Knowledge Mining with Scene Text for Fine-Grained Recognition

Hao Wang[1]*, Junchao Liao[1]*, Tianheng Cheng[1], Zewen Gao[1], Hao Liu[2], Bo Ren[2], Xiang Bai[1], Wenyu Liu[1]†
[1]Huazhong University of Science and Technology, [2]Tencent YouTu Lab
{wanghao4659,liaojc,thch,gaozw,xbai,liuwy}@hust.edu.cn, {ivanhliu,timren}@tencent.com

Abstract

Recently, the semantics of scene text has been proven to be essential in fine-grained image classification. However, the existing methods mainly exploit the literal meaning of scene text for fine-grained recognition, which might be irrelevant when it is not significantly related to objects/scenes. We propose an end-to-end trainable network that mines implicit contextual knowledge behind scene text image and enhance the semantics and correlation to fine-tune the image representation. Unlike the existing methods, our model integrates three modalities: visual feature extraction, text semantics extraction, and correlating background knowledge to fine-grained image classification. Specifically, we employ KnowBert to retrieve relevant knowledge for semantic representation and combine it with image features for fine-grained classification. Experiments on two benchmark datasets, Con-Text, and Drink Bottle, show that our method outperforms the state-of-the-art by 3.72% mAP and 5.39% mAP, respectively. To further validate the effectiveness of the proposed method, we create a new dataset on crowd activity recognition for the evaluation. The source code and new dataset of this work are available at this repository[1].

1. Introduction

The text conveys the information, knowledge, and emotion of human beings as a significant carrier. Texts in natural scene images contain sophisticated semantic information that can be used in many vision tasks such as image classification, visual search, and image-based question answering.

Several approaches [2,15,22,23,26,39] were proposed to incorporate semantic cues of scene text for image classification or retrieval and achieved significant performance improvements. These methods follow a general pipeline that first spots the text by a scene text reading system, then converts the spotted word into text features to combine it with image features for the subsequent tasks.

(a) (b) (c)

Text	Entity	Description of entity
"party"	political party	organized group of people who have the same ideology
"party"	party	social event
"Leninade"	Leninade	Soviet-themed lemonade soda

(d): Knowledge in Wikipedia about text from (c)

Figure 1. The three images belong to the category of "Soda". (d) shows the knowledge behind scene text embodied in the image (c) from knowledge base Wikipedia. Each text instance contains one or more entities stored in the knowledge base. The associated descriptions further explain the precise meaning of entity. Only the entities of two text instances are listed for simplicity.

This paper explores how to dig deeper into background knowledge and extract context information of scene text for the fine-grained image classification task. Unlike document text, in our observation, the natural scene text is often sparse, appearing as a few keywords rather than complete sentences. Moreover, these few keywords may be vague and give no clue to the classification model when their semantic cues are not directly related to the precise meaning that the image conveys.

As shown in Fig. 1 (a) and (b), the literal meaning of the keyword "Soda" explicitly expresses that the bottles in the two images belong to the category *Soda* despite their intra-class visual variance. However, we hardly understand the object in Fig. 1 (c) by solely fetching the semantic cues of scene text. To understand the image certainly, getting more relevant contextual knowledge about the image is crucial. Therefore, we explore how to dig extra background knowledge and mine the contextual information to enhance the correlation between scene text and a picture. For example, the table in Fig. 1 (d) exhibits related information or knowl-

*Authors contribute equally.
†Corresponding author.
[1]https://github.com/lanfeng4659/KnowledgeMiningWithSceneText

edge of scene text embodied in (c). The description of the entity *Leninade* informs that it is a Soda beverage bottle. Thus, the knowledge extracted in this manner complements the literal meaning of the raw text and reduces the semantics loss caused by using the literal meaning of scene text only.

Specifically, after extracting the text from the image by a scene text reading system [20, 40], we retrieve relevant knowledge from databases such as (e.g., WordNet [25] and Wikipedia) that store rich human-curated knowledge with all possible correlation to the target. As shown in Fig. 1 (d), the possible entities (e.g., party and political party) can be extracted for the text instance "party" from the knowledge databases. However, all the retrieved contextual knowledge may not necessarily provide helpful semantic cues to understand the visual contents. In order to filter relevant contextual information from irrelevant, we design an attention module that focuses on very pertinent knowledge for the semantics of objects or scenes.

We evaluate the performance of our method on two public benchmark datasets, Bottles [2] and Con-Text [16]. The results demonstrate the usage of contextual knowledge behind scene text can significantly promote fine-grained image classification models performances. To further prove the effectiveness of our method, we developed a new dataset consisting of 21 categories and 8785 natural images. Furthermore, the dataset mainly focuses on crowd activity, while most images contain multiple scene text instances. To the best of our knowledge, the existing crowd activity datasets do not contain scene text instances. However, everyday human activities are highly related to scene text presences, for example, procession, exhibitions, press briefing, and sales campaigns. This dataset will be a valuable asset for exploring the role of scene text on crowd activity.

