Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification

Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo

Abstract—Text in natural images contains rich semantics that are often highly relevant to objects or scenes. In this paper, we focus on the problem of fully exploiting scene text for visual understanding. The main idea is combining word representations and deep visual features into a globally trainable deep convolutional neural network. First, the recognized words are obtained by a scene text reading system. Then, we combine the word embedding of the recognized words and the deep visual features into a single representation, which is optimized by a convolutional neural network for fine-grained image classification. In our framework, the attention mechanism is adopted to reveal the relevance between each recognized word and the given image, which further enhances the recognition performance. We have performed experiments on two datasets: Con-Text dataset and Drink Bottle dataset, that are proposed for fine-grained classification of business places and drink bottles, respectively. The experimental results consistently demonstrate that the proposed method combining textual and visual cues significantly outperforms classification with only visual representations. Moreover, we have shown that the learned representation improves the retrieval performance on the drink bottle images by a large margin, making it potentially useful in product search.

Index Terms—Scene text, fine-grained classification, convolutional neural networks, attention mechanism, product search.

I. INTRODUCTION

Combining multi-modal features for the recognition or index of visual data is the inevitable way to automatically understanding image/video content in the recent years. The multi-modal visual data are often correlated and complementary to each other in visual understanding, especially when they originate from the same source. In this paper, we are concerned on how to efficiently integrate visual and textual cues for fine-grained image classification with convolutional neural networks. The goal of fine-grained classification is to assign the labels to the images having subtle differences in visual appearance such as animal species, product types, and place types. Such a problem is quite challenging, as relevant differences that are not obvious may not be accurately distinguished by a typical classification model. Even for human perception, domain specific knowledge is often required for fine-grained classification.

Scene text, which frequently appears in natural scenes including road signs, street views, product packages, and license plates, often possesses rich and precise meanings that are highly related to semantics of the object or scene in the same image. For instance, texts on the buildings or wrapping bags are quite useful for distinguishing among their categories. With the recent developments in text detection and recognition in the wild [1], [2], it has been shown that textual cues are very beneficial to fine-grained classification, especially in the classification of business places such as bakery, cafe and bookstore [3].

Scene text detection and recognition has been an active research area in both computer vision and document analysis communities. A large number of novel approaches are proposed for the localization and recognition of the text embodied in natural images and videos. Such approaches enable us to automatically extract the text information, which can be considered as an additional information for image classification. In this paper, the recognized words in natural scenes are used as a kind of input features for an image classification model. However, given an image, not every recognized word must have the relation to it. As shown in Figure 1, the recognized words such as ‘restaurant’, ‘bakery’, ‘pharmacy’, ‘bread’, ‘hotel’ are very relevant to the scene semantics, but some words like ‘grand’, ‘great’, ‘edison’ don’t have the direct connection to them. This requires a model to measure the relevance of each recognized word to the semantics of objects or scenes, in order to make full use of textual information in image classification. It should be mentioned that the existing robust reading system can not guarantee that all the words embodied in the images can be correctly detected and/or recognized. Yet, even under such a situation, the recognized words are still very helpful for image classification as demonstrated by our experiments.

To the best of our knowledge, there has been little research on combining textual and visual cues for fine-grained image classification. Therefore, it is imperative to develop such technologies for many fine-grained classification tasks such as the recognition or indexing of films, products, logos, places, athletes, advertisements, and so on, to fulfill the needs of many real-world applications. Since such scenarios often contain texts that are highly relevant and complementary to the visual appearances of objects or scenes, the text semantics are quite helpful for enhancing the performances of fine-grained classification.

Recently, deep neural networks have been the mainstream techniques for visual understanding, thanks to their great successes in various vision tasks. In this paper, we focus on designing a proper deep neural network architecture for fine-grained classification, which is able to efficiently fuse the
II. RELATED WORKS

A. Fine-grained Classification

Recently, fine-grained image classification is an active topic in some domains, such as animal species \[7\], plant species \[8\] and man-made objects \[9\]. Different from the base-class image classification, fine-grained image classification needs to distinguish images with more subtle differences among object classes. Discerning the subtle differences among similar classes is more challenging. Many recent works rely on relevant parts of fine-grained classes for domain specific knowledge. Examples are the recent works in \[10, 11\] that use salient parts in images. Semantic object parts are used in \[7\] to isolate the subtle differences. Xiao et al. \[12\] used three attention models to first propose candidate patches, then to localize discriminative parts by selecting relevant patches of a specific object. In \[13\], the authors proposed to select many useful parts from multi-scale part proposals of objects in each image without learning expensive object/part detectors. A weakly supervised strategy that using Multi-Instance Learning (MIL) to learn discriminative patterns for image representation is proposed in \[14\]. Zhang et al. \[15\] develop a pose-normalized representation for image patches to provide a high degree of pose invariance.

