng

***Abstract*—Fine-grained image classification is to recognize hundreds of subcategories belonging to the same basic-level category, which is a highly challenging task due to the quite subtle visual distinctions among similar subcategories. Most existing methods generally learn part detectors to discover discriminative regions for better classification accuracy. However, not all localized parts are beneficial and indispensable for classification, and the setting for number of part detectors relies heavily on prior knowledge as well as experimental results. As is known to all, when we describe the object of an image into text via natural language, we only focus on the pivotal characteristics, and rarely pay attention to common characteristics as well as the background areas. This is an involuntary transfer from human visual attention to textual attention, which leads to the fact that textual attention tells us how many and which parts are discriminative and significant. So textual attention of natural language descriptions could help us to discover visual attention in image. Inspired by this, we propose a visual-textual attention driven fine-grained representation learning (VTA) approach, and its main contributions are: (1) Fine-grained visual-textual pattern mining devotes to discovering discriminative visual-textual pairwise information for boosting classification through jointly modeling vision and text with generative adversarial networks (GANs), which automatically and adaptively discovers discriminative parts. (2) Visual-textual representation learning jointly combine visual and textual information, which preserves the intra-modality and inter-modality information to generate complementary fine-grained representation, and further improve classification performance. Comprehensive experimental results on the widely-used CUB-200-2011 and Oxford Flowers-102 datasets demonstrate the effectiveness of our VTA approach, which achieves the best classification accuracy compared with state-of-the-art methods.**

***Index Terms*—Fine-grained image classification, visual-textual attention, fine-grained visual-textual pattern mining, visual-textual representation learning.**

I. Introduction

FINE-GRAINED image classification aims to recognize similar subcategories in the same basic-level category. It is one of the most challenging and significant open problems of multimedia and computer vision, which achieves great progress as well as attracts extensive attention of academia and industry. The progress incarnates in four aspects: (1) More fine-grained domains have been covered, such as animal species [1], [2], plant breeds [3], [4], car types [5] and aircraft models [6]. (2) Methodologies of fine-grained image classification have achieved promising performance in recent years [7]–[11], due to the powerful modeling ability of deep neural networks (DNNs). (3) Workshop on fine-grained visual categorization[1] has been organized at the conference of computer vision and pattern recognition (CVPR) every two years since 2011, which promotes the development of fine-grained image classification. (4) Some information technology companies, such as Microsoft and Baidu, begin to turn fine-grained image classification technologies into their applications[2] [3].

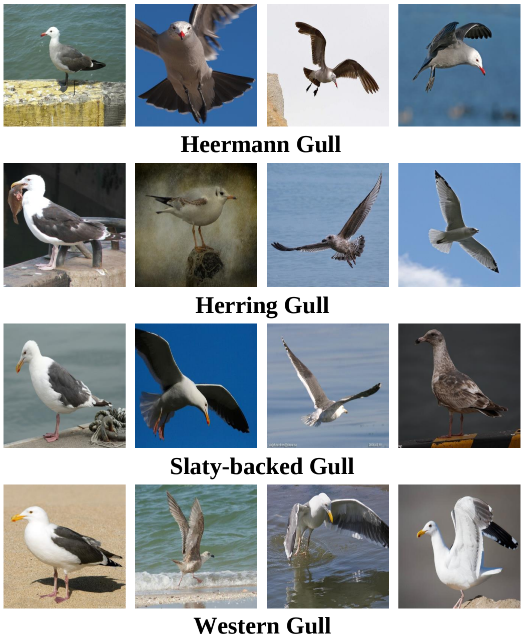


Fig. 1: Examples from CUB-200-2011 dataset [1]. Note that fine-grained image classification is a technically challenging task even for humans to recognize these subcategories, due to large variances in the same subcategory and small variances among different subcategories.

Fine-grained image classification lies in the continuum between basic-level image classification (e.g. object recognition) and identification of individuals (e.g. face recognition). The main challenges of fine-grained image classification are summarized as the two following aspects: (1) Variances among similar subcategories are subtle and local, because they belong to the same genus. (2) Variances in the same subcategory are large and diverse, due to different poses and views, and for animals or plants also because of different living

This work was supported by National Natural Science Foundation of China under Grants 61771025 and 61532005.

The authors are with the Institute of Computer Science and Technology, Peking University, Beijing 100871, China. Corresponding author: Yuxin Peng (e-mail: pengyuxin@pku.edu.cn).

[1] https://sites.google.com/view/fgvc4/home/

[2] https://www.microsoft.com/en-us/research/project/flowerreco-cn/

[3] http://image.baidu.com/?fr=shitu/

environments and growth periods. For example, as shown in Fig. 1, “Heermann Gull” and “Herring Gull” look similar in global appearance, but “Herring Gulls” look different in the pose, view and feather color. So it is hard for a person without professional knowledge to classify them. Consequently, fine-grained image classification is difficult to address with today’s general-purpose object recognition machinery.

The subcategories look generally the same in global appearance, and distinguished by the subtle and local variances, such as the color of abdomen, the shape of toe and the texture of feather for bird. These subtle variances are located at the discriminative parts of object, so the localization of object and its discriminative parts is crucial for fine-grained image classification. Researchers generally adopt a two-stage learning framework: the first stage is to localize the object or its discriminative parts, and the second is to extract the deep features of the object or its parts through Convolutional Neural Network (CNN) and train classifiers for the final prediction. Zhang et al. [12] utilize R-CNN [13] with geometric constraints to detect object and its parts first, then extract features for the object and its parts, finally train one-versus-all linear SVMs for classification. Simon and Ronder [14] propose a constellation model to localize parts of objects, which utilizes CNN to find the constellations of neural activation patterns. Then part detectors are trained for localizing discriminative parts for better classification. However, not all parts are beneficial and indispensable for distinguishing among subcategories. The conclusive distinctions among subcategories generally locate at a few specific parts, such as red beak or black tail. So the classification performance depends on the number of part detectors and whether the detected parts are discriminative or not. However, mainstream methods generally set the detector number due to their prior knowledge and the experimental results, which is highly empirical and limited especially in practical applications. Huang et al. [15] show that the classification accuracy declines when the number of parts increases from 8 to 15 in the experiments of their Part-stacked CNN method. Zhang et al. [16] pick 6 parts for CUB-200-2011 dataset [1] and 5 parts for Stanford Dog dataset [2] in the experiments for achieving the best classification accuracy. He and Peng [17] only use 2 discriminative parts for classification in their method. This is highly limited in flexibility, and difficult for generalizing to the other datasets or domains.

