Zero-Shot Object Detection with Textual Descriptions

Zhihui Li,[1] Lina Yao,[1] Xiaoqin Zhang,[2] Xianzhi Wang,[3] Salil Kanhere,[1] Huaxiang Zhang[4]*
[1]School of Computer Science and Engineering, University of New South Wales
[2]College of Mathematics and Information Science, Wenzhou University
[3]School of Software, University of Technology Sydney
[4]School of Information Science and Engineering, Shandong Normal University

Abstract

Object detection is important in real-world applications. Existing methods mainly focus on object detection with sufficient labelled training data or zero-shot object detection with only concept names. In this paper, we address the challenging problem of zero-shot object detection with natural language description, which aims to simultaneously detect and recognize novel concept instances with textual descriptions. We propose a novel deep learning framework to jointly learn visual units, visual-unit attention and word-level attention, which are combined to achieve word-proposal affinity by an element-wise multiplication. To the best of our knowledge, this is the first work on zero-shot object detection with textual descriptions. Since there is no directly related work in the literature, we investigate plausible solutions based on existing zero-shot object detection for a fair comparison. We conduct extensive experiments on three challenging benchmark datasets. The extensive experimental results confirm the superiority of the proposed model.

Introduction

In the last decade, researchers have made promising progress in object detection Girshick (2015); Ren et al. (2015, 2017); Lin et al. (2017). Most of these achievements rely on the collection of large-scale labeled training data. Although researchers have struggled to acquire larger datasets with a broader set of categories, the processing procedure is time-consuming and tedious. Furthermore, it is impossible to collect enough training data for rare concepts, *i.e.* Okapia. Therefore, a challenging problem is how to simultaneously recognize and locate these novel object instances with no training samples.

Zero-shot learning has been widely used to tackle the problem of data scarcity Akata et al. (2016); Frome et al. (2013); Lampert, Nickisch, and Harmeling (2009); Zhang, Xiang, and Gong (2017); Zhang and Saligrama (2015); Rahman, Khan, and Porikli (2018); Mikolov et al. (2013). Most of these works focus on the concept classification problem. Although it remains a challenging and unsolved problem, there is still a large gap between the problem setting and real-world scenarios in the following aspects. Firstly, in most zero-shot learning benchmark datasets Welinder et al. (2010); Nilsback and Zisserman (2008); Russakovsky et al. (2015), each image has only one dominant object, while in real-world applications, multiple objects may appear in a single image. Secondly, most of the zero-shot classification methods are based on attributes and semantic descriptions, which cannot be directly applied to zero-shot detection in the entire scene image. Thirdly, the setting of zero-shot learning does not consider occlusions and clutter, which commonly exist in real-world applications. To close this gap, Rahman, Khan, and Porikli (2018) introduced a new "zero-shot object detection" (ZSD) problem setting method, which aims at concurrently detecting and recognizing novel instance in the absence of any training examples.

Although the data-scarcity challenge exists for a large number of categories in real-world applications, there is a massive amount of textual data for these categories. These data arrive in the form of dictionary entries, online encyclopedias and other online resources. For example, English Wikipedia has 5,645,010 articles on its site, which provides a rich knowledge base for different topics.

The major problem we focus on in this paper is how to simultaneously recognize and locate novel object instances using purely unstructured textual descriptions with no training samples. In other words, our goal is to concurrently link visual image features with the semantic label information where the descriptions of novel concepts are presented in the form of natural languages, *i.e.* online encyclopedias. We design a novel deep learning framework for zero-shot object detection with textual description. The proposed network takes a description and an image as input and outputs the affinities between the description and the object proposals in the image. We process the textual description in a word-by-word fashion with word-LSTM. For each word in the description, we achieve unit-level attentions for different units using the LSTM. Each unit determines whether a specific object pattern exists in the object proposal. The contributions of different units are weighted by the visual-unit attention mechanism. To step further, we also study word-level attention which learns the importance of different words for adaptive word-level weighting. We achieve the final affinity by averaging over all units' responses for all words. We conduct experiments to confirm that both visual unit-level attention and word-level attention contribute to the good performance of the proposed

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

model. Compared with the related works in the literature, we go beyond the traditional object recognition setting, and explore the knowledge in the natural language descriptions to detect and recognize novel (unseen) concepts.