In this paper, we propose a method that mines contextual knowledge behind scene text to improve the performance of the multi-modality understanding task. To this end, we design a deep-learning-based architecture that combines three modality features, including visual contents, scene text, and knowledge for fine-grained image recognition. Our method achieves significant improvements and can be applied to other tasks, such as visual grounding [33] and text-visual question answering [3] beyond the fine-grained image classification task. In addition, we propose a new dataset where each image contains multiple scene text instances, which promotes the study of multi-modal crowd activity analysis.

2. Related Work

2.1. Fine-Grained Image Classification

The task of fine-grained image classification needs to distinguish images with subtle visual differences among object classes in some domains, such as animal species [12, 18], plant species [24] and man-made objects [19]. Previous methods [6, 10] classify objects with only visual cues and aim at finding a discriminative image path. Recently, some approaches have shown a growing interest in employing textual cues to combine the visual cues for this task. Movshovitz *et al*. [26] first propose to leverage scene text for the fine-grained image classification task by using the visual cues of scene text. However, extracting robust visual cues of scene text is challenging due to blur and occlusion of text instances. Karaoglu *et al*. [15] employ the textual cues of scene text as a discriminative signal and combine the visual features that are obtained by the GoogLeNet [38] to distinguish business place. To fully exploit the complementarity of visual information and textual cues, several methods [2, 22] propose to fuse features of the two modalities with an attentional module. Bai *et al*. [2] propose an attention mechanism to select textual features from word embeddings of recognized words. To overcome optical character recognition errors, Mafla *et al*. [22] leverage the usage of the PHOC [1] representation to construct a bag of textual words along with the fisher vector [29] that models the morphology of text. Despite the promising progress, the existing methods exploit the literal meaning of scene text and overlook the meaningful human-curated knowledge of text.

2.2. Knowledge-aware Language Models

The pre-trained language models such as ELMo [30] and BERT [8] are optimized to either predict the next word or some masked words in a given sequence. Petroni *et al*. [32] find that the pre-trained language models, such as BERT, can recall factual and commonsense knowledge. Such knowledge is stored implicitly in the parameters of the language model and useful for downstream tasks such as visual question answering [17]. This knowledge is usually obtained either from the latent context representations produced by the pre-trained model or by using the parameters of the pre-trained model to initialize a task-specific model for further fine-tuning. To further enhance the language model awareness of human-curated knowledge better, some works [31, 34] explicitly integrate the knowledge in knowledge bases into the pre-trained language model. In our method, we employ both BERT [8] and KnowBert [31] as a knowledge-aware language model and apply them to extract knowledge features. Although previous methods [36] extract knowledge features from sentences on vision-language tasks, they require the annotation of image-text pairs.

3. Methodology

As shown in Fig. 2, the proposed network accepts as input an image, a knowledge base, and scene text spotted by a scene text reading system such as [20, 40]. The part of extracting features in our framework consists of three branches, the visual features extraction branch, the knowledge extraction branch for retrieving relevant knowledge,

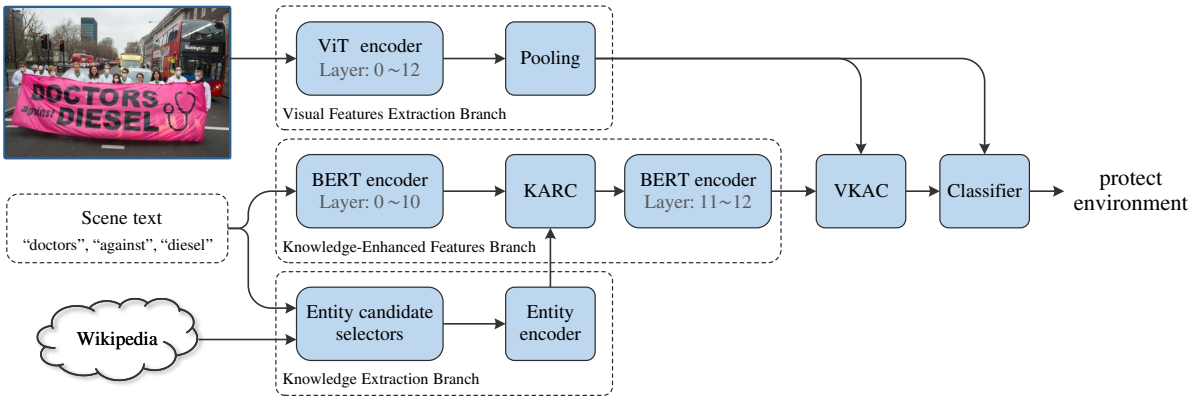


Figure 2. The framework of our method. The proposed model combines visual cues and textual cues for classification. The input text instances are spotted by a scene text reading system. KARC and VKAC mean the knowledge attention and recontextualization component and the visual-knowledge attention component, respectively.