Therefore, it is significant to automatically learn and mine how many and which parts really make sense to fine-grained image classification. When a person see two images from two different subcategories, human visual attention mechanism allows him to focus on the pivotal distinctions between them. Attention is the behavioral and cognitive process of selectively concentrating on a discrete aspect of information, whether deemed subjective or objective, while ignoring other perceivable information. Inspired by this, researchers begin to apply human visual attention mechanism in their works, aiming to find the most discriminative characteristics for classification. Xiao et al. [18] propose a two-level attention model: object-level attention is to select relevant region proposals to a certain object, and part-level attention is to localize discriminative parts of object. Fu et al. [19] propose a recurrent attention convolutional neural network (RA-CNN), recursively learning discriminative region attention and region-based feature representation. These works simulate human visual attention mechanism to find discriminative parts for classification only from visual information.

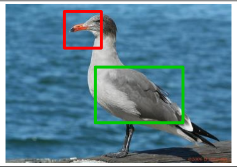
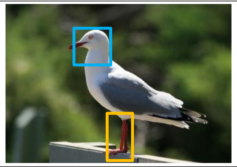
Subcategory	Vision	Text
Heermann Gull		(1)A large bird with different shades of grey all over its body, white and black tail feathers, and a long sharp orange beak (2)This bird is gray and black in color, with a orange beak (3)This bird has black outer retices and white inner retires and an orange beak ...
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)This bird has a white head, nape, breast and belly, with gray wings, the underside of which are black-tipped. ...

Fig. 2: Examples of attention for vision and text. The images in the column of vision come from CUB-200-2011 dataset [1], and natural language descriptions in the column of text are collected by Reed et al. [20] through Amazon Mechanical Turk (AMT) platform.

As is known to all, when human beings give the interpretation of the visual data into textual information via natural language descriptions, they have pointed out how many and which parts really make sense to classifying the visual data, such as an image of “Heermann Gull”. This is an involuntary transfer from human visual attention to textual attention. In this transfer process, common characteristics of object and background areas of images are eliminated naturally. As shown in Fig. 2, images and their descriptive texts are shown. From the image of “Red Legged Kittiwake”, we can find that the most discriminative parts to distinguish it from “Heermann Gull” are “red feet” and “white head”. Therefore, when we describe this image, the above two discriminative parts will occur in the texts. Comparing with part locations and attributes annotations, textual information provides more pivotal and fine-grained visual descriptions. Part locations could not point out which parts are the most discriminative parts for classification or tell the discriminative characteristics, such as the color of bill and the shape of wing. Attribute annotation provides the information of whether contains a certain part, e.g. belly, and what color the part is, but they do not provide the location information of parts in images.

Therefore, how to exactly relate textual attention to visual attention and mine the discriminative part are pivotal to fine-grained image classification. This paper proposes a visual-textual attention driven fine-grained representation learning (VTA) approach, and its main contributions are:

- **Fine-grained visual-textual pattern mining** devotes to discovering discriminative visual-textual parts for classification through jointly modeling vision and text with generative adversarial networks (GANs). Different from existing methods, the localized discriminative parts in this paper could not only tell us how many and which parts are significant for classification, but also which attributes of parts distinguish this subcategory from others. The parts number is determined automatically and adaptively by textual attention information.

- **Visual-textual representation learning** is proposed, which jointly combines visual and textual information. Visual stream focuses on the locations of the discriminative parts, while textual stream focuses on the discrimination of the regions. It preserves the intra-modality and inter-modality information to generate complementary fine-grained representation, and further improve classification accuracy.

Our previous conference paper CVL [11] proposes a two-stream model combing vision and language for learning fine-grained representations. Vision stream learns deep representations from visual information and language stream utilizes textual information to encode salient visual aspects for distinguishing subcategories. The main differences between the proposed VTA approach and CVL can be summarized as the following two aspects: (1) Our VTA approach employs textual pattern mining to localize textual attention for exploiting the human visual attention transfered into textual information, which points out how many and which parts are significant and indispensable for classification. While CVL directly utilizes the whole textual information, do not mine fine-grained textual attention information. (2) Our VTA approach employs visual pattern mining based on discovered textual patterns to localize discriminative parts, so that discriminative parts and objects are both exploited to learn multi-grained and multi-level representations and boost fine-grained classification. While CVL only exploits objects, which ignores complementary and semantic fine-grained clues provided by the discriminative parts. (3) Our VTA approach employs fine-grained visual-textual pattern mining to discover the discriminative and significant visual-textual pairwise information via jointly modeling vision and text with GANs, which mines the correlation between textual and visual attention. While CVL combines vision and text, ignoring to exploit the visual and textual attention which is significant to classification. Comparing with state-of-the-art methods on two widely-used fine-grained image classification datasets, the effectiveness of our VTA approach is verified by the comprehensive experimental results.

The rest of this paper is organized as follows: Section II briefly reviews related works on fine-grained image classification, frequent pattern mining and multi-modal analysis. Section III presents our proposed VTA approach, and Section IV introduces the experiments as well as the results analyses. Finally Section V concludes this paper.

II. Related Work

A. Fine-grained Image Classification

Most existing methods follow the pipeline: first localize the object or its parts, and then extract discriminative features for fine-grained image classification. An intuitive idea is directly using the annotations for the locations of object and its parts. For example, CUB-200-2011 [1] provides the detailed annotations: object annotation (i.e. bounding box of object) and parts annotations (i.e. 15 part locations). Object annotation is used in the works of [21], [22] to learn part detectors in an unsupervised or latent manner. And even part annotations are used in these methods [23], [24]. Since the annotations of the testing image are not available in practical applications, some researchers use the object or part annotations only at training phase and no knowledges of any annotations at testing phase. Object and part annotations are directly used in training phase to learn a strongly supervised deformable part-based model [25] or directly used to fine-tune the pre-trained CNN model [26]. Further more, Krause et al. [27] only use object annotation at training phase to learn the part detectors, then localize the parts automatically in the testing phase. Recently, there are some promising works attempting to learn the part detectors under the weakly supervised condition, which means that neither object nor part annotations are used in both training and testing phases. These works make it possible to put the fine-grained image classification into practical applications. Simon et al. [14] propose a neural activation constellations part model (NAC) to localize parts with constellation model. Xiao et al. [18] propose a two-level attention model, which combines two level attentions to select relevant proposals to the object and the discriminative parts. Zhang et al. [16] incorporate deep convolutional filters for both part detection and description. The aforementioned methods mostly set the detector number due to the prior knowledge and the experimental results, which is highly limited in flexibility and difficult for generalizing to the other domains. Therefore, we attempt to automatically learn and mine how many and which parts really make sense to classification via fine-grained visual-textual pattern mining.