To sum up, we make the following contributions in this work.

- We pose and address a challenging problem of zero-shot object detection with textual descriptions, which aims to simultaneously detect and recognize unseen objects by exploring natural language description. To the best of our knowledge, this task has not yet been explored in the computer vision and machine learning communities.
- We propose a novel deep learning framework to jointly learn the visual units, visual-unit attention, and word-level attention, which are combined to achieve word-proposal affinity by an element-wise multiplication.
- We investigate plausible solutions based existing zero-shot object detection, and establish baselines on the zero-shot object detection with textual descriptions.
- We conduct extensive experiments and component studies to demonstrate the superiority of the proposed model for zero-shot object detection with textual descriptions.

The rest of this paper is organized as follows. We first review the related works on zero-shot learning and object detection. Then we introduce the proposed framework for zero-shot object detection with unstructured textual description. After that, we present the experimental results, followed by the conclusion and a discussion of future work.

Related Works

We briefly review the related work on zero-shot learning, object detection and language and vision. Due to space limitations, we cannot do justice to the entire body of literature.

Zero-Shot Learning: Existing zero-shot learning algorithms exploit different methods to transfer the knowledge from seen concepts to unseen ones. These algorithms can be generally grouped into two categories: projection-based and similarity-based methods. The projection-based methods measure the relatedness between the test samples and the unseen concepts with the projected features of the visual features in the semantic space Lampert, Nickisch, and Harmeling (2009); Akata et al. (2016); Zhang and Saligrama (2015); Romera-Paredes and Torr (2015). Specifically, Akata et al. (2016) first project the visual features and the semantic embeddings of concepts onto a common space, and then measure the relatedness between the data point and the concept. Romera-Paredes and Torr (2015) combine a linear model together with a principled choice of regularizers to achieve a simple yet efficient method. In contrast, the similarity based methods Norouzi et al. (2013) exploit the discrete classifiers trained on the seen concepts in the visual feature space to determine how close the novel instance is to the seen concepts.

Object Detection: Researchers have demonstrated the superiority of object proposal based methods for detecting objects within an image Girshick et al. (2014); Girshick (2015); Ren et al. (2015). Specifically, Girshick et al. (2014) employ an off-the-shelf object detector to generate object proposals, which are cropped and warped by a R-CNN framework. To take this one step further, the authors introduce a Region-of-Interest (RoI) pooling method in Girshick (2015). Their method shares the feature computation for all the proposal regions, which greatly improves the effectiveness of their algorithm. In Ren et al. (2015), Ren *et al.* replace the off-the-shelf object detector with a region proposal network (RPN), which shares full-image convolutional features with the detection network. Although these object detection algorithms work well on pre-defined concepts, they cannot be directly applied to novel concepts.

Vision and Language: Recent years have witnessed the rapid progresses of recurrent neural networks (RNN) for vision and language tasks, *e.g.*, image/video caption generation, visual question answering, visual-semantic embedding, *etc.* Mao et al. (2016) propose a framework for joint generation and comprehension, which can generate an unambiguous description for objects or regions in an image. In Yu et al. (2016), a new model is developed to incorporate detailed context into referring expression models.

The goal of visual-semantic embedding is to project both images/videos and languages onto a common space for subsequent classification or retrieval tasks Frome et al. (2013); Karpathy and Li (2015); Reed et al. (2016); Liu et al. (2015). For example, Reed et al. (2016) propose training an end-to-end CNN-RNN network to project the images and fine-grained visual descriptions onto a common feature space for zero-shot learning. In Liu et al. (2015), a multitask deep visual-semantic embedding model is trained to learn the multi-view distance between the video side semantic information and visual content.

Recent Progress on ZSD: We have noticed that there are three contemporary works on zero-shot object detection Bansal et al. (2018); Rahman, Khan, and Porikli (2018); Demirel, Cinbis, and Ikizler-Cinbis (2018). Bansal et al. (2018) develop background-aware models to solve the ZSD problem. Rahman, Khan, and Porikli (2018) propose a sematnic clustering loss. Demirel, Cinbis, and Ikizler-Cinbis (2018) employ hybrid region embedding to improve the performance. Although these methods share a common theme of zero-shot detection with us, the proposed model significantly differ from them. All of these methods focus on exploiting the visual or semantic similarity between seen classes and unseen classes, while we propose to detect unseen concepts by exploring their natural language description.