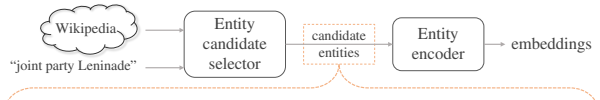
and the knowledge-enhanced features branch that employs the retrieved knowledge to enhance the presentations of scene text. Then, the visual-knowledge attention component (VKAC) inputs the visual features and the knowledge-enhanced text features and outputs the attended features. Moreover, the concatenation of visual features and attended features is fed to the subsequent classifier.

In our method, we employ ViT [9] to extract the global visual features of the input image. We mainly detail the knowledge extraction branch, Knowledge-enhanced features branch, and the visual-knowledge attention component in the following subsections.

3.1. Knowledge extraction branch

The goal of this branch is to extract relevant knowledge from Wikipedia and embed them into features. Such knowledge is stored via entities in a knowledge base, and relevant entities can be queried by scene text instances in our method. However, most text instances can map to multiple entities due to the uncertainty of the meaning of the text. For example, the text "apple" can denote the entity of either fruit apple or Apple company. This requires an entity candidate selector that takes as input a sentence and returns a list of C potential entities.

Inspired by [11], we use an entity prior for entity candidate selection. The prior means the probability of a text instance being an entity, which is computed by averaging hyperlink count statistics from Wikipedia, a large Web corpus [37], and the YAGO dictionary [14]. As depicted in Fig. 3, first, we combine all scene text instances to sentence according to the spotting order. Then, the tokens of this sentence are obtained as BERT does. The entity candidate selector generates the top C entity candidates of each text instance based on the prior. Finally, the entity embeddings are obtained via the precomputed entity encoder in KnowBert. Specifically, the entity encoder adopts a skip-gram like objective to learn 300-dimensional embeddings of Wikipedia page titles from Wikipedia descriptions. As a result, such entity embeddings encode the factual knowledge mined from Wikipedia descriptions.



Text instance	Entity	Span	Priors	Description of entity
joint	-	-	-	-
party	political party	[1,1]	0.20	organized group of people who have the same ideology
party	party	[1,1]	0.15	social event
Leninade	Leninade	[2,2]	1.00	Soviet-themed lemonade soda

Figure 3. The process of knowledge extraction branch. The span is the [start index, end index] of the token inside the sentence.

3.2. Knowledge-enhanced features branch

This branch aims at using the retrieved entity embeddings to enhance the representations of text. The architecture is adapted from KnowBert that incorporates knowledge bases into BERT by inserting a knowledge attention and recontextualization component (KARC) at a particular layer. Following KnowBert, we insert Wikipedia into the 10^{th} layer of the encoder of BERT.

The brief pipeline of this branch is given in Fig. 2. Formally, a sequence of word piece tokens is fed to the former 10 successive encoder layers of BERT, outputting the contextual representations H_i. Then, the KARC takes as inputs H_i and candidate entity embeddings and outputs knowledge enhanced representations $H_i^{'}$. Finally, these enhanced representations are fed to the remainder of the encoder of

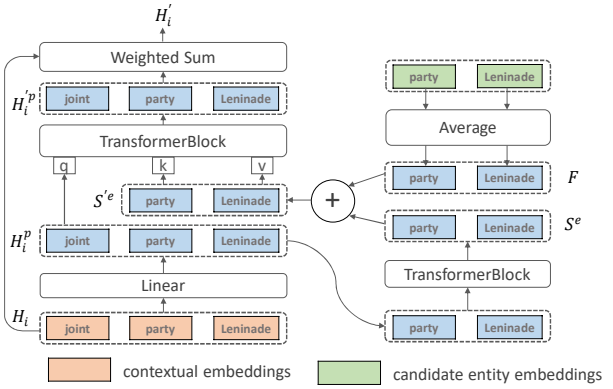


Figure 4. The architecture of the knowledge attention and recontextualization component.

BERT, generating the final knowledge enhanced features. The module in each encoder layer of BERT is the TransformerBlock formulated as

$$H_i = \text{TransformerBlock}(H_{i-1}, H_{i-1}, H_{i-1}). \quad (1)$$

This block uses H_{i-1} as the query, key, and value to allow each vector to attend to each other.

The KARC is the key component for integrating the retrieved entity embeddings to H_i. Different from the one in KnowBert, the width of the span is restricted as 1 in our KARC. Namely, these entities named as more than one text instance are ignored due to the sparsity of scene text. The details of KARC are given in Fig. 4, the word piece representations (H_i) are first projected to H_i^p by a linear layer. The representations of those word pieces that link to at least one entity are contextualized into contextual word representations S^e by a TransformerBlock. Meanwhile, the C candidate entity representations of each token are averaged to form weighted entity embeddings F. Specifically, as KnowBert does, we disregard all candidate entities with scores below a fixed threshold, and softmax normalize the remaining scores to weight the corresponding candidate entity representations. Then, S^e are updated by adding entity embeddings F to form word-entity representations $S^{'e}$. The $S^{'e}$ is employed to recontextualize the H_i^p with a TransformerBlock, where we substitute H_i^p for the query, and $S^{'e}$ for both the key and value:

$$H_i^{'p} = \text{TransformerBlock}(H_i^p, S^{'e}, S^{'e}), \quad (2)$$

Finally, a residual connection is adapted to fuse the $H_i^{'p}$ and H_i, forming the knowledge enhanced representations $H_i^{'}$:

$$H_i^{'} = g(H_i^{'p}) + H_i, \quad (3)$$

where, g is a linear function. The fully connected layer is employed in our method.