Besides these part-based methods, many other methods have been developed. For example, in \[16\], the authors adopted a deep ranking model to learn similarity metric directly from images. A multi-stage distance metric learning framework is proposed in \[17\] to learn a more discriminative similarity. Xie et al. \[18\] proposed a novel data augmentation approach to address the scarcity of data and a novel regularization technique to improve the generalization performance. In \[11\] and \[19\], the authors proposed to include humans in loop to select discriminative features.

The most related work with our proposal in this paper is the one in \[3, 20\], where scene texts in the business place images are considered as the domain knowledge. The authors proposed to combine visual and textual cues for fine-grained business place classification. More precisely, they train two one-versus-rest SVM classifiers on visual and respectively text features. Then these two trained SVM kernels are added together to form the final SVM kernel matrix. In this way, the visual and text cues are combined together. In \[20\], Bag-of-Words are used to represent visual cues. Deep visual cues given by GoogleNet \[5\] are used in \[3\]. Sharing the same spirit as part-based methods, the authors proposed in \[3\] to extract 100 EdgeBoxes \[21\]. Then the final visual feature is given by the combination of the GoogleNet visual features for the 100 boxed regions and the GoogleNet visual features for the whole image. In our work, we also propose to integrate both visual and text cues for fine-grained classification. We rely on an attention-based model to select the relevant words and get rid of those irrelevant ones. Hopefully, the combination of the attended text cues and visual feature of the whole image are more discriminative. Besides, as shown in Figure 2, our proposed model is end-to-end trainable, which is more elegant and efficient.

B. Scene text detection and recognition

Scene text detection and recognition have been extensively studied in computer vision and document analysis communities \[22–47\] (see \[1, 2\] for comprehensive surveys). Since...
the proposed method mainly leverages the recognized words to improve fine-grained classification, this section briefly revisits the most relevant works about end-to-end word recognition. In the detection stage, connected components [30], [32], [33] or sliding windows [31], [34] are usually used to extract character candidates. Then, the character candidate regions are further verified and grouped into word candidates based on some geometric rules or learning models. In the recognition stage, character-based methods [31], [32], [48]–[50] and word-based methods [6], [37], [51] are the two mainstream recognition approaches. Recently, following the popular general object detection approaches such as RCNN [52], Faster RCNN [53], YOLO [54], and SSD [55], the scheme of end-to-end word proposal generation followed by proposal classification has been a trend for word recognition in natural images [4], [37], [38], [44], [46]. In this paper, we adopt a recent algorithm for end-to-end word recognition named TextBoxes [4], which is the fastest among the existing text reading systems.

C. Multi-modal Fusion

In many tasks, information about the same phenomenon can be easily obtained from various sources, such as video, audio, images, and text. Generally, information from different modalities is complementary and offers useful knowledge to each other. Multi-modal approaches have shown superior performances over single-modal ones for many tasks [56]. For example, in [3], [20], the authors proposed to combine visual and textual cues to improve fine-grained classification of business place images. Similarly, Lu et al. [57] discussed some methods that combine visual cues and video captions to improve video classification. In [58], the authors presented a page classification application in a banking workflow by merging visual and textual descriptions.

The main challenge of multi-modal approaches lies on how to optimally combine the multi-modal information. The combination can take place at different levels. The most widely used strategy is to fuse the information at the feature level, which is known as early fusion. This kind of system merges the extracted features using integration mechanisms, such as concatenation [59], neural network architectures [60], etc. Another combination strategy is at decision level or late fusion [61], which merges multiple modalities in the semantic space. A hybrid combination of the early and late fusion is also proposed [62].

Following [3], [20], we also propose to combine visual and textual cues to improve fine-grained classification. We combine the visual cues and textual cues at the feature level.

D. Attention Mechanism

Attention mechanism is widely used in natural language processing, such as speech recognition [63], machine translation [64], and dialog management [65] to improve the network’s ability of finding relevant features from corresponding input parts. Recently, visual attention attracts much attention in computer vision society. Ba et al. [66] presented an attention-based model for recognizing multiple objects in images. In [67], the authors proposed a novel network which combines the attention mechanism and an auto-encoding framework for iterative image construction. The attention mechanism is applied to video description in [68], and [69], [70] studied the spatial attention for image captioning.