B. Frequent Pattern Mining

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold [28]. For example, a set of items, such as diaper and beer, that appear frequently together in sales data of a supermarket, is a frequent pattern. Frequent pattern mining is first proposed by Agrawal et al. [29] for market basket analysis. Agrawal and Strikant propose Apriori algorithm [30] to mine frequent patterns in a large transaction database. For textual mining, frequent patterns may be sequential patterns, frequent itemsets, or multiple grams, which can represent the textual information. While for visual mining, frequent patterns may be middle-level feature representation or high-level semantic representation, which can describe the content of visual information. Han et al. [31] propose to mine visual patterns from images using low-level features. Li et al. [32] propose to combine CNN features and association rule mining for discovering visual patterns. Li et al. [33] propose a novel multi-modal pattern mining method, which takes textual pattern and visual pattern into consideration at the same time. In this paper, we first utilize Aprior algorithm to discover the textual patterns, and then employ generative adversarial networks (GANs) to mine the relationships between part

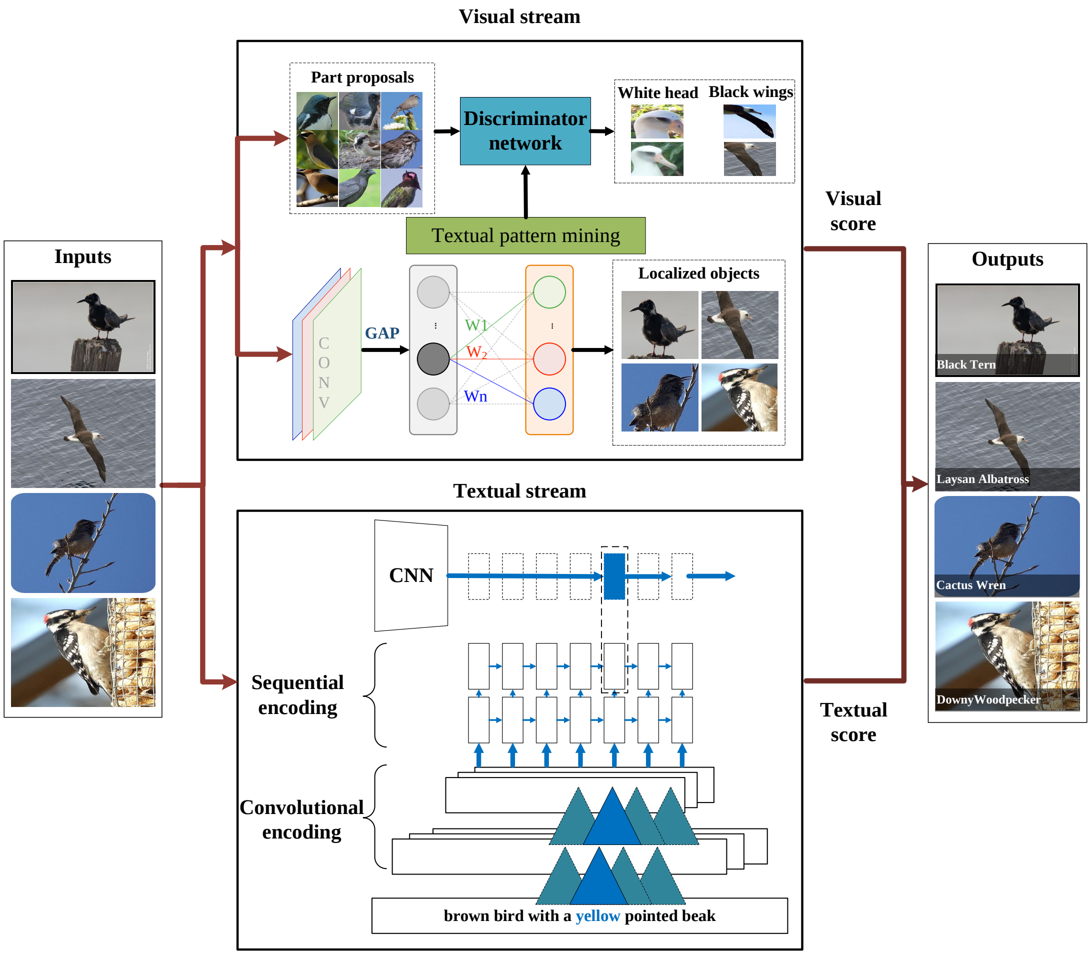


Fig. 3: Overview of our VTA approach.

proposals and textual patterns for better classification accuracy, which discovers visual and textual patterns at the same time as well as mines the intrinsic correlation between them.

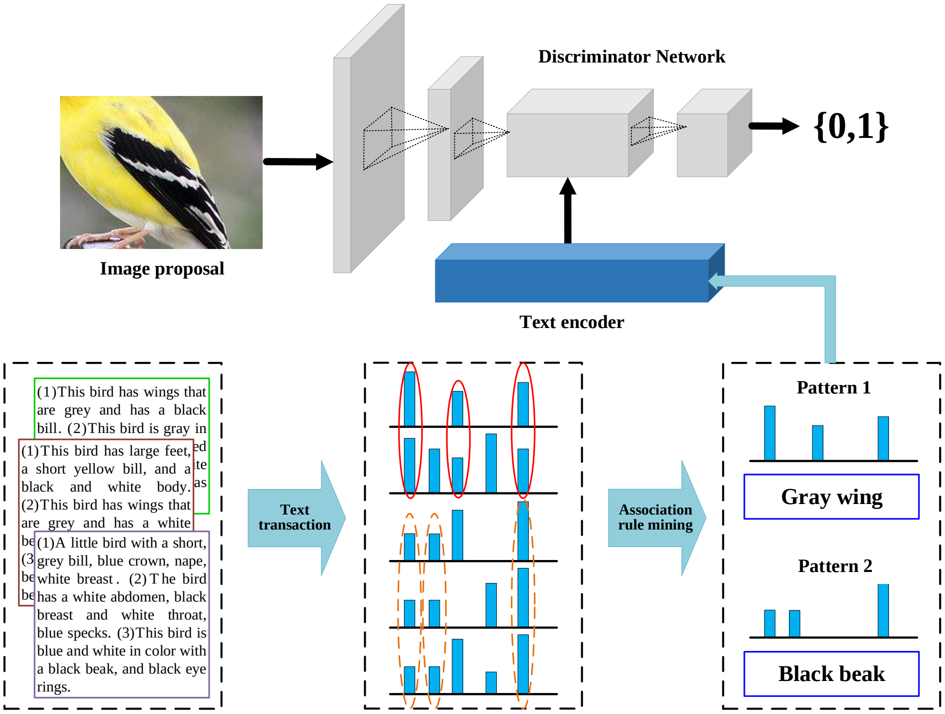
C. *Multi-modal Analysis*

With the rapid growth of multi-modal data, e.g. image, text, video and audio, has been the main form of the big data. Multi-modal data carries different kinds of information, which needs to be integrated to get comprehensive results in many real-world applications. How to learn multi-modal representation is a fundamental research problem. A traditional representative method is the canonical correlation analysis (CCA) [34], which learns a subspace to maximize the correlation among data of different media types, and is widely used for modeling multi-modal data [35]–[37]. Zhai et al. [38] propose to learn projection functions by the metric learning, and this method is further improved as joint representation learning (JRL) [39] by adding other information such as semantic categories and semi-supervised information. Inspired by the progress of deep neural networks, some works begin to focus on deep multi-modal representation learning. Ngiam et al. [40] propose a multi-modal deep learning (MDL) method to combine the audio and video into an autoencoder, which improves the speech signal classification for noisy inputs as well as learns a shared representation across modalities. Recently, a surge of progress has been made in image and video captioning. Long Short-Term Memory (LSTM) [41] is widely used in modeling captions at word level. Besides LSTM, character-based convolutional networks [42] have been used for text modeling. In this paper, we apply the extension of Convolutional and Recurrent Networks (CNN-RNN) to learn a visual semantic embedding. In this paper, we bring the multi-modal representation learning into fine-grained image classification to boost the performance, and jointly modeling vision and text.