3.3. Visual-knowledge attention component

Generally, not all knowledge of text in an image must have semantic relations to the object or scene. Some retrieved knowledge may have strong correlations with the image, others may be not relevant at all. Therefore, we design an attention component that focuses on very pertinent knowledge for the semantics of objects or scene. The basic idea is that we take the global visual feature $f_v \in \mathbb{R}^{1\times D}$ as query and retrieve those knowledge features that are highly similar to f_v from all knowledge features $H \in \mathbb{R}^{N\times D}$. The parameter D is the feature dimension.

Formally, given f_v and H, we first calculate their similarities, which is defined by:

$$W = softmax(\frac{\theta(f_v)\cdot(\phi(H))^T}{\sqrt{D}}), \quad (4)$$

where both θ and ϕ are a single linear function that projects the features into a feature space, $W \in \mathbb{R}^{1\times N}$ is the outputted similarity matrix. Then, W is used for weighting knowledge features. Finally, the weighted features are fed to a residual connection block. The implemented process is defined as follow:

$$H_{att} = W \cdot \psi(H), \quad (5)$$

$$H_{out} = \kappa(H_{att}) + H_{att}, \quad (6)$$

where $H_{out} \in \mathbb{R}^{1\times D}$ is the attended knowledge features, κ is a linear function.

3.4. Classifier and loss function

The classifier consisting of a fully connected layer and a softmax layer performs the classification task, inputting the concatenation of the global visual features and the knowledge-enhanced features. The objective function is formulated as

$$L = -\frac{1}{M}\sum_{m=1}^{M}\mathbf{1}(m=y)\log p_m, \quad (7)$$

where M is the number of categories, p_m is the probability of predicting the sample as the m^{th} category, y is the associated label.

4. Experiments

First, we introduce the datasets used in our experiments and the new dataset created by us. Then, the implementation details are given. Third, we evaluate our method on our proposed Crowd Activity dataset and make comparisons with the state-of-the-art approaches. Last, we conduct the ablation studies. We compare with previous methods under the metric of mAP as most existing methods do.

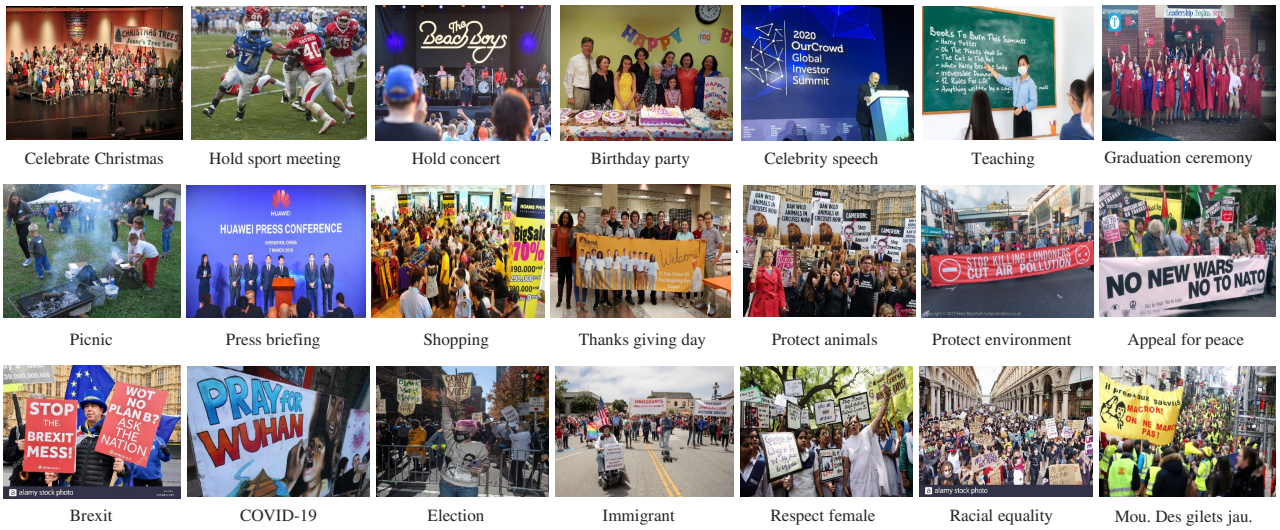


Figure 5. Examples of 21 categories from the Crowd Activity dataset.