In our case, even though a powerful text reading system TextBoxes [4] is used to spot words in natural images, there are
still some errors (see Figure 1 due to large variations in both foreground texts and background objects, and uncontrollable lighting conditions). To reduce the influence of false detections and irrelevant words, we introduce attention mechanism to automatically analyze the correlation between texts and images. Hopefully, the attention mechanism helps to select most related words in images for the fine-grained classification.

### III. FINE-GRAINED CLASSIFICATION INTEGRATING TEXT AND VISUAL APPEARANCE

#### A. Proposed pipeline

In this paper, we present an end-to-end neural-network-based model for improving the fine-grained classification by combining visual and text features in the image. The proposed network structure is depicted in Figure 2. It consists mainly of three parts: visual and text feature extraction, semantic attention mechanism for noisy text filtering, and multi-modal feature fusion. The first part aims at transforming each input source into a feature representation. The second one is dedicated to reduce noisy texts brought in during text spotting. The last part combines different modality representations into a single representation. The whole framework is end-to-end trainable. In the following, we detail the main components of the proposed method.

#### B. Visual Representation

In computer vision tasks, convolutional neural networks have become a popular method to extract image features. In [71], it has been shown that CNN models trained on a large image dataset can learn common features, which can be shared in many different tasks. In [3], the authors have also shown that fine-tuning from a trained model can obtain better performance than training from scratch in some tasks. Based on this property, we use GoogleNet [5] trained on millions of ImageNet [2] images as initial CNN model and fine tune its parameters with other parts of our model in an end-to-end way. Given an image I, we use the output of the last average pooling layer of the fine tuned GoogLeNet as the representation of visual features \( f_v \), a 1024-dimensional feature vector.

#### C. Text Representation

Word2Vec [73] and GloVe [74] are two popular word embedding algorithms to construct vector representations for words. These vector-based models represent words in a continuous vector space where semantically similar words are embedded nearby each other. Pennington et al. [74] noted that GloVe consistently outperforms word2vec for the same corpus, vocabulary, window size and training time. So we adopt GloVe in this paper to transform recognized texts into text features. More specifically, we represent each spotted word \( T_i \) by the pre-trained 300-dimensional word vector \( f_t \), of GloVe. For an image with \( N \) spotted words, this results in a text feature \( f_t \) whose size is \( 300 \times N \). To train the network efficiently with batch, we fix the size \( N \) with the maximum number of spotted words in each training image denoted as \( N_{\text{max}} \). For the training and testing images having less than \( N_{\text{max}} \) spotted words, we use zero padding. For the testing images having more than \( N_{\text{max}} \) spotted words, we sort the words in decreasing order of their recognition scores and use the first \( N_{\text{max}} \) ones to form the text representation. Finally, the text information in each image is represented by a \( 300 \times N_{\text{max}} \) feature.

#### D. Attention Unit

Generally, not all the words in an image must have semantic relations to the scene. Some words may have strong correlations with the image, others may be not relevant at all. Consequently, different words should have different weights representing their correlations with the image. For example, it is reasonable that a word being more related to the image or having a high recognition score should have a large weight, and vice versa. Besides, the text reading system could make some mistakes and result in some noisy texts, which should be discarded. These noisy texts should have a very small weight. Inspired by the work in [70] which proposed a semantic attention mechanism to selectively attend on rich semantic attributes detected from the image, we apply a similar semantic attention mechanism to assign different weights over words in our case. The architecture of the proposed attention model is depicted in Figure 3.

In the attention unit, a bilinear function is used to estimate the weight \( \omega_i \) for each spotted word \( T_i \) in the \( N_{\text{max}} \) words reflecting the image’s text feature. This weight \( \omega_i \) measures the correlation between the spotted word \( T_i \) and the image. It is given by:

$$
\omega_i \propto \exp\left(f_t^T \cdot \bar{U} \cdot f_t\right),
$$

where the exponent is taken to normalize over all words \( T = \{T_i\} \) in a softmax fashion, and \( \bar{U} \) is the 1024 \( \times \) 300 bilinear parameter matrix. We then use weighted sum-pooling to fuse all the \( N_{\text{max}} \) word features. This fused word feature named attended text representation \( f_a \) is given by:

$$
f_a = \sum_{i=1}^{N_{\text{max}}} w_i \cdot f_t,
$$

where the exponent is taken to normalize over all words \( T = \{T_i\} \) in a softmax fashion, and \( \bar{U} \) is the 1024 \( \times \) 300 bilinear parameter matrix. We then use weighted sum-pooling to fuse all the \( N_{\text{max}} \) word features. This fused word feature named attended text representation \( f_a \) is given by:
E. Multi-Modal feature fusion

Multi-modal feature fusion has shown superior performance over unimodal approaches in many tasks [75]. A common multi-modal feature fusion method takes two or more modalities to form a preferable feature. In our problem, visual and attended text features provide important and complementary information about the image. The fine-grained classification would benefit from a combination of these two features.