III. Our Visual-Textual Attention Driven Fine-Grained Representation Learning Approach

A. *Overview of Our VTA Approach*

Our method is based on a very promising and interesting intuition: natural language descriptions (text) can point out the discriminative parts or characteristics of images (vision), and are complementary with visual information. Therefore, we propose a visual-textual attention driven fine-grained representation learning (VTA) approach, which takes the advantages of vision and text jointly as well as exploits the intrinsic correlation between them. Fig. 3 shows our VTA approach.



(1)This bird has wings that are grey and has a black bill. (2)This bird is gray in

(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

(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.

Fig. 4: Overview of our fine-grained visual-textual pattern mining approach.

First, we conduct fine-grained visual-textual pattern mining to discover the discriminative vision-text parts for boosting classification. Then, we localize the object region of image for boosting the visual analysis. Final, we propose a visual-textual representation learning approach to model visual and textual information jointly for better classification accuracy.

B. Fine-grained Visual-textual Pattern Mining

Since human visual attention is described into the form of text, we first conduct textual pattern mining to localize the textual attention, which points out the discriminative parts, and then based on the textual pattern find the discriminative visual information with generative adversarial networks (GANs). In the following paragraphs, we will describe our fine-grained visual-textual pattern mining approach from three aspects: 1) definition of pattern mining, 2) textual pattern mining and 3) visual pattern mining via GANs. The overview of our approach is shown in Fig. 4.

1) Definition of Pattern Mining: For describing our fine-grained visual-textual pattern mining approach clearly, we first introduce the basic definitions for pattern mining. Assume that there is a set of n items, which is denoted as $X = \{x_1, x_2, ..., x_n\}$, and transaction T is a subset of X, which means $T \subseteq X$. We also define a transaction database $D = \{T_1, T_2, ..., T_K\}$ that contains K transactions. Our goal is to discover a particular subset T^* of transactions database X, which can predict the presence of some target item $y \in T_y$, and $T^* \subset T_y$ as well as $y \cap T^* = \emptyset$. T^* is referred to frequent itemset in pattern mining literature. The support of T^* denotes how often T^* appears in D and its definition is as follow:

$$supp(T^*) = \frac{|\{T_y | T^* \subseteq T_y, T_y \in D\}|}{K} \tag{1}$$

An association rule $T^* \rightarrow y$ defines a relationship between T^* and a certain item y. Therefore, we aim to find patterns that appears in a transaction there is a high likelihood that y. We define the confidence as follow:

$$conf(T^* \rightarrow y) = \frac{supp(T^* \cup y)}{supp(T^*)} \tag{2}$$

2) Textual Pattern Mining: In this paper, we devote to discovering textual pattern, which contains the human visual attention informations. Textual transaction is necessary for pattern mining algorithm.

First, we remove stop words and punctuations from each textual descriptions for each image. Then we select the words, which appear in at least 10 captions in our dataset, as vocabulary which is used for generating transactions. It is noted that each word do not have duplicate in the vocabulary. In order to generate transaction for each textual description, we map each word back to its corresponding word in the vocabulary, then include that corresponding word index in the transaction. After obtaining the transactions, we perform association rule mining to find the words that frequently appear in textual descriptions, which also means that these words can represent the characteristics of this subcategory.

The transactions generated through the above processing form the transaction database D, we utilize the Aprior algorithm [30] to find a set of patterns P through association rule mining. Each patter $p \in P$ must satisfy the following criteria:

$$supp(p) > supp_{min} \tag{3}$$

$$conf(p \rightarrow c) > conf_{min} \tag{4}$$

where $supp_{min}$ and $conf_{min}$ are thresholds for the support value and confidence value, and c means the subcategory label. After association rule mining, each discovered pattern p contains a set of words.

We want to find some patterns that points out the discriminative parts of the image, which has the semantic meaning. Therefore, we conduct distance constraint on association rule mining as follows:

$$supp(p) > supp_{min} \tag{5}$$

$$conf(p \rightarrow c) > conf_{min} \tag{6}$$

$$dis(w_i, w_j) < dis_{min} \tag{7}$$

where w_i means the i-th word in pattern p, w_j means the j-th word in pattern p and $dis(\cdot)$ means the distance between i-th word and j-th word in the same textual description, which make sure that the discovered patterns have the semantic meanings. Final, we discover a set of patterns P, which mean the textual attention in the textual descriptions that contains the information of human visual attention.

3) Visual Pattern Mining via GANs: After obtaining the textual attention, we devote to mining the relationship between visual and textual attention, meaning that localize the discriminative parts of images via the guidance of textual attention. Due to the great progress made by generative adversarial networks (GANs), which can generate images based on textual information. In this paper, we employ GAN-CLS [43] to break through the gap between visual and textual information, localize the discriminative parts which correspond to the discovered textual patterns. The network architecture and training strategy follow Reed et al. [43]. In the following paragraphs, we introduce our visual pattern mining approach.

First, for each image we perform bottom-up process to generate part proposals. In this paper, we utilize selective search method [44] to generate 1000 part proposals for each image. Then we take the part proposals and discovered textual patterns as the inputs of discriminator network, which is utilized to relate the discovered textual patterns with the corresponding part proposals. For each part proposal, discriminator network outputs score vector that refers to the degree of correlations between part proposal and textual patterns. We select the part proposal with highest score for each textual pattern, which is one of the most discriminative parts in the images and beneficial for classification. They will be utilized as the inputs of visual-textual representation learning.