Method	Activities of daily living											Demonstrations										mAP
	c.c.	h.s.	h.c.	b.p.	c.s.	teac.	g.c.	pic.	p.b.	shop.	t.g.	p.a.	p.e.	a.p.	brex.	cov.	elec.	imm.	r.f.	r.e.	m.d.	
R152 [13]	59.6	92.5	70.4	71.1	48.4	88.7	89.6	80.5	83.2	86.4	68.8	68.5	62.4	74.3	84.7	50.6	59.6	56.5	68.8	48.2	91.0	71.6
ViT [9]	76.0	**98.7**	81.1	83.6	57.1	85.3	93.5	84.7	93.0	88.6	73.6	73.6	71.5	78.8	85.7	75.6	74.6	75.8	84.3	60.0	91.0	80.3
fastText [4]	58.3	46.9	55.3	56.0	33.4	46.1	59.6	31.7	52.0	47.9	27.1	87.2	82.7	76.7	78.9	57.6	69.6	69.8	73.0	55.0	75.4	59.1
KB [31]	62.8	56.3	55.3	59.5	51.4	54.1	70.3	46.1	54.8	45.9	45.8	89.9	79.2	78.5	78.1	72.8	73.8	67.0	77.2	64.2	74.9	64.7
Mafla *et al*. [22]	60.0	90.9	75.6	76.4	49.4	89.0	86.4	83.5	79.0	94.2	67.1	83.1	76.2	82.4	88.7	65.5	72.4	71.4	74.7	67.7	95.5	77.6
Mafla *et al*. [23]	72.3	87.5	78.1	80.7	50.3	91.6	86.5	81.1	73.0	89.1	62.6	87.4	79.2	86.5	85.6	75.5	79.1	73.1	80.3	67.9	97.0	79.2
Ours	**83.0**	98.5	**88.8**	**86.1**	**60.5**	**89.4**	**95.7**	**89.1**	**94.0**	**94.5**	**78.2**	**92.4**	**92.4**	**89.6**	**95.4**	**83.0**	**82.1**	**84.7**	**90.1**	**73.7**	**98.1**	**87.5**
Gain	**3.9**											**11.1**										**7.2**

Table 1. Classification performance for baselines and the proposed method on the Crowd Activity dataset. KB denotes KnowBert.

4.1. Datasets

Con-Text dataset is introduced by Karaoglu [16] and is a subset of ImageNet dataset [7]. This dataset is constructed by selecting the sub-categories of "building" and "place of business", consisting of 24,255 images classified into 28 categories that are visually similar.

Drink Bottle dataset is presented by Bai [2] and consists of various types of drink bottle images contained in soft drink and alcoholic drink sets in ImageNet dataset [7]. The dataset has 18,488 images divided into 20 categories.

All categories within the existing two datasets are about products or places of business. The textual cues of those categories are obvious, and most images can be understood by the apparent meaning of scene texts rather than the knowledge behind them. Therefore, we create a new dataset that concentrates on the activities of the crowd for a fine-grained image classification task, named as **Crowd Activity** dataset, as automatically understanding crowd activity is meaningful for social security. This dataset is newly collected, where the images are mainly searched on the Internet and collected from streets by mobile phones. All images in this dataset contain at least one text instances. The categories come from activities of daily living and demonstrations stimulated by hot events in recent years. Specifically, this dataset consists of 21 categories and 8785 images in total. As shown in Fig. 5, the 21 categories broadly fall into two types: activities of daily living(*i.e*., *celebrating Christmas*, *holding sport meeting, holding concert*, *celebrating birthday party*, *celebrity speech*, *teaching*, *graduation ceremony*, *picnic*, *press briefing*, *shopping*, *celebrating Thanks giving day*) and demonstrations (*i.e*., *protecting animals*, *protecting environment*, *appealing for peace*, *Brexit*, *COVID-19*, *election*, *immigrant*, *respecting female*, *racial equality*, *mouvement des gilets jaunes*).

4.2. Implementation Details

Before training, we first extract scene text by Google OCR or E2E-MLT. Then, the model of our method is trained

Method	Vision	Text Spotter	Embedding	Con-Text	Bottles	Activity
Karao. *et al*. [16]	BOW	Custom	BoB	39.00	-	-
Karao. *et al*. [15]	BOW+GoogLeNet	Jaderberg	Probs	77.30	-	-
Bai *et al*. [2]	GoogLeNet	Textboxes	GloVe	78.90	-	-
Bai† *et al*. [2]	GoogLeNet	Google OCR	GloVe	80.50	74.50	-
Mafla *et al*. [22]	ResNet-152	E2E-MLT	GloVe	77.58	74.91	72.58
Mafla *et al*. [22]	ResNet-152	E2E-MLT	fastText	77.77	75.40	73.01
Mafla *et al*. [22]	ResNet-152	SSTR-PHOC	PHOC	77.45	75.93	73.84
Mafla *et al*. [22]	ResNet-152	SSTR-PHOC	FV	80.21	77.38	77.57
Mafla *et al*. [23]	ResNet-152	E2E-MLT	fastText	82.36	78.14	75.31
Mafla *et al*. [23]	ResNet-152	SSTR-PHOC	PHOC	82.77	78.27	75.45
Mafla *et al*. [23]	ResNet-152	SSTR-PHOC	FV	83.15	77.86	77.54
Mafla *et al*. [23]	ResNet-152	Google OCR	fastText	85.81	79.87	79.25
Ours	ResNet-152	E2E-MLT	KnowBert	84.93	79.32	81.91
Ours	ViT	E2E-MLT	KnowBert	87.28	84.01	85.68
Ours	ViT	Google OCR	KnowBert	**89.53**	**85.26**	**87.45**