Since the statistical properties are usually not shared across modalities, a transformation of features is required before their fusion. We use a 512-dimensional fully connected layer to map the visual feature into a suitable space $f_v$ having an approximate dimensionality as the attended text feature. Moreover, as shown in many applications, batch-normalization [76] could improve the performance of the proposed classifier integrating two features. Consequently, this would benefit from a combination of these two features.

F. Supervision Over Different Features

Multitask Learning (MTL) [77] is an inductive transfer mechanism which leverages the domain-specific information to improve the performance of generalization. It relies on training multi-tasks in parallel using a shared representation, which enables them help each other to improve their performances. Inspired by MTL, we add two extra classifiers on visual feature $f_v$ and attended text feature $f_a$, in parallel with the classifier based on multi-modal feature $f_c$. Simply put, we train in parallel three classifiers for fine-grained classification: 1) a classifier based on combined visual and attended text feature $f_c$; 2) a classifier using only visual feature $f_v$; 3) a classifier relying on attended text feature $f_a$.

Note that for the last classifier trained on $f_a$ only, we add an extra class representing images without text to train the network using only text information. Although we could train a visual baseline, we fine-tune the pre-trained GoogLeNet with a 28-category softmax classifier; a) Textual baseline. We compare the proposed method with two baselines: a) Visual baseline. We fine tune the pre-trained GoogLeNet with a 28-category softmax classifier; b) Textual baseline. As described in Section III-F, we have to add an extra class representing images without spotted text to train the network using only text information. Although we could

Let $\mathcal{O} = \{\Phi_i \mid i = 1, 2, 3\}$ denotes the parameters of the three classifiers, then the final loss function is simply given by:

$$\mathcal{L}(\mathbf{W}, \mathcal{O}) = \sum_{i=1}^{3} \beta_i \mathcal{L}_i(\mathbf{W}, \Phi_i),$$

where $\beta_i$ is the loss weight for the $i$-th classifier. We minimize this loss function $\mathcal{L}(\mathbf{W}, \mathcal{O})$ via standard (back-propagation) stochastic gradient descent (SGD):

$$(\mathbf{W}, \mathcal{O})^* = \arg\min \mathcal{L}(\mathbf{W}, \mathcal{O}))$$

IV. EXPERIMENTAL RESULTS

A. Experimental setup

1) Datasets: We evaluate the proposed method on two datasets: Con-Text dataset proposed in [20] and our self-collected Drink Bottle dataset. The Con-Text dataset is dedicated for fine-grained classification of business places, e.g., Cafe, Bookstore, and Pharmacy. It is composed of 24,255 images classified into 28 categories. All the images in this dataset are split into three folds, where two folds are used for training and the other one is used for testing. Each type of experiment on this dataset are repeated three times using respectively the three possible training and testing settings. The average performance is used as the reported results.

To further verify the generality of the proposed method, we have collected a new image dataset named Drink Bottle dataset which will be released soon. It consists of various types of drink bottle images contained in soft drink and alcoholic drink sets in ImageNet dataset [72]. The dataset has 18,488 images divided into 20 categories: Root Beer, Ginger ale, Coca Cola, Pepsi Cola, Cream Soda, Egg Cream, Birch Beer, Tonic, Sarsaparilla, Orange Soda, Pulque, Kvass, Bitter, Guinness, Ouzo, Slivovitz, Drambuie, Vodka, Chablis and Sauterne. The other settings follow Con-Text dataset. As shown in some image samples illustrated in Figure 4, the images in this dataset have low-resolutions and dense words. Moreover, a lot of fonts are in exotic styles. All these aspects make the fine-grained classification on this dataset more challenge than on the Con-Text dataset.

2) Evaluation measures: The Average Precision (AP) of each category and the mean of AP (mAP) over all categories are used to measure the performance. Following the work in [8], we first sort the classification score in descending order for each category, then the AP is given by:

$$AP = \frac{1}{n} \sum_{k=1}^{n} P(k) \ast (R(k) - R(k - 1)),$$

where $P(k)$ and $R(k)$ stands for precision and respectively recall at cut-off $k$ in the sorted list, $R(0) = 0$, and $n$ is the number of images in current category.