C. Object Localization

For better classification performance, we apply an automatic object localization method based on CAM [45]to localize the object in a weakly supervised manner, which means that neither object nor part annotations are used in both training and testing phases. Through CAM, we can generate a subcategory activation map M_c for each subcategory c, in which the spatial value is calculated as follow:

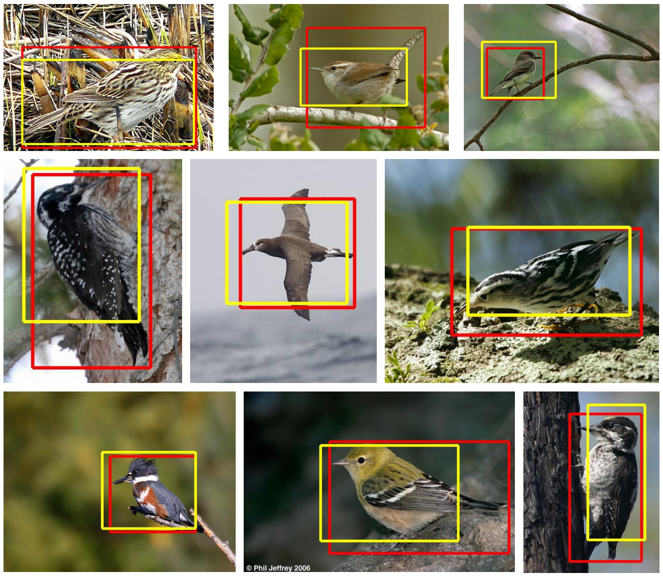


Fig. 5: Examples of object localization results in this paper. The red rectangles indicate the ground truth object annotations, i.e. bounding boxes of objects, and the yellow rectangles indicate the object regions generated by our approach.

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \tag{8}$$

where $f_k(x, y)$ denotes the activation of unit k in the last convolutional layer at spatial location (x, y), and w_k^c is the weight corresponding to subcategory c for unit k. The subcategory label information is not available in testing phase, so we set subcategory c by the predicted subcategory. After obtaining the activation map for each image, we conduct OTSU algorithm [46] to binarize the image and take the bounding box that covers the largest connected area as the localization of object. The localized object is utilized as the inputs of visual-textual representation learning along with the localized discriminative parts via fine-grained visual-textual pattern mining. Examples of object localization results are shown in Fig. 5.

D. Visual-textual Representation Learning

Considering that two different forms, i.e. visual and textual information, are complementary with each other, we jointly model them with a two-stream model to learn deep representations for better classification accuracy. The two-stream model consists of: 1) visual stream and 2) textual stream.

1) Visual Stream: A natural candidate for the visual classification function f is a CNN model, which consists of a hierarchy of convolutional and fully connected layers. We can benefit from model pre-training due to the additional training data, such as 1.3M training images of ImageNet 1K dataset [47]. This has been proved by a large amount of computer vision tasks, such as object detection [13], object segmentation [48] and fine-grained image classification [13], [49]–[51]. Therefore, we use a CNN model pre-trained on the ImageNet 1 K dataset [47] as the basic model in our experiments. Then we fine-tune the pre-trained CNN model on the fine-grained dataset.

For a given image I, its object region b and n discriminative parts $Pa = \{Pa_1, Pa_2, ..., Pa_n\}$ are automatically generated at object localization stage and fine-grained visual-textual pattern mining stage respectively, then the object region and parts regions are clipped from the original image and saved as images I_b and $I_{Pa} = \{I_{Pa_1}, I_{Pa_2}, ..., I_{Pa_n}\}$. We feed the original image I and its object image I_b as well as its part images $I_{Pa} = \{I_{Pa_1}, I_{Pa_2}, ..., I_{Pa_n}\}$ to the CNN model to obtain the predictions. For the predictions of parts images, we calculate their mean value as the final part prediction. Then we calculate the weighted mean of original prediction, object prediction and part prediction to obtain the result of the visual stream, whose weights are set by cross-validation method.

2) Textual Stream: In textual stream, we aim to measure the similarity between visual and textual information. We first apply the deep structured joint embedding method [20] to jointly embed vision (i.e. images) and text (i.e. natural language descriptions for images). It learns a compatibility function of vision and text, which can be seen as an extension of the multi-modal structured jointed embedding [52]. Instead of using a bilinear compatibility function, we use the inner product of features generated by deep neural encoders, and maximize the compatibility between a description and its matching image as well as minimize compatibility with images from other classes.

Given data $D = (v_n, t_n, y_n), n = 1, ..., N$, in which $v \in V$ indicates visual information, $t \in T$ indicates textual information and $y \in Y$ indicates the subcategory label, then the visual and textual classifier functions $f_v : V \to Y$ and $f_t : T \to Y$ are learned by minimizing the empirical risk:

$$\frac{1}{N}\sum_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)) \tag{9}$$

where $\Delta : y \times y \to \mathbb{R}$ is the 0-1 loss and

$$f_v(v) = arg \max_{y \in Y} \mathbb{E}_{t \sim T(y)}[F(v, t)] \tag{10}$$

$$f_t(t) = arg \max_{y \in Y} \mathbb{E}_{v \sim V(y)}[F(v, t)] \tag{11}$$

We then define the compatibility function $F : V \times Y \to \mathbb{R}$ that uses features from the learnable encoder functions $\theta(v)$ for vision and $\phi(t)$ for text:

$$F(v, t) = \theta(v)^T \phi(t) \tag{12}$$

We apply the GoogleNet [53] as vision encoder model, and Convolutional Recurrent Net (CNN-RNN) [20] as the text encoder model which will be discussed in the next paragraph.

A mid-level temporal CNN hidden layer is at the bottom of CNN-RNN model, and a recurrent network is stacked on it. We extract the average hidden unit activation over the sequence as the textual feature, as shown in Equation 13. The score function is defined as a linear accumulation of evidence for compatibility with the image which needs to be recognized.

$$\phi(t) = \frac{1}{L}\sum_{i=1}^{L} h_i \tag{13}$$

where h_i indicates the hidden activation vector for the i-th frame and L indicates the sequence length.

E. Final Prediction

Given an Image I, the two-stream model conducts on the original images and their object localizations as well as their part localizations. The visual stream gives the prediction from the view of the visual information, while the textual stream gives the prediction via measuring the visual and textual information with the shared compatibility function. Due to the fact that joint learning of visual and textual information preserves the intra-modality and inter-modality information to generate complementary information, we fuse the predicted results of the two streams as the final prediction to utilize the advantages of the two via the follow equation:

$$f(I) = f_v(v) + \beta * f_t(t) \tag{14}$$

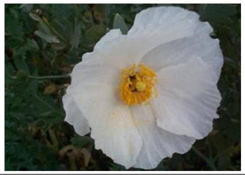
where $f_v(v)$ and $f_t(t)$ are the visual and textual classifier functions as mentioned above, and β is selected by the cross-validation method.

IV. Experiments

A. Datasets

This subsection presents two fine-grained image classification datasets adopted in the experiments, including CUB-200-2011 and Oxford Flowers-102 datasets, and their detailed information is described as follows:

- **CUB-200-2011**. It is the most widely-used dataset for fine-grained image classification task. The visual information comes from the original dataset of CUB-200-2011 [1]. It contains 11,788 images of 200 subcategories belonging to birds, 5,994 for training and 5,794 for testing. Each image has detailed annotations: 1 subcategory label, 15 part locations, 312 binary attributes and 1 bounding box. The textual information comes from Reed et al. [20]. They expand the CUB-200-2011 dataset by collecting fine-grained natural language descriptions. Ten single-sentence descriptions are collected for each image, as shown in Fig. 6. The natural language descriptions are collected through the Amazon Mechanical Turk (AMT) platform, and are required at least 10 words, without any information of subcategories and actions.
- **Oxford Flowers-102**. The same with CUB-200-2011 dataset, textual information comes from Reed et al. [20], and visual information comes from the original dataset of Oxford Flowers-102 [4], as shown in Fig. 6. It has 8,189 images of 102 subcategories belonging to flowers, 1,020 for training, 1,020 for validation and 6,149 for testing. Each subcategory consists of between 40 and 258 images.