Zero-Shot Object Detection using Textual Description

In this section, we describe the building of our framework for zero-shot object detection using textual description. The proposed framework is agnostic to the choice of base object detection models. We use Faster R-CNN Ren et al. (2015) as the backbone architecture for its simplicity and state-of-the-art performance. We firstly introduce the structure of the proposed framework and then discuss its training procedure.

The Model Architecture

The challenge for solving the problem of zero-shot detection with textual descriptions is to effectively build word-proposal relations. Given each word in the textual description, the network should be able to determine whether the word with its context fit the object proposal Li et al. (2017). For a textual description, we investigate all these word-proposal relation, weight the confidences of all the relations, and aggregate them to achieve the final description-proposal affinity.

Based on the intuition mentioned above, we propose a novel deep learning framework to explore word-proposal relations and determine the affinity between a textual description and an object proposal. We show the overall network architecture in Figure 1. The network contains three branches: a visual branch, a language branch and a bounding box regression branch. The visual branch outputs visual unit activations, each of which determines whether certain object patterns exist in the object proposal. The language branch employs a recurrent neural network with long short-term memory units (LSTM). For each word in the description, it generates unit-level attention and word-level attention to weight the visual units from the visual branch. The unit-level attention determines which units should be paid more attention to. We weight all the units' activations by both unit-level attention and word-level attention, and achieve the final affinity.

Visual Feature Encoding: Given an arbitrary-sized image $\mathbf{x}$, the Faster R-CNN framework employs a ConvNet (*e.g.* VGG or ResNet) to extract the intermediate convolutional activations, which are used as feature maps. The RPN operates on these feature maps and outputs a set of candidate rectangular object proposals, each with a confidence score. Since these high-scoring object candidate proposals may be of different sizes, the framework uses a RoI pooling layer to project them to a fixed dimensional representation. We represent the fixed dimensional representation for each candidate object proposal as $\mathbf{f}$. The RPN is generic because it outputs object proposals based on an objectness measure. Thus, a pre-trained RPN on seen concepts can be directly applied to generate object proposals for unseen novel concepts. In the remainder of our section on network architecture, we use these feature representations to learn useful representations for both seen and unseen concepts.

Visual-Language Gated Attention: We propose a visual-language gated attention mechanism to effectively mine word-proposal relations. Intuitively, for a given word, the network should assign larger weights to the visual parts which have similar semantic meanings.

We employ an LSTM network Hochreiter and Schmidhuber (1997) to train the language part of the framework because of its ability to capture the temporal relations of sequential data. For a description of each semantic concept, the LSTM network outputs attentions for different visual units word by word. We first encode these words into K-dimensional one-hot vectors, where K is the size of the vocabulary. Given the description of a concept, we use a fully connected layer ("w-fc1") to transform the t-th word to its corresponding embedded feature vector x_w^t. Furthermore, we add two fully connected layers ("v-fc1" and "v-fc2") after the fixed dimensional representation for each object proposal, resulting in the visual features x_v for the word-LSTM. Then, at every step, we feed the concatenation of x_v and x_t^w to the LSTM as input.

The LSTM model contains a memory cell c_t, which is used to memorize the history of the previous steps. It has three controlling gates, *i.e.* input gate i_t, forget gate f_t, and output gate o_t. The LSTM model uses these controlling gates to control the update and flow direction of information. In the literature, there are many variants of LSTM and we use the LSTM proposed in Zaremba and Sutskever (2015), which iterates in the following fashion:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \tag{1}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \tag{2}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \tag{3}$$

$$g_t = \phi(W_{xg}x_t + W_{hg}h_{t-1} + b_g), \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \quad h_t = o_t \odot h(c_t), \tag{5}$$

where σ represents the sigmoid function, $\odot$ denotes the element-wise product, and W and b are parameters to be optimized.