Table 2. Classification performance of state-of-the-art methods on the Con-Text, Drink-Bottle, and Activity datasets. BOW denotes bag of visual words. BoB denotes Bag of Bigrams. FV denotes Fisher Vector.

in an end-to-end manner. For the data augmentation on images, we first randomly crop an image patch on the original image with the scale from 0.05 to 1.0 while keeping the ratio in a range of [0.75, 1.33]. Next, the image patch is resized to 224×224. Finally, we perform normalization on the image by setting both the mean and the standard deviation as $(0.5, 0.5, 0.5)$. As for training BERT and KnowBert, no data augmentation is used other than shuffling the order of scene text before grouping them into a sentence, as both BERT and KnowBert can overfit quickly when the input text is not so abundant. We adapt AdamW [21] to optimize the whole network with an initial learning rate of 3e-5. The learning rate warmup for 500 iterations and the cosine annealed warm restart strategy are adopted at the same time. All models are trained on the dataset for 10 epochs.

We conduct all experiments based on PyTorch [27]. The codes of ResNet-152 [13] and ViT [9] are from [22] and the timm package [41]. For both ResNet-152 and ViT, the pre-trained models on ImageNet are used for finetuning. The implementation of BERT [8] and KnowBert are from the huggingface transformers [42] and [31]. The BookCorpus [43] and English Wikipedia pre-trained model are loaded on BERT. In addition, we use `torchtext`, which is a package from PyTorch for the GloVe [28] and fastText [4].

During testing, the shorter side of the image is resized to 224. Then a 224×224 image patch is cropped from the image center. As for the spotted scene text, we keep their original order for BERT and KnowBert.

4.3. Baselines on the crowd activity dataset

We compare our method with several baseline methods, including visual baseline (ResNet-152 and ViT), textual/-knowledge baseline (fastText and KnowBert), and multi-modal baseline ([22]and [23]) on our proposed crowd activity dataset. We conduct two types of experiments using two different dataset settings. 1) The visual baseline and multi-modal baseline models are trained on all training images and tested on all testing images. 2) The textual/knowledge baseline models are trained and tested on the subset of images consisting of spotted texts. The textual cues used in [22]and [23] are from fastText.

Tab. 1 displays the quantitative comparisons on the crowd activity dataset. Among previous methods, ViT achieves state-of-the-art performances, while our method outperforms ViT by 7.2% mAP. In particular, the improvements of the subset of demonstrations reach more than 11.0% mAP, which is the highest gain than activities of daily living. The reason is that the visual cues on those demonstration activities are incredibly subtle. For example, most scenarios are that protest marchers hold flags and slogans and walk on the street. Such subtle visual cues require valuable knowledge for better understanding those scenes. Thus, the performance improvement confirms the significance of scene text instances in datasets such as Crowd Activity for the robust classification of fine-grained images.

4.4. Comparisons with state-of-the-art method

Bai *et al*. [2] take GoogLeNet [38] as visual backbone while the most recent state-of-the-art methods [22, 23] employ ResNet-152 [13]. For a fair comparison, we first evaluate our method with ResNet-152 and take E2E-MLT [5] as text spotter. Then, we conduct experiments under the setting of ViT and Google OCR.

As shown in Tab. 2, our model achieves the best performance on the three datasets. The method [23] outperforms previous methods by using the features of general

Figure 6. Some examples of classification results. GT denotes Ground Truth. The Top-1 prediction and its probability are shown below each picture. The names of some categories are abbreviated.

Figure 7. Visualization results. The top two are ResNet-152 grad-CAM [35] results, and the bottom two are ViT attention maps.