Subcategory	Vision	Text	Subcategory	Vision	Text
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. ...	Primula		(1)This flower is white and yellow in color, with petals that are heart shaped. (2)This white color flower has the simple row of heart shaped petals shaded with orange color at the end. (3)This flower has thick heart shaped white petals and a very yellow star shaped center. ...
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. ...	Silverbush		(1)The flower has petals that are white with yellow centers. (2)This flower has large, flat white petals that connect to each other and have a yellow center. (3)This flower is white and yellow in color, with petals that are connected to each other. ...
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 ...	Tree Poppy		(1)This flower has a yellow center surrounded by several large, overlapping white petals with ruffled edges. (2)This flower is white and yellow in color, with petals that are ruffled on the edges. (3)This pretty little flower has large white petals and a yellow center ...

CUB-200-2011 **Oxford Flowers-102**

Fig. 6: Some examples of vision and text in CUB-200-2011 dataset and Oxford Flowers-102 dataset.

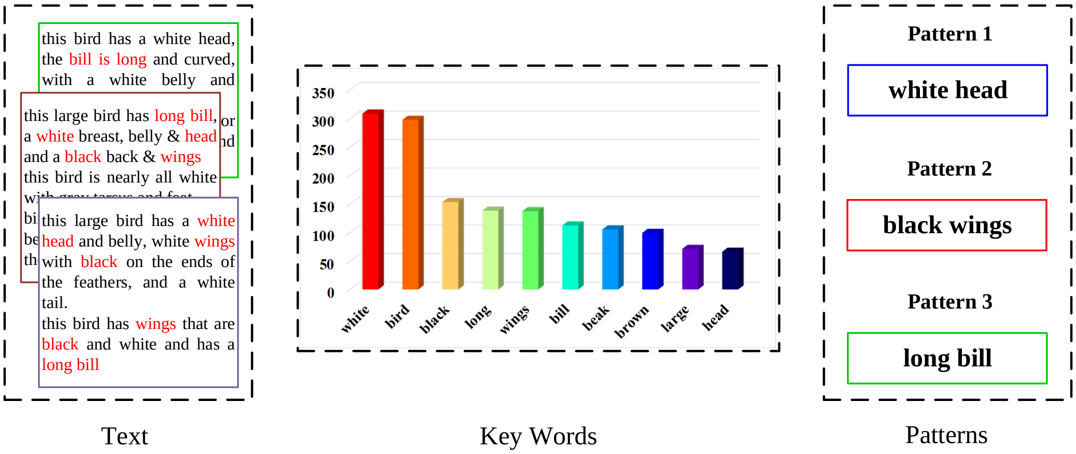

this bird has a white head, the bill is long and curved, with a white belly and
this large bird has long bill, a white breast, belly & head and a black back & wings this bird is nearly all white
this large bird has a white head and belly, white wings with black on the ends of the feathers, and a white tail.
this bird has wings that are black and white and has a long bill
350
300
250
200
150
100
50
0
white
bird
black
long
wings
bill
beak
brown
large
head
Pattern 1
white head
Pattern 2
black wings
Pattern 3
long bill
Text
Key Words
Patterns

Fig. 7: Examples of textual pattern mining.

B. Evaluation Metric

Accuracy is adopted to comprehensively evaluate the classification performances of our VTA approach as well as compared state-of-the-art methods, which is widely used in fine-grained image classification [8], [12], and its definition is as follow:

$$Accuracy = \frac{R_a}{R} \tag{15}$$

where R denotes the number of images in testing set, and R_a denotes the number of images that are correctly classified.

C. Implementation Details

Fine-grained Visual-textual Pattern Mining. First, we calculate the frequencies of words in the textual information for each subcategory, and select the top-10 words as keywords, and then discover frequent patterns from textual informations via Aprior algorithm [30]. From this phase, we obtain textual patterns that could describe the characteristics of object, such as "white head", "black wings" and "long bill", as shown in Fig. 7. Second, we conduct selective search [44] on each image to generate image proposals. Finally, we employ discriminator network to relate textual patterns to image proposals, then select the proposal with highest score as the discriminative part for each textual pattern. For each subcategory, the number of parts is set automatically and adaptively by discovered textual patterns. The selected parts can be seen in Fig. 8.

Visual-textual Representation learning. For textual stream, we apply GoogleNet [53] with batch normalization [54] as vision encoder and CNN-RNN [20] as text encoder. All the

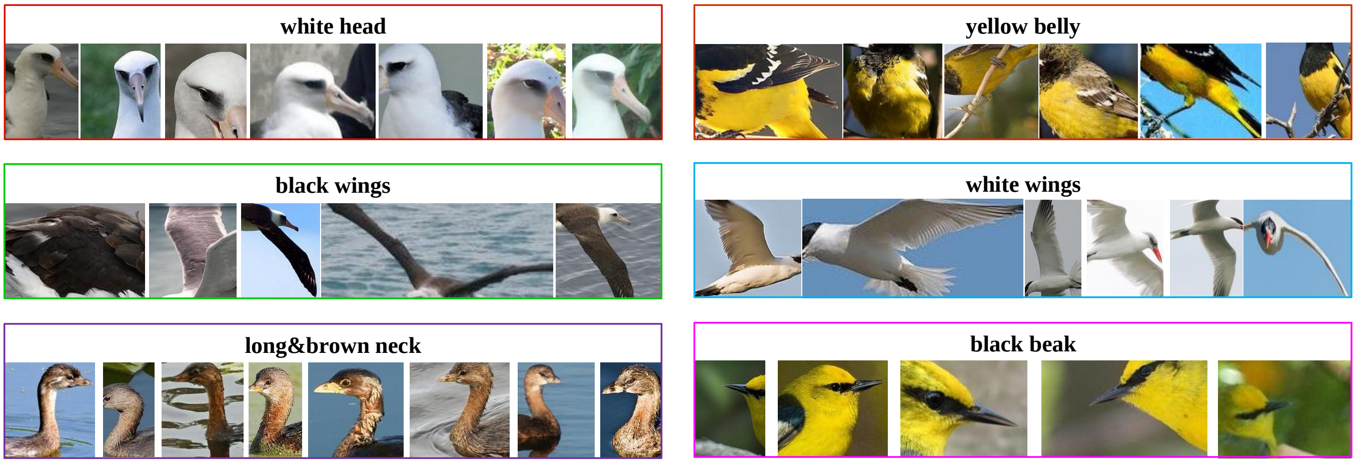


Fig. 8: Examples of discovered visual-textual patterns by our fine-grained visual-textual pattern mining approach.

configurations and source codes [4] used for training and testing follow the work of Scott Reed et al. [20]. For visual stream, we apply the widely-used model of 19-layers VGGNet [55] with batch normalization. The model is first pre-trained on ImageNet 1K dataset, and then fine-tuned on fine-grained dataset. Inspired by the strategy adopted by Xiao et al [18], we utilize the pre-trained CNN model as a filter net to select proposals relevant to the object from the generated image proposals by selective search method. With the selected proposals, we further fine-tune the pre-trained model.