For each word, we propose to generate the unit-level attentions by feeding the output hidden state h_t into a fully connected layer with a LeakyReLU Xu et al. (2015) function and a fully connected layer with a softmax function, resulting in the attention vector $A_t \in \mathbb{R}^{512}$. Note that the visual units v and the attention vector have the same dimension. Hence, we have the affinity between the concept description and the object proposal at the t-th word as follows:

$$a_t = \sum_{n=1}^{512} A_t(n) v_n, \tag{6}$$

$$\textit{s.t.} \sum_{n=1}^{512} A_t(n) = 1, \tag{7}$$

where $A_t(n)$ is the attention value for the n-th visual unit. Since each visual unit denotes the existence of certain visual object patterns and the attention values A_t tell us which visual units should more attention be placed, it is intuitive to obtain the affinity between the concept description and the object proposal at the t-th word by element-wise multiplication of the two vectors. Finally, we compute the description-proposal affinity by summing up the affinity of all words, *i.e.* $a = \sum_{t=1}^{T} a_t$, where T is the number of words in the concept description.

To this end, we have used the unit-level attention to associate the most related units to each word in the concept description. To force different visual units' attentions to compete against each other, we use a softmax non-linearity function in the framework. In our work, we find this solution works well for learning effective unit-level attentions.

However, we still have ignored the varying importance of different words for learning the description-proposal affinity. For example, the word "stripe" conveys more information than the word "the". Therefore, it is important for us to design an effective approach to learn the weights of different words. We propose learning the word-level attention by mapping the

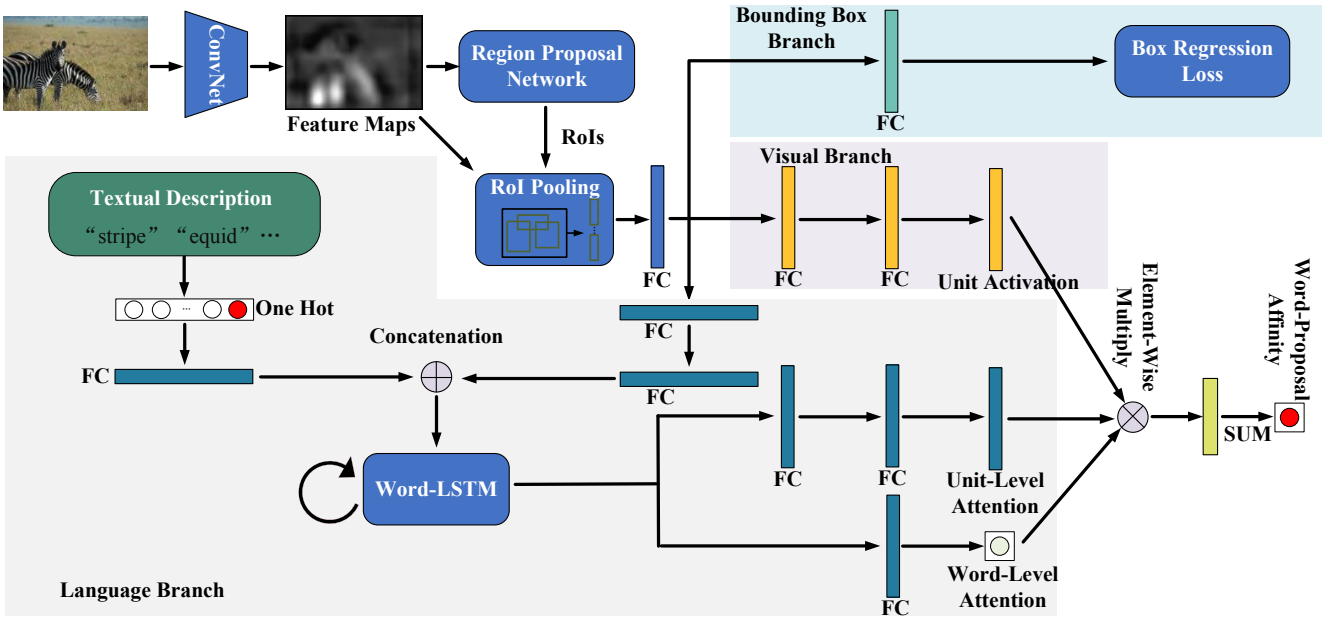

Figure 1: Overview of the proposed framework. It contains three branches: a visual branch, a language branch, and a bounding box regression branch. Different background colors are used to denote three branches. We use the visual branch to generate visual units, which capture the object patterns of the proposals. The language branch learns visual unit-level and word-level attentions for weighting visual units for each word in the description.

hidden state h_t of the LSTM via a fully-connected layer with sigmoid non-linearity function $g_t = \sigma(W_g h_t + b_g)$, where W_g and b_g are the parameters of the fully-connected layer to be optimized.