	Vision	Emb.	C.T.	Bottles	Activity
vis.	R152	-	70.96	73.41	71.58
	ViT	-	79.24	80.81	80.29
vis. + text	R152	GloVe	73.97	76.67	74.75
	R152	fastText	73.66	76.67	74.89
	ViT	GloVe	79.79	80.56	81.25
	ViT	fastText	79.82	81.18	80.71
vis. + text + know.	R152	BERT	81.59	77.94	81.68
	R152	KB	85.42	80.17	83.79
	ViT	BERT	86.51	82.81	85.34
	ViT	KB	**89.53**	**85.26**	**87.45**

Table 3. Performances of different vision and embedding models combinations on three datasets. R152 denotes ResNet-152. The abbreviated names, vis. and know., mean visual and knowledge. The text + know. means the features containing knowledge of texts. C.T. means Con-Text. (Metric: mAP)

objects within images. However, our model surpasses it, by 5.39% and 3.72% on Drink Bottle and Con-Text datasets, respectively. The method [22] does not use the information of general objects. Consequently, our method achieves superior performance on the two public datasets over the method [22]. The consistent outperformance of our proposed model over existing methods demonstrates the significance and effectiveness of integrating the knowledge behind scene text for better understanding the objects or scene. To further validate the significance of introducing knowledge to this task, we compare our method with [22] and [23] on our Crowd Activity dataset. Specifically, we train the model with their officially released codes[2,3]. As depicted in Tab. 2, our method outperforms the method [23] by 8.20% mAP, which further illustrates that mining knowledge is vital to understand the meanings of natural images fully.

As some qualitative results of our method are shown in Fig. 6, the proposed method can identify these visually alike images on Drink Bottle and Con-Text datasets. As illustrated in Sec. 4.3, the visual cues and the literal meaning of scene text in images are highly subtle on the crowd activity dataset. Yet, our method still classifies them very well.

4.5. Ablation study

This section provides detailed ablation studies to validate the effect of different modules included in the proposed model for mining knowledge. Thus, we present the performances on the three datasets under various combinations

[2]http://github.com/DreadPiratePsyopus/Fine_Grained_Clf
[3]https://github.com/AndresPMD/GCN_classification

	KARC	VKAC	C.T.	Bottles	Activity
Baseline			86.51	82.81	85.34
model A	√		87.25	83.59	86.16
model B	√	√	**89.53**	**85.26**	**87.45**

Table 4. Ablation studies of KARC and VKAC components. ViT is applied to extract features from images. C.T. denotes the Con-Text dataset. (Metric: mAP)

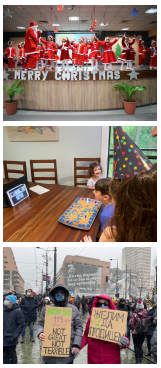
Model	Con-Text	Bottles	Activity
ViT	79.24	80.81	80.29
KnowBert	47.07	53.28	64.66
model A	81.47	82.26	81.79
model B	**89.53**	**85.26**	**87.45**

Table 5. The model A is trained in a separated manner. The model B is trained in an end-to-end manner. (Metric: mAP)

of visual features and textual features. Then, we discuss the impact of KARC and VKAC components. Finally, we show the advantage of jointly optimizing the whole network.

The impact of visual features As shown in Tab. 3, introducing textual cues (*i.e.*, Glove and fastText) to the ResNet-152 model can significantly improve the performance up to 3% mAP. However, ViT model performance improvement is not more than 1% mAP. We further compare the two models with qualitative examples via visualizing the attention map of both models (only trained with image data) of ResNet-152 and ViT. As depicted in Fig. 7, the ResNet-152 model mainly focuses on the visual contents. However, the ViT model captures the visual contents and harvest the textual cues from the image by self-attention mechanism. Thus, embedding features provides complementary information to boost the performances of ViT instead of solely exploiting the literal meaning of scene text.

The impact of knowledge-enhanced features As mentioned before, a direct way to mine knowledge is to exploit the BERT encoder output features. As shown in Tab. 3, the employment of knowledge-enhanced features from BERT achieves significant improvements than the typical word embedding features (GloVe/fastText). The ViT+BERT model surpasses the performance of the ViT+fastText model by 6.69%, 1.63%, 4.63% on Con-Text, Drink Bottle, and Crowd Activity. This superior performance proves that the explicit knowledge in knowledge bases significantly enriches the semantics of scene text for understanding objects. Furthermore, unlike BERT, KnowBert explicitly introduces knowledge from a knowledge base into the model. The experimental results show that the KnowBert model consistently outperforms the BERT model. Therefore, introducing knowledge behind scene text to neural network feature learning enhances understanding natural images. As shown in Fig. 8, the employment of knowledge substantially enriches the classification accuracy, as the knowledge behind "PM2.5" tells that the third image is about environment.



GT: Celebrate Christmas
ViT: Celebrate Christmas (0.9997)
ViT + fastText: Celebrate Christmas (1.0000)
ViT + KnowBert: Celebrate Christmas (0.9998)

GT: Birthday party
ViT: Teaching (0.9136)
ViT + fastText: Birthday party (0.8939)
ViT + KnowBert: Birthday party (0.8141)

GT: Protect environment
ViT: Appeal for peace (0.3469)
ViT + fastText: Racial equality (0.4324)
ViT + KnowBert: Protect environment (0.8439)

Figure 8. The classification results of different models.