D. Comparisons with state-of-the-art methods

In this subsection, we present the experimental results on our proposed approach as well as all the compared methods. Tables I and II show the accuracies on two datasets, and demonstrate that our proposed approach achieves the best classification accuracy compared with state-of-the-art methods on both two datasets. As shown in Table I, our proposed approach has improved the classification accuracy from 85.30% to 86.31% on CUB-200-2011 dataset. We divide the compared methods into three groups due to the amount of object and parts annotations used in these methods.

- *Neither object nor parts annotations are used.* Nowadays, researchers begin to focus on how to get better classification accuracy under the weakly supervised setting, which means neither object nor parts annotations are used. Most of these methods utilize the attention property of convolutional neural layers to localize the discriminative parts of object for better accuracy, such as RA-CNN [19], PNA [8], TSC [17] and TL Atten [18]. They simulate human visual attention mechanism only from visual information. In our approach, we exploit visual and textual attention simultaneously as well as mine the complementary between them, which make our proposed approach more effective and obtain a 1.01% higher accuracy than the best performing result of RA-CNN [19].
- *Only one of object and parts annotations is used.* These methods utilize object annotation (i.e. bounding box) to train an object detector or learn parts detectors, which is to learn more representative features for classification. In our approach, we utilize CAM [45] to automatically localize the object region of image, which avoids using object annotation, and the result of object localization can be seen in Fig. 2. Even using object annotation, these methods achieve lower accuracies than our proposed approach.
- *Both object and parts annotations are used.* In order to obtain better classification accuracy, some methods utilize both object and parts annotations at training phase as well as testing phase. However, these annotations are heavy labor-consuming. In our approach, we get object region and discriminative parts automatically via object localization and fine-grained visual-textual pattern mining respectively without using any annotations. We promote the classification performance through discovering the discriminative and representative object and its parts.

Besides, classification results on Oxford Flowers-102 dataset are shown in Table II, which also have the similar trends as CUB-200-2011 dataset, while our proposed VTA approach still keeps the best.

E. Performances of components in our VTA approach

In this subsection, we conduct two baseline experiments as follows to verify the separate contribution of each component in our proposed VTA approach. Tables III to V show the accuracies of our proposed VTA approach as well as the baseline approaches on CUB-200-2011 dataset at the following two aspects.

1) Effectivenesses of Fine-grained Visual-textual Pattern Mining and Object Localization: In our VTA approach, fine-grained visual-textual pattern mining and object localization generate discriminative parts and object for promoting the classification accuracy. They make sense to the visual stream and then further impact whole approach. Therefore, Tables

[4] https://github.com/reedscot/cvpr2016

Method	Train Annotation		Test Annotation		Accuracy (%)
	Bbox	Parts	Bbox	Parts	
Our VTA Approach					**86.31**
RA-CNN [19]					85.30
PNA [8]					84.70
TSC [17]					84.69
FOAF [56]					84.63
Low-rank Bilinear [57]					84.21
Spatial Transformer [58]					84.10
Bilinear-CNN [59]					84.10
Multi-grained [60]					81.70
NAC [14]					81.01
PIR [61]					79.34
TL Atten [18]					77.90
MIL [62]					77.40
VGG-BGLm [63]					75.90
Dense Graph Mining [64]					60.19
Coarse-to-Fine [65]	√				82.50
PG Alignment [27]	√		√		82.80
Triplet-A (64) [66]	√		√		80.70
Webly-supervised [67]	√	√			78.60
PN-CNN [26]	√	√			75.70
Part-based R-CNN [12]	√	√			73.50
SPDA-CNN [68]	√	√	√		85.14
Deep LAC [69]	√	√	√		84.10
PBC [70]	√	√	√		83.70
SPDA-CNN [71]	√	√	√		81.01
PS-CNN [15]	√	√	√		76.20
PN-CNN [26]	√	√	√	√	85.40

TABLE I: Comparisons with state-of-the-art methods on CUB-200-2011, sorted by amount of annotation used. “Bbox” and “Parts”indicate the object and parts annotations (i.e. bounding box and parts locations) provided by the dataset.

Method	Accuracy (%)
Our VTA Approach	**96.88**
PBC [70]	96.10
NAC [14]	95.34
RIIR [72]	94.01
Deep Optimized [73]	91.30
SDR [73]	90.50
MML [74]	89.45
CNN Feature [51]	86.80
Generalized Max Pooling [75]	84.60
Efficient Object Detection [3]	80.66

TABLE II: Comparisons with state-of-the-art methods on Oxford Flowers-102.

Method	Accuracy (%)
VTA-visual	**85.54**
VTA-visual(w/o object)	83.21
VTA-visual(w/o parts)	84.79
VTA-visual(w/o object&parts)	80.82

TABLE III: Effects of fine-grained pattern mining and object localization for visual stream.

Method	Accuracy (%)
Our VTA Approach	**86.31**
VTA(w/o object)	85.17
VTA(w/o parts)	85.83
VTA(w/o object&parts)	84.05

TABLE IV: Effects of fine-grained pattern mining and object localization for our proposed VTA approach.

Method	Accuracy (%)
Our VTA Approach	**86.31**
VTA-textual	81.81
VTA-visual	85.54
VTA(only original image)	80.82

TABLE V: Effects of different components of our proposed approach on CUB-200-2011.

III and IV show the effects of fine-grained visual-textual pattern mining and object localization to visual stream and our proposed VTA approach respectively. “object” means that object localization is conducted, and “parts” means that fine-grained visual-textual pattern mining is employed. We can observe that considering object localization can achieve better classification accuracy than considering fine-grained visual-textual pattern mining. This is because that objects contain the global and local features simultaneously, while discriminative parts focus subtle and local characteristics. However, jointly considering object localization and fine-grained visual-textual pattern mining can further improve the classification accuracy.

Fine-grained visual-textual pattern mining aims to select the image proposals that corresponding to the discovered textual patterns. The relations between image proposals and textual patterns ensure the discrimination and representativeness of selected parts. Some examples of discovered visual-textual patterns are shown in Fig. 8.