Then we use both visual unit-level attention and the word-level attention to compute the description-proposal affinity at the t-th word by $\hat{a}_t = g_t \sum_{n=1}^{512} A_t(n) v_n$. The final affinity can be computed by the aggregation of affinities of all words $\hat{a} = \sum_{t=1}^{T} \hat{a}_t$.

The third branch is for bounding box regression, the goal of which is to predict precise location of objects by adding proper offsets to the proposals to align them with the ground truths. We implement this branch in the same way as in Faster-RCNN Ren et al. (2015) except we change the parameterizations of the 4 coordinates as follows:

$$t_x = \log(\frac{|x - x_a|}{w_a}), \quad t_y = \log(\frac{|y - y_a|}{h_a}), \tag{8}$$

$$t_w = \log(\frac{w}{w_a}), \quad t_h = \log(\frac{h}{h_a}), \tag{9}$$

$$t_x^* = \log(\frac{|x^* - x_a|}{w_a}), \quad t_y^* = \log(\frac{|y^* - y_a|}{h_a}), \tag{10}$$

$$t_w^* = \log(\frac{w^*}{w_a}), \quad t_h^* = \log(\frac{h^*}{h_a}). \tag{11}$$

Different from the widely used version in Girshick et al. (2014), we transform the first two elements using a $\log(\cdot)$ function to count more on close-by objects. We make this change based on the intuition that we need to model distant objects while original bounding box regression only considers closer objects.

Training Procedure

We adopt a three-step training mechanism to optimize the parameters. In the first step, we train the backbone Faster-RCNN with the seen concepts. The weights of shared layers are initialized using the Inception-ResNet v2 model (pre-trained on ImageNet) Szegedy et al. (2017). Then we train the RPN, classification and detection networks. In the second step, we train the visual-branch on the training set. This step is used to generate a series of visual unit activations, each of which encodes certain object patterns. When we jointly train the whole network, only the newly added fully connected layers are optimized. In the third step, we jointly optimize the proposed model.

We minimize an objective function following the multi-task loss in Faster R-CNN Ren et al. (2015). We define the loss function for an image as follows:

$$\mathcal{L}(b_i, y_i, b_i^*, \hat{a}_i, \theta) = \frac{1}{T} \sum_i (\mathcal{L}(\hat{a}_i, y_i) + \mathcal{L}(b_i, b_i^*)), \tag{12}$$

where θ denotes the parameters of the deep neural network, $\hat{a}_i$ is the output of the classification branch, $T = N \times R$ is the total number of RoIs in the N-image training set. b_i is a vector denoting the four parameterized coordinates of the predicted bounding box, and b_i^* is that of the ground-truth box associated with a positive anchor. $\hat{a}_i$ denotes the predicated affinity for the i-th object proposal. y_i represents the ground-truth of the i-th object proposal.

Table 1: Experiment comparisons for zero-shot object detection with textual description on the ILSVRC-2017 detection dataset. Mean average precision (mAP) is used as the evaluation metric. We present the results in percentages. A larger mAP indicates better performance.