3) Baselines: We compare the proposed method with two baselines: a) Visual baseline. We fine tune the pre-trained GoogLeNet with a 28-category softmax classifier; b) Textual baseline. As described in Section III-F, we have to add an extra class representing images without spotted text to train the network using only text information. Although we could
evaluate the textual baseline using 29 categories, it is seriously imbalanced between the 28 original categories and the added category. Because most images in this dataset do not have text or the text reading system fails to spot words in images. Consequently, we train and test the textual baseline using 28 categories on a subset of images having spotted words. Since only text information is involved in the textual baseline, instead of using a weighted sum-pooling of word vectors described in Section II-D we take the classical average of the word vectors given by all spotted words as text feature for this baseline. For a fair comparison, we also train a visual baseline model and a model of the proposed method on this subset.

4) Implementation details: Our method is built on several pre-trained networks: GoogLeNet\cite{5} to extract image features, TextBoxes\cite{4} to detect texts in images, incremental CNN \cite{78} to recognize the detected words, and GloVe\cite{74} to generate the word embedding. During training, the mini-batch size for SGD is set to 64. The learning rate is set to 0.01 for the first 7K iterations, and then is divided by 10 every 10K iterations. The models are trained for up to 30K iterations. We use a weight decay of 0.0001 and a momentum of 0.9.

https://github.com/kmatzen/googlenet-caffe2torch
https://github.com/MhLiao/TextBoxes
http://www.robots.ox.ac.uk/~vgg/research/text/
https://github.com/stanfordnlp/GloVe

Fig. 4: Some image samples in the Drink Bottle dataset. They have very similar visual appearances. Distinguishing them is very challenge without textual cues.
are responsible for their fails. Some failure cases (fourth row) features being very close to the incorrectly predicted classes and eliminate irrelevant ones, even for the failure cases in the proposed attention unit can accurately select the relevant words spotted words in several images are given in Figure 5. The classical average-pooling scheme.

sent the textual information related to the scene. As depicted
weighted sum-pooling of word vectors is then used to repre-
assign appropriate weights to different spotted words. The
 These words may be similar in appearance but semantically far
away, such as ‘hotpot’ and ‘hotspot’, ‘barbera’ and ‘barber’,

For the text representation described in Section III-C we only consider spotted words having more than 3 letters. Because the words having less than 3 letters are more likely to be false positives or represent irrelevant information for fine-grained classification.

The loss weights in Eq. (4) are set respectively to $\beta_1 = 1.0$, $\beta_2 = 0.5$, and $\beta_3 = 0.5$. In the training stage, we use classical dataset augmentation strategy with different scales and aspect ratios. More precisely, we re-size each training image to $255 \times 255$. Then, we randomly crop it with re-size scale in the interval $[0.08, 1]$ and aspect ratio in $[3/4, 4/3]$. Finally, the cropped patch is re-sized to $224 \times 224$. Then randomly horizontal flip is also used with a probability of 0.5. In the testing stage, images are directly re-sized to $224 \times 224$ and the standard ten-crop testing is adopted.

The proposed method is implemented using Torch deep learning framework. All the experiments are conducted on a NVIDIA GTX TitanX GPU. The materials to reproduce all the reported results will be released soon.

### B. Importance of attention unit and multi-supervisions

1) **Importance of attention unit:** The state-of-the-art text reading system may result in some noisy texts including words that are irrelevant to the scene and incorrectly spotted words. These words may be similar in appearance but semantically far away, such as ‘hotpot’ and ‘hotspot’, ‘barbera’ and ‘barber’, and etc. We propose to use the attention mechanism to assign appropriate weights to different spotted words. The weighted sum-pooling of word vectors is then used to represent the textual information related to the scene. As depicted in Table I on the Con-Text dataset, the proposed attention-based method achieves nearly 1% performance gain over the classical average-pooling scheme.

Some weights given by the attention mechanism over spotted words in several images are given in Figure 5. The proposed attention unit can accurately select the relevant words and eliminate irrelevant ones, even for the failure cases in the third row. The attended weights are reasonable. Their visual features being very close to the incorrectly predicted classes are responsible for their fails. Some failure cases (fourth row) are due to the performance of the text reading system. The relevant words are not detected or incorrectly recognized, such as ‘LAUNDROMAT’ in (m), ‘TEA’ in (n), ‘ROOT BEER’ on the right can in (o), and ‘SODA’ in (p). Yet the words having large weights are semantically close to the scene class.

