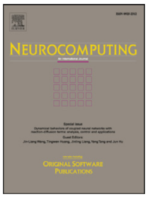
Neurocomputing 424 (2021) 117–124

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Brief papers

Instance search based on weakly supervised feature learning

Jie Lin, Yu Zhan, Wan-Lei Zhao*

Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Xiamen 361005 Fujian, China

ARTICLE INFO

Article history:
Received 14 June 2019
Revised 17 October 2019
Accepted 29 November 2019
Available online 4 December 2019

Communicated by Dr. Xinmei Tian

Keywords:
Instance search
Convolutional neural network
Weakly supervised learning

ABSTRACT

Instance search has been conventionally addressed as an image retrieval issue. In the existing solutions, traditional hand-crafted features and global deep features have been widely adopted. Unfortunately, since the features are not directly derived from the exact area of an instance in an image, satisfactory performance from most of them is undesirable. In this paper, a compact instance level feature representation is proposed. The scheme basically consists of two convolutional neural network (CNN) pipelines. One is designed for localizing potential instances from an image, while another is trained to learn object-aware weights to produce distinctive features. The sensitivity to the unknown categories, the distinctiveness to different instances, and most importantly, the capability of localizing an instance in an image are all carefully considered in the feature design. Moreover, both pipelines only require image level annotations, which makes the framework feasible for large-scale image collections with variety of instances. To the best of our knowledge, this is the first piece of work that builds the instance level representation based on weakly supervised object detection.

1. Introduction

Instance search is a retrieval task that allows users to launch a query with specified visual instance in an image. It requires the system to return the images/videos as well as the location (usually given as a bounding box), where the specified instance is in presence. In comparison to conventional image search, it reflects better the real needs from practice. For instance, a typical query could be a bag worn by a woman on a shopping website, a person in the streetview or a logo on a bottle, etc. Instance search also facilitates the task such as image hyper-linking [1], in which images are hyper-linked to each other via the instances shared in common. In the state-of-the-arts, due to the lack of instance level feature representation, this issue has been addressed by the approaches that are originally designed for content-based image retrieval [2–7]. Most of the features are built either on the image level or on the sub-region level despite they are hand-crafted features or trained deep features.

In the last decade, instance search has been treated as a subimage retrieval task. The solution to the instance search is dominated by hand-crafted local features such as scale invariant feature transform (SIFT) [8] and speeded-up robust features (SURF) [9], etc. These features are extracted from saliency regions of images. They are usually invariant to basic geometric transformations such as scaling and rotation. The instance search is therefore addressed as a point-to-point feature matching problem between the query and candidate images. Although encouraging results are achieved [10], the latent difficulties are hard to overcome. First of all, there are usually several hundreds to several thousands local features extracted from one image. The computation cost of point-to-point matching would be prohibitively high given there are millions of images to be compared. Although this issue has been alleviated by the encoding schemes such as bag-of-visual word (BoVW) [11], vector of locally aggregated descriptors (VLAD) [12] and Fisher vector (FV) [13], the features from different instances are embedded into one vector, which makes the instance level comparison hard to achieve or simply impossible. In addition, image local features are vulnerable to object deformation and out-plane rotation that are widely observed in the real world.

In recent years, due to the great success of convolutional neural networks (CNNs) in many computer vision tasks such as image classification [14–16], object detection [17–19] and instance segmentation [20,21], CNNs have been gradually introduced to image retrieval [22–27]. In the common practice of recent research, CNNs are trained to be more sensitive to object regions [23,25,26,28]. During the feature extraction, higher weights are assigned to objects regions on the feature maps. Such that resulting features are more representative for the latent objects in an image. Unfortunately, this type of feature representation only produces a single

[*] Corresponding author.
E-mail address: wlzhao@xmu.edu.cn (W.-L. Zhao).

https://doi.org/10.1016/j.neucom.2019.11.029

vector for one image. Features from different objects are simply mixed up.

In the recent literature, several attempts have been made to produce instance level representation via deep learning framework. Approaches presented in [24,29] are able to produce instance level feature representation by pooling features from the detected object region. Although satisfactory performance is achieved, the training of the network requires object level or pixel level annotations, which are too expensive to be deployed in the real applications. In addition, the strong reliance on the visual category annotations also makes the CNNs only sensitive to known classes. While in practice, the instance representation is required to be sensitive to the known as well as the unknown categories[1], given the fact that one cannot restrict the query instances to the annotated categories only.

In this work, motivated by the recent advances of the weakly supervised object detection methods [30–33], a novel instance level feature representation is proposed. In our solution, two convolutional neural network pipelines are integrated. One is designed to localize the visual instances from images, another is designed to derive features from the regions supplied by the first pipeline. Both pipelines require only the image level annotations. The advantages of such design are at least two folds.

- In the first pipeline, the bounding boxes of the visual instances are produced with the weak guidance of class information, which makes it still sensitive to instances of unknown categories.
- In the second pipeline, the trained network is able to capture the dissimilarity between different visual objects. Such that the produced features remain discriminative to each other even the corresponding instances are from the same category.

The reminder of this paper is organized as follows. The works related to visual object detection are reviewed in Section 2. Our solution to the weakly supervised instance search is presented in Section 3. The performance evaluation about our method in comparison to the state-of-the-art methods is presented in Section 4. Section 5 concludes the paper.

2. Related work

Instance-wise feature representation is preferred over global feature in instance search task since it requires instance level comparison and localization. Visual object detection therefore becomes the key step in the instance level feature design. In this section, the fully supervised and weakly supervised object detection, which are the most relevant to our work, are reviewed.

In general, there are two popular deep learning frameworks in fully supervised object detection. One is two-stage detection framework. It first produces a set of region proposals and then refine them by CNNs. Region convolutional neural network (R-CNN) [34], Fast R-CNN [17] and Faster R-CNN [18] are the representative methods in this category. Another one is one-stage detection framework, which gets increasingly popular in recent years. The representative methods are You Look Only Once (YOLO) [35] and single shot multibox detector (SSD) [36]. They are more efficient over two-stage methods since no proposal generation step is involved. Generally, two-stage methods outperform one-stage methods in terms of detection accuracy. Encouraging performance has been observed when fully supervised object detection or instance segmentation is adopted for instance level feature representation [24,29]. Nevertheless, the object level or pixel level annotations are required for the training, which is too expensive to be tractable in the large-scale context. In addition, the trained model becomes insensitive to unknown classes.

As a result, weakly supervised object detection (WSOD), which only requires image level annotations, is preferred. There are several popular WSOD methods [30–33] in recent literature. They all follow the pipeline of multiple instance learning (MIL) [37]. In the pipeline, an image is taken as a bag of proposals. Each proposal in the image is fed to the networks to check whether it keeps an visual object that is of the same category as image class label. The disadvantage