2) Effectiveness of Visual-textual Representation Learning: We also present the baseline experiments to verify the effectiveness of visual-textual representation. The results are shown in Table V, where “VTA-textual” means textual stream, “VTA-visual” means visual stream and “VTA(only original

Subcategory	Vision	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.

Fig. 9: Some results of the textual stream.

image" means only a fine-tuned CNN model is used. We can observe that classification result of textual stream is promising. From the first line of each row in Fig. 9, we can find that textual description with the highest score always points out the discriminative parts or characteristics of the object. The red words are the important textual descriptions for distinguishing subcategories, and blue ones are the descriptions of easily confused characteristics with other subcategories. Combing visual and textual information can further achieve more accurate classification result, which demonstrates that the two types of information are complementary: visual information focuses on the global and local features, and textual information further points the importance of these features.

From the above baseline results, the separate contribution of each component in our proposed VTA approach can be verified. First, object localization and fine-grained pattern mining discover the discriminative and representative information of image via visual-textual attention. Second, the complementarity between visual and textual information is fully captured by visual-textual representation learning.

V. Conclusions

In this paper, the visual-textual attention driven fine-grained representation learning approach has been proposed. Based on textual attention, we employ fine-grained visual-textual pattern mining to discover discriminative information for classification through jointly modeling vision and text with GANs. Then, visual-textual representation learning jointly considers visual and textual information, which preserves the intra-modality and inter-modality information to generate complementary fine-grained representation, and further improve classification performance. Experimental results on two widely-used fine-grained image classification datasets demonstrate the superiority of our method compared with state-of-the-art methods.

The results are promising, and point out a few future directions. First, combining visual and textual information can boost classification accuracy, but the two streams are trained respectively, we will focus on the work of training the two streams end-to-end. Second, we will exploit exact and effective methods on relating textual attention and visual attention for more accurate discriminative parts localization as well as better classification performance.

References

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

[2] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011.

[3] Anelia Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 811–818, 2013.

[4] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[5] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[6] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[7] Chen Huang, Zhihai He, Guitao Cao, and Wenming Cao. Task-driven progressive part localization for fine-grained object recognition. *IEEE Transactions on Multimedia (TMM)*, 18(12):2372–2383, 2016.

[8] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking neural activations for fine-grained recognition. *IEEE Transactions on Multimedia (TMM)*, 2017.

[9] Yan Wang, Sheng Li, and Alex C Kot. Deepbag: Recognizing handbag models. *IEEE Transactions on Multimedia (TMM)*, 17(11):2072–2083, 2015.

[10] Yan Wang, Sheng Li, and Alex C Kot. On branded handbag recognition. *IEEE Transactions on Multimedia (TMM)*, 18(9):1869–1881, 2016.

[11] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European Conference on Computer Vision (ECCV)*, pages 834–849, 2014.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[14] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1143–1151, 2015.

[15] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1173–1182, 2016.

[16] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1134–1142, 2016.

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

[18] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, 2015.

[19] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[20] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning deep representations of fine-grained visual descriptions. *arXiv preprint arXiv:1605.05395*, 2016.

[21] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 321–328, 2013.

[22] Shulin Yang, Liefeng Bo, Jue Wang, and Linda G Shapiro. Unsupervised template learning for fine-grained object recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3122–3130, 2012.

[23] Thomas Berg and Peter Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 955–962, 2013.

[24] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1641–1648, 2013.

[25] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 729–736, 2013.

[26] 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.

[27] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5546–5555, 2015.

[28] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[29] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[30] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *International Conference on Very Large Data Bases (VLDB)*, volume 1215, pages 487–499, 1994.

[31] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM, 2000.

[32] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Mid-level deep pattern mining. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 971–980, 2015.

[33] Hongzhi Li, Joseph G Ellis, Heng Ji, and Shih-Fu Chang. Event specific multimodal pattern mining for knowledge base construction. In *ACM on Multimedia Conference (ACM MM)*, pages 821–830. ACM, 2016.

[34] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[35] Hervé Bredin and Gérard Chollet. Audio-visual speech synchrony measure for talking-face identity verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages II–233, 2007.

[36] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[37] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4437–4446, 2015.

[38] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

[39] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 24(6):965–978, 2014.

[40] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, pages 689–696, 2011.

[41] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[42] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 649–657, 2015.

[43] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, pages 1060–1069, 2016.

[44] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013.

[45] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[46] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[49] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014.

[50] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)*, pages 647–655, 2014.

[51] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[52] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2927–2936, 2015.

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[54] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

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

[56] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, and Qi Tian. Fused one-vs-all features with semantic alignments for fine-grained visual categorization. *IEEE Transactions on Image Processing (TIP)*, 25(2):878–892, 2016.

[57] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[58] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2017–2025, 2015.

[59] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, 2015.

[60] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *International Conference on Computer Vision (ICCV)*, pages 2399–2406, 2015.

[61] Yu Zhang, Xiu-Shen Wei, Jianxin Wu, Jianfei Cai, Jiangbo Lu, Viet-Anh Nguyen, and Minh N Do. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing (TIP)*, 25(4):1713–1725, 2016.

[62] Zhe Xu, Dacheng Tao, Shaoli Huang, and Ya Zhang. Friend or foe: Fine-grained categorization with weak supervision. *IEEE Transactions on Image Processing (TIP)*, 26(1):135–146, 2017.

[63] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. *arXiv preprint arXiv:1512.02665*, 2015.

[64] Luming Zhang, Yang Yang, Meng Wang, Richang Hong, Liqiang Nie, and Xuelong Li. Detecting densely distributed graph patterns for fine-grained image categorization. *IEEE Transactions on Image Processing (TIP)*, 25(2):553–565, 2016.

[65] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Coarse-to-fine description for fine-grained visual categorization. *IEEE Transactions on Image Processing (TIP)*, 25(10):4858–4872, 2016.

[66] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. *arXiv preprint arXiv:1512.05227*, 2015.

[67] Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[68] 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. pages 1143–1152, 2016.

[69] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1666–1674, 2015.

[70] Chao Huang, Hongliang Li, Yurui Xie, Qingbo Wu, and Bing Luo. Pbc: Polygon-based classifier for fine-grained categorization. *IEEE Transactions on Multimedia (TMM)*, 19(4):673–684, 2017.

[71] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1143–1152, 2016.

[72] Lingxi Xie, Jingdong Wang, Weiyao Lin, Bo Zhang, and Qi Tian. Towards reversal-invariant image representation. *International Journal of Computer Vision (IJCV)*, 123(2):226–250, 2017.

[73] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. From generic to specific deep representations for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–45, 2015.

[74] Qi Qian, Rong Jin, Shenghuo Zhu, and Yuanqing Lin. Fine-grained visual categorization via multi-stage metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3716–3724, 2015.

[75] Naila Murray and Florent Perronnin. Generalized max pooling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2473–2480, 2014.