2) **Effect of multi-supervisions:** Although the main idea in this paper is to use combined text and visual feature $f_c$ to improve fine-grained classification, we rely on multisupervisions on $f_c$, $f_v$, and $f_a$ respectively during the training stage to boost the performance of the proposed classifier based on $f_c$. Without adjusting carefully the weights over different supervisions in Eq. (4), we simply set the loss weights for combined text and visual classifier, visual classifier, and text classifier as described in Section IV-A4. As depicted in Table II on the Con-Text dataset, the multi-supervisions provide about a 0.7% performance gain compared to single-supervision. Note that the two extra supervision classifiers are only used in training stage. During test stage, we simply use the classifier based on combined visual and text feature $f_c$. Consequently, the 0.7% performance gain thanks to multi-supervisions is achieved without any loss in runtime during the test.

### C. Comparison with baselines

We compare the proposed method with several baseline methods described in Section IV-A3 We have conducted two types of experiments using two different dataset settings: 1) Visual baseline and the proposed method trained on all training images and tested on all testing images; 2) Visual baseline, textual baseline, and the proposed method trained on the subset of training images and tested on the subset of testing images having spotted texts in each dataset. We add * after corresponding name to distinguish these experiments from previous ones.

Table III displays the quantitative comparisons on the Con-Text dataset. In general, the textual baseline achieves results on par with visual baseline on images having spotted texts in this dataset. This shows that the text feature is as useful as visual feature for fine-grained classification on images containing texts. The proposed method using the combination of text and visual features significantly outperforms the baseline based only on text or visual feature. A large improvement is observed particularly for the images having similar visual layout but discriminative textual cues, e.g., Discount House, Steak House, and Pawn Shop. For the images having fewer texts or no texts, the improvement is less significant. The observations on the Con-Text dataset also hold for the Drink Bottle dataset. As depicted in Table IV the proposed method also significantly outperforms the visual and textual baselines. Yet, the improvement over visual baseline (about 10%) is less important as compared to the improvement (about 20%) on the Con-Text dataset. The main reason is likely to be that recognition of the exotic texts in the low-resolution drink bottle images is more challenge. The text reading system would yield more noisy texts. The fact that less accurate text recognition gives less significant improvements also demonstrates the importance and interestingness of integrating text appearances into visual appearances for fine-grained classification.

To further analyze the benefit of integrating text information into visual appearance for fine-grained classification, we

| Text fusion method       | mAP(%) |
|-------------------------|--------|
| Average-pooling         | 78.6   |
| Weighted sum-pooling    | 79.6   |

**TABLE I:** Importance of semantic attention mechanism.

| Supervision method      | mAP(%) |
|-------------------------|--------|
| Single-Supervision      | 78.9   |
| Multi-Supervisions      | 79.6   |

**TABLE II:** Effect of multi-supervisions. In Single-Supervision, only a classifier trained on the combined feature $f_c$ is used. Multi-Supervisions denotes the use of the three classifiers described in Section III-F respectively on $f_c$, $f_v$, and $f_a$ features.

| Supervision method | mAP(%) |
|--------------------|--------|
| Supervision        | 78.9   |
| Multi-Supervisions | 79.6   |

**TABLE III:** Importance of semantic attention mechanism.
TABLE III: Classification performance for baselines, our method, and two state-of-the-art methods on the Con-Text dataset. Models with * are trained and tested on the subset of images containing spotted texts. For clearness, the words in the first row are massage parlor, pet shop, restaurant, computer store, theatre, repair shop, bookstore, bakery, medical center, barbershop, pizzeria, diner, hotspot, bistro, teahouse, school, pharmacy, funeral, country store, tavern, motel, packinghouse, cafe, tobacco shop, dry cleaner, discount house, steakhouse, and pawnshop respectively.

| Method          | Visual Baseline | Textual Baseline | Visual Baseline | ours* |
|-----------------|-----------------|------------------|-----------------|-------|
| Textual Baseline* | 81.4 84.3 72.4 74.3 | 86.1 72.9 85.5 80.4 | 74.2 85.9 79.9 75.7 69.0 20.9 | 55.7 66.6 72.7 70.6 60.3 33.8 |
| ours*           | 81.8 89.5 78.6 80.6 | 92.4 80.1 94.2 89.6 | 83.6 95.8 90.4 86.7 80.3 32.4 68.9 | 81.3 88.4 84.4 78.5 52.2 93.3 |