	p.box	syringe	harmonica	maraca	burrito	pineapple	electric-fan	iPod	dishwasher	can-opener	plate-rack	bench	bow-tie	s.trunks	scorpion	snail	hamster	tiger	ray	train	unicycle	g.ball	h.bar	**mAP**
SAN Rahman, Khan, and Porikli (2018)	5.9	1.5	0.3	0.2	40.6	2.9	7.7	28.5	13.3	5.1	7.8	5.2	2.6	4.6	68.9	6.3	53.8	77.6	21.9	55.2	21.5	31.2	5.3	20.3
SB Bansal et al. (2018)	6.8	1.8	0.8	0.5	43.7	3.8	8.3	30.9	**15.2**	6.3	8.4	6.8	3.7	6.1	71.2	7.2	**58.4**	79.4	23.2	58.3	23.9	34.8	6.5	22.0
DSES Bansal et al. (2018)	7.4	2.3	1.1	0.6	46.2	**4.3**	8.7	32.7	14.6	6.9	9.1	7.4	4.9	6.9	73.4	7.8	56.8	80.8	24.5	59.9	**25.4**	33.1	7.6	22.7
LAB Bansal et al. (2018)	6.5	1.6	0.7	0.5	44.1	3.6	8.2	30.1	14.9	6.4	8.8	6.4	4.1	4.8	69.9	6.9	57.1	80.2	23.6	58.2	25.1	35.6	7.2	21.9
Ours	**7.8**	**3.1**	**1.9**	**1.1**	**49.4**	4.0	**9.4**	**35.2**	14.2	**8.1**	**10.6**	**9.0**	**5.5**	**8.1**	**73.5**	**8.6**	57.9	**82.3**	**26.9**	**61.5**	24.9	**38.2**	**8.9**	24.1

For the detection-classification branch, we minimize the following cross-entropy loss:

$$\mathcal{L}(\hat{a}_i, y_i) = -(y_i \log \hat{a}_i + (1 - y_i) \log(1 - \hat{a}_i)). \quad (13)$$

The part of regression loss is similar to faster-RCNN regression loss, except we make changes to the coordinates as in Equation (4) and Equation (5).

Experiments

In this section, we describe the extensive experiments which were conducted to evaluate the proposed framework. We first discuss the detailed experimental setup, then we discuss quantitative results on three challenging datasets. After this, we conduct component studies to analyze the effects of different components. Finally, we present some qualitative results on the used datasets.

Experimental Setup

Dataset Description

Three challenging datasets have been used in this paper to evaluate the performance.

ILSVRC-2017 detection dataset Russakovsky et al. (2015) constitutes of 200 basic-level object categories. These categories were carefully selected in terms of different factors such as object scale, level of image clutterness, and many others. Following Rahman, Khan, and Porikli (2018), we choose 23 categories as unseen concepts, and the rest are as unseen concepts.

MSCOCO Lin et al. (2014) was collected for object detection and semantic segmentation tasks. There are multiple object instances per image with variations in occlusion, clutter, views, *etc.* We use training samples from the 2014 training set. Since we do not have ground-truth for the test images, we randomly select images from the validation set for testing.

VisualGenome (VG) Krishna et al. (2017) was designed primarily for visual relationship understanding. The authors also provide bounding boxes for multiple objects in the images. We use this dataset because it contains bounding box information for a large number of classes. Following Bansal et al. (2018), we use images from part-1 for training, and randomly sample from part-2 for testing.

Train/Test Split: For the zero-shot learning setting, we are not allowed to use any visual examples of unseen concepts. In terms of the ILSVRC-2017 detection dataset, we use the same train/test split as in Russakovsky et al. (2015). Please refer to Russakovsky et al. (2015) for details of the split. For the MSCOCO and VisualGenome datasets, we follow the same procedure as in Bansal et al. (2018). Specifically, we use 48 training classes and 17 test classes for MSCOCO, and 478 training classes and 130 test classes for VisualGenome.

Compared Methods: To the best of our knowledge, this is *the first work* for zero-shot object detection using textual description. Hence, there are no directly related algorithms with which to compare with. However, we have tried our best to implement two state-of-the-art baselines for comparison. We extend Rahman, Khan, and Porikli (2018) and Bansal et al. (2018) to our problem setting by replacing the word embedding with textual description embedding using fastText Grave et al. (2017).

Evaluation Metric: For ILSVRC-2017 detection dataset, we use the commonly used evaluation metric, mean Average Precision (mAP), to evaluate the performance of novel object detection. We use this evaluation metric because it has been widely used in supervised object detection tasks. Following Bansal et al. (2018), recall is used as the evaluation metric for the MSCOCO and Visual Genome datasets. This is because that it is infeasible to exhaustively label bounding box annotations for all instances of an object. mAP is very sensitive to missing annotations and will count these detection results as false positives.