The impact of KARC and VKAC components As shown in Tab. 4 Model B is the default model equipped with KARC and VKAC, while model A only employed KARC. The integration of KARC only in model A improves the performance on all datasets. Moreover, integrating VKAC on top of KARC in model B increases the recognition performance mAP by 2.28%, 1.67%, and 1.29% on Con-Text, Bottles, and Crowd Activity datasets, respectively. The experimental results demonstrate the effectiveness of fusing multi-modal features for this task.

Joint optimization Integrating the process of mining knowledge, feature extraction, and classification in a unified network makes it feasible to optimize them jointly. The model that is jointly optimized could achieve better performance than the one with separated feature extraction and classifier, as those processes are complementary to each other. To confirm this assumption, we first train the models of ViT and KnowBert with image data and scene text, respectively, at the supervision of the classification task. Then, the classifier and VKAC are trained, accepting as input visual features and knowledge-enhanced features extracted from the pre-trained models. As reported in Tab. 5, the model trained in an end-to-end manner significantly outperforms the one trained separately, showing the necessity of integrating knowledge mining process into the network.

5. Conclusion

In this paper, we have confirmed that the usage of the knowledge behind scene text can improve the performance of the fine-grained image classification task. Experiments on the two benchmark datasets and the proposed Crowd Activity dataset have verified the effectiveness and efficiency of our method for product recognition and crowd activity analysis. In the future, we will further explore the usage of knowledge mining of scene text on other tasks of multi-modal fusion, such as scene text, visual question and answering, and visual grounding.

Acknowledgements This work was supported by the National Natural Science Foundation of China 61733007.

References

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE TPAMI*, 2014. 2

[2] Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018. 1, 2, 5, 6

[3] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4290–4300, 2019. 2

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 2017. 5, 6

[5] Michal Busta, Yash Patel, and Jiri Matas. E2E-MLT - an unconstrained end-to-end method for multi-language scene text. In Gustavo Carneiro and Shaodi You, editors, *ACCV*, 2018. 6

[6] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. Differentiable patch selection for image recognition. In *CVPR*, pages 2351–2360, 2021. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2, 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 5, 6

[10] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017. 2

[11] Octavian-Eugen Ganea and Thomas Hofmann. Deep joint entity disambiguation with local neural attention. In *EMNLP*, 2017. 3

[12] ZongYuan Ge, Chris McCool, Conrad Sanderson, Peng Wang, Lingqiao Liu, Ian D. Reid, and Peter I. Corke. Exploiting temporal information for dcnn-based fine-grained object classification. In *DICTA*, 2016. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[14] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *EMNLP*, 2011. 3

[15] Sezer Karaoglu, Ran Tao, Theo Gevers, and Arnold W. M. Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE TMM*, 2017. 1, 2, 6

[16] Sezer Karaoglu, Jan C. van Gemert, ., and Theo Gevers. Context: text detection using background connectivity for fine-grained object classification. In *ACM MM*, 2013. 2, 5, 6

[17] Aisha Urooj Khan, Amir Mazaheri, Niels da Vitoria Lobo, and Mubarak Shah. MMFT-BERT: multimodal fusion transformer with BERT encodings for visual question answering. In *EMNLP*, 2020. 2

[18] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on FGVC*, 2011. 2

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 2

[20] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *ECCV*, 2020. 2

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[22] Andrés Mafla, Sounak Dey, Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Fine-grained image classification and retrieval by combining visual and locally pooled textual features. In *WACV*, 2020. 1, 2, 5, 6, 7

[23] Andrés Mafla, Sounak Dey, Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval. In *WACV*, 2021. 1, 5, 6, 7

[24] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, 2013. 2

[25] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 1995. 2

[26] Yair Movshovitz-Attias, Qian Yu, Martin C. Stumpe, Vinay D. Shet, Sacha Arnoud, and Liron Yatziv. Ontological supervision for fine grained classification of street view storefronts. In *CVPR*, 2015. 1, 2

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019. 6

[28] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 6

[29] Florent Perronnin and Christopher R. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007. 2

[30] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, 2018. 2

[31] Matthew E. Peters, Mark Neumann, Robert L Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *EMNLP*, 2019. 2, 5, 6

[32] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *EMNLP-IJCNLP*, 2019. 2

[33] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2

[34] Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul N. Bennett, and Saurabh Tiwary. Knowledge-aware language model pretraining. *CoRR*, 2020. 2

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 7

[36] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *ICCV*, 2019. 2

[37] Valentin I. Spitkovsky and Angel X. Chang. A cross-lingual dictionary for english wikipedia concepts. In *LREC*, 2012. 3

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2, 6

[39] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. In *CVPR*, 2021. 1

[40] Hao Wang, Pu Lu, Hui Zhang, Mingkun Yang, Xiang Bai, Yongchao Xu, Mengchao He, Yongpan Wang, and Wenyu Liu. All you need is boundary: Toward arbitrary-shaped text spotting. In *AAAI*, 2020. 2

[41] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 6

[42] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *EMNLP*, 2020. 6

[43] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015. 6