Karaoglu et al. [20] | 34.9 45.2 51.1 33.8 | 48.5 17.8 60.0 37.8 | 48.5 55.2 55.2 43.4 65.7 | 9.1 12.5 44.3 60.8 35.2 10.6 |
Karaoglu et al. [3] | - - - - - - - - - - - - - - - - - - - - - - - - - - - |

TABLE IV: Classification performance for baselines and the proposed method on the Drink Bottle dataset. The category of every logogram in this table is depicted in Section IV-A1. Similar to Table III, models with * are trained and tested on the subset of images containing spotted texts.

| Method          | r.b. g.a. coke pepsi c.s. e.c. b.c. tonic sarsap. o.s. pulque kvass bitter guin. ouzo sliv. dramb. vodka chab. saut. | mAP       |
|-----------------|-------------------------------------------------------------------------------------------------|-----------|
| Visual Baseline | 64.1 63.2 94.8 98.1 27.5 40.4 33.1 59.6 29.5 77.8 | 58.3 26.4 76.4 97.0 64.8 47.3 88.4 82.2 90.9 87.1 65.4 |
| Textual Baseline | 54.7 61.9 34.1 76.5 27.0 0.2 22.3 33.6 18.6 54.5 37.4 23.2 40.2 98.0 74.0 96.1 98.0 | 54.4 41.1 47.4 69.1 48.9 85.4 65.6 40.2 84.7 98.5 97.8 59.5 91.7 90.8 93.5 90.9 74.5 45.0 |
| ours*           | 61.9 54.2 94.1 95.6 32.4 60.7 35.7 53.4 41.1 78.5 51.7 31.6 66.5 93.3 45.3 43.6 75.3 75.9 85.1 84.4 63.1 |
| ours            | 76.8 70.3 95.8 96.9 52.5 73.2 42.9 65.6 57.8 87.4 61.7 39.4 77.3 96.7 63.8 54.1 79.9 85.7 90.2 88.7 72.8 |

TABLE V: Average precision (in %) on the subset of images having spotted words and the rest images in the Con-Text dataset. Both models are trained on all training images.

| Method          | image with texts | image without texts |
|-----------------|------------------|---------------------|
| Visual Baseline | 52.8             | 65.8                |
| ours            | 79.0             | 71.7                |

D. Comparison with other methods

We compare the performance of the proposed method and visual baseline on the subset of testing images having spotted texts and subset without texts separately in Con-Text dataset. In this dataset, about 42% images contain texts. The quantitative comparison is given in Table V. The proposed method improves dramatically the performance on images having texts as observed in the above experiments. Surprisingly, it also boosts slightly the performance for images without texts. One reason for this improvement may lie on the multi-supervisions of the network using shared parameters.

E. Application to Product Retrieval

We present an application of the proposed framework for **Product Retrieval** to further demonstrate its potential. Compared to image classification, image retrieval attracts more attention in e-commerce. Given a query image $I$, the objective

\[ \text{query} = \text{image} \]
Fig. 5: Some classification results on the Con-Text dataset and Drink Bottle dataset (better visualization in electronic version). The ground truth and predicted class (in blue and red for correct and respectively incorrect predictions) are given in the first line at the bottom of each image. The spotted words and their corresponding weights are shown in the rest lines.
of product retrieval is to select all product images from an image set within the category as $I$. We have conducted this experiment on the Drink Bottle dataset. We simply extract the input vector of the softmax-based classifier in fine-grained classification model. This input vector is considered as product feature for retrieval. The cosine similarity is used to rank images during matching. Following FlickrLogos-32’s evaluation settings [80], the test set and train set are used as queries and respectively search set. We evaluate this application based on mAP. As shown in Table VI, the proposed method boosts consistently and significantly the retrieval performance.

V. CONCLUSION

In this paper, we have introduced a unified, end-to-end trainable framework which integrates visual and text cues for fine-grained image classification. We have also incorporated the attention mechanism to better exploit the extracted text in images by only selecting the most relevant words. Using several state of the art components in an effective pipeline, our experimental results on two datasets have demonstrated both the value and performance of the proposed method. Furthermore, the product retrieval experiment shows that the learned representation provides a potentially powerful tool for product search. It is worth noting that while it certainly makes sense to employ the state of the art in scene text and visual recognition as the components, our main motivation and contribution are to fill a gap in fine-grained image classification by exploiting multi-modal complementary information when present, and to do so with an end-to-end trainable neural network with the attention mechanism for implicitly exploiting the semantic relationships between textual and visual cues. The main drawback of using textual cues in the proposed framework for fine-grained image classification is clearly the absence of texts in some images. Besides the scene texts in images, the additional textual cues could be user tags, images attributes, and so on. In the future, we would like to incorporate other textual cues to handle the problem of no texts in images.