Implementation Details: We adopt a three-step training mechanism to optimize the parameters. Following Rahman, Khan, and Porikli (2018), we rescale the shorter size of images to 600 pixels. To reduce redundancy, we employ non-maximum suppression (NMS) on proposals class probability with IoU threshold equals 0.7. During training, we use the Adam optimizer with learning rate 10^{-5}. We implement the proposed model based on the open-source package PyTorch. The code and models will be released.

Quantitative Analysis

Extensive experiments have been conducted to evaluate the performance of the proposed model against the state-of-the-art baselines on three challenging benchmark datasets. We report the mAP for all the compared models on the ILSVRC-2017 detection dataset in Table 1. *Note that after careful re-implementation, all the baselines have used description embedding instead of original word embedding, for a fair comparison.* From these experimental results, we have the following observations: (1) The proposed model generally performs better than the other compared models on most of the objects, achieving 1.4% improvement against the second best baseline. This is very significant in this challenging

Table 2: Experimental comparisons for zero-shot object detection with textual description on the MSCOCO and Visual Genome (VG) datasets. Recall@100 is used as the evaluation metric. We show the performance in percentages. (Larger recall is better.)

	MSCOCO			Visual Genome		
IoU	0.4	0.5	0.6	0.4	0.5	0.6
SAN Rahman, Khan, and Porikli (2018)	35.7	26.3	14.5	6.8	5.9	3.1
SB Bansal et al. (2018)	37.8	28.6	15.4	7.2	5.6	3.4
DSES Bansal et al. (2018)	42.6	31.2	16.3	8.4	6.3	3.3
LAB Bansal et al. (2018)	35.2	22.4	12.1	8.6	6.1	3.3
Ours	**45.5**	**34.3**	**18.1**	**9.7**	**7.2**	**4.2**

Table 3: Experimental results of different components on the three challenging benchmark datasets. mAP is used as the evaluation metric for ILSVRC and Recall@100 is used for the other two datasets. Larger value indicates better performance for both evaluation metrics.

	MSCOCO			Visual Genome			ILSVRC
	0.4	0.5	0.6	0.4	0.5	0.6	
Ours w/o pre-train	30.8	21.2	8.6	2.4	1.1	0.6	12.2
Ours w/o word-level att.	36.5	28.1	13.2	4.6	3.2	1.8	18.4
Ours w/o visual att.	22.7	15.1	6.3	1.8	0.8	0.3	9.2
Ours full model	**45.5**	**34.3**	**18.1**	**9.7**	**7.2**	**4.2**	**24.1**

zero-shot object detection setting. (2) The novel concepts (*i.e.* iPod, scorpion, tiger) which have similar concepts in the training set have much better performance than those (*i.e.* pineapple, bowtie, maraca) without any similar concepts. For example, zero-shot object detection on iPod of the proposed model achieves 35.2%, while it only achieves 5.5% on bowtie. This indicates the challenge of zero-shot object detection. (3) We also notice that the performances of our re-implemented Rahman, Khan, and Porikli (2018) and Bansal et al. (2018) are generally better than the results reported in the original paper. We think this improvement is achieved because there is more information contained in the textual description than single concept name. This phenomenon confirms the benefits of exploring textual description, *i.e.* online encyclopedias.

We also report the extensive experiment results in terms of Recall@100 for all the compared baselines on the MSCOCO and VG datasets in Table 2. Three different IoU overlap thresholds (*i.e.* 0.4, 0.5, 0.6) were used in these experiments. In this part, we use the threshold of IoU ≥ 0.5 as an example for convenient discussion. From the experimental results, we observe that the proposed model generally perform much better than the other compared baselines. Specifically, the proposed model increases the recall to 34.3% from 26.3% achieved by the baseline method SAN Rahman, Khan, and Porikli (2018) on MSCOCO, and increases the recall to 7.2% from 5.9% on Visual Genome, which indicates that Visual Genome is much more challenging than MSCOCO. We also make the observations that by exploring textual description we obtain much better performance than the results reported in the original paper, especially on the challenging dataset Visual Genome. For example, the performance of LAB Bansal et al. (2018) improves from 5.4% to 7.2% by exploring the embedding of textual description instead of a single concept name.

Component Studies

In this part, we describe the extensive experiments to study the effects of different components in our model, and report the experiment results. Following previous settings, we still use mAP as the evaluation metric for the ILSVRC dataset and Recall@100 for the MSCOCO and Visual Genome datasets. As discussed in the model section, we first pre-train the network. From the experimental results, we can see that without pre-training, the performance of zero-shot object detection dropped dramatically on all the datasets. This indicates that the pretraining significantly affects the final performance. To evaluate the effectiveness of the visual-unit-level attention and word-level attention, we compare with two variants. For the variant without word-level attention, we treat all the words in the description equally. For the variant without visual-unit-level attention, we use average pooling instead. From the experimental results, it can be seen that both components contribute greatly to the whole model.

Qualitative Results

In this section, we present the detection results by the proposed model. Visual Genome is used as an example here. The detection results are shown in Figure 2. These examples confirm that the proposed model is capable of detecting novel concepts with textual descriptions, although there are some false positives. For example, the "deck table" was recognized as a "chair" in the fourth example. We think this is because the chair and the deck table in this image are visually similar, which makes the system hard to distinguish. In the future, we will continue improving the system to solve this problem.

Figure 2: Selected examples of zero-shot object detection. We can see that the detection results are reasonable, which demonstrates the effectiveness of the proposed model for object detection in a zero-shot setting.

Conclusions and Future Work

In this paper, we have made the first attempt at zero-shot object detection with natural language description. To solve this challenging problem, we have proposed a joint model to learn both visual-unit-level and word-level attention. In the experiment part, we described the extensive experiments which were conducted to demonstrate the superiority of the proposed model, and investigated the effectiveness of different components. In the future, we will incorporate the relational network into the framework to further improve the performance.

Acknowledgements

This work is supported in part by the National Natural Science Joint Fund Key Project (NSFC-General Technology Fundamental Research Joint Fund) [grant no. U1836216], in part by the National Natural Science Foundation of China [grant no. 61772322, and grant no. 61472285], in part by the Zhejiang Provincial Natural Science Foundation [grant no. LR17F030001], in part by the project of science and technology plans of Zhejiang Province (Grants no. 2015C31168), in part by the Key Innovative Team Support and Project of science and technology plans of Wenzhou City [grant nos. G20150017, ZG2017016].

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2016. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(7):1425–1438.

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Di-

vakaran, A. 2018. Zero-shot object detection. *CoRR* abs/1804.04340.

Demirel, B.; Cinbis, R. G.; and Ikizler-Cinbis, N. 2018. Zero-shot object detection by hybrid region embedding. *CoRR* abs/1805.06157.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*, 2121–2129.

Girshick, R. B.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Girshick, R. B. 2015. Fast R-CNN. In *ICCV*.

Grave, E.; Mikolov, T.; Joulin, A.; and Bojanowski, P. 2017. Bag of tricks for efficient text classification. In *EACL*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Karpathy, A., and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *CVPR*.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: common objects in context. In *ECCV*.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature pyramid networks for object detection. In *CVPR*.

Liu, W.; Mei, T.; Zhang, Y.; Che, C.; and Luo, J. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Nilsback, M., and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *ICVGIP*.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.; and Dean, J. 2013. Zero-shot learning by convex combination of semantic embeddings. In *NIPS*.

Rahman, S.; Khan, S. H.; and Porikli, F. 2018. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. *CoRR* abs/1803.06049.

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

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*.

Ren, S.; He, K.; Girshick, R. B.; Zhang, X.; and Sun, J. 2017. Object detection networks on convolutional feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(7):1476–1481.

Romera-Paredes, B., and Torr, P. H. S. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Xu, B.; Wang, N.; Chen, T.; and Li, M. 2015. Empirical evaluation of rectified activations in convolutional network. *CoRR* abs/1505.00853.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *ECCV*.

Zaremba, W., and Sutskever, I. 2015. Learning to execute. In *ICLR*.

Zhang, Z., and Saligrama, V. 2015. Zero-shot learning via semantic similarity embedding. In *ICCV*.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In *CVPR*.