REFERENCES

[1] Q. Ye and D. Doermann, “Text detection and recognition in imagery: A survey,” *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 37, no. 3, pp. 1480–1500, 2015.

[2] Y. Zhu, C. Yao, and X. Bai, “Scene text detection and recognition: Recent advances and future trends,” *Frontier of Computer Science*, vol. 10, no. 1, pp. 19–36, 2016.

[3] S. Karaoglu, R. Tao, T. Gevers, and A. W. Smeulders, “Words matter: Scene text for image classification and retrieval,” *IEEE Transactions on Multimedia*, 2016. DOI: 10.1109/TMM.2016.2638622

[4] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, “Textboxes: A fast text detector with a single deep neural network,” in AAAI, 2017.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[6] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. DOI: 10.1109/TPAMI.2016.2646371

[7] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based r-cnn for fine-grained category detection,” in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[8] A. Angelova and S. Zhu, “Efficient object detection and segmentation for fine-grained recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 811–818.

[9] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3d object representations for fine-grained categorization,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[10] S. Yang, L. Bo, J. Wang, and L. G. Shapiro, “Unsupervised template learning for fine-grained object recognition,” in Advances in neural information processing systems, 2012, pp. 3122–3130.

[11] J. Deng, J. Krause, and L. Fei-Fei, “Fine-grained crowdsourcing for fine-grained recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 580–587.

[12] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 842–850.

[13] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do, “Weakly supervised fine-grained categorization with part-based image representation,” *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1713–1725, 2016.

[14] P. Tang, X. Wang, B. Feng, and W. Liu, “Learning multi-instance deep discriminative patterns for image classification,” *IEEE Transactions on Image Processing*, 2016. DOI: 10.1109/TIP.2016.2642781

[15] N. Zhang, R. Farrell, and T. Darrell, “Pose pooling kernels for sub-category recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3665–3672.

[16] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[17] Q. Qian, R. Jin, S. Zhu, and Y. Lin, “Fine-grained visual categorization via multi-stage metric learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3716–3724.

[18] S. Xie, T. Yang, X. Wang, and Y. Lin, “Hyper-class augmented and regularized deep learning for fine-grained image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2645–2654.

[19] M. Wilber, I. S. Kwak, D. Kriegman, and S. Belongie, “Learning concept embeddings with combined human-machine expert in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 981–989.

[20] S. Karaoglu, J. C. van Gemert, and T. Gevers, “Con-text: text detection using background connectivity for fine-grained object classification,” in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 757–760.

[21] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *European Conference on Computer Vision*. Springer, 2014, pp. 391–405.

[22] X. Chen and A. L. Yuille, “Detecting and reading text in natural scenes,” in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004. CVPR 2004., vol. 2, pp. 366–373.

[23] B. Epshtein, O. Ofek, and Y. Wexler, “Detecting text in natural scenes with stroke width transform,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2963–2970.

[24] K. Wang and S. Belongie, “Word spotting in the wild,” in *European Conference on Computer Vision*. Springer, 2010, pp. 501–604.

[25] Y.-F. Pan, X. Hou, and C.-L. Liu, “A hybrid approach to detect and locate texts in natural scene images,” *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 800–813, 2011.
[73] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[74] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[75] L. Deng, D. Yu et al., “Deep learning: methods and applications,” Foundations and Trends® in Signal Processing, vol. 7, no. 3–4, pp. 197–387, 2014.

[76] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.

[77] R. Caruana, “Multitask learning,” in Learning to learn. Springer, 1998, pp. 95–133.

[78] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Synthetic data and artificial neural networks for natural scene text recognition,” arXiv preprint arXiv:1406.2227, 2014.

[79] R. Collobert, K. Kavukcuoglu, and C. Farabet, “Torch7: A matlab-like environment for machine learning,” in BigLearn, NIPS Workshop, no. EPFL-CONF-192376, 2011.

[80] S. Romberg, L. G. Pueyo, R. Lienhart, and R. Van Zwol, “Scalable logo recognition in real-world images,” in Proceedings of the 1st ACM International Conference on Multimedia Retrieval. ACM, 2011, p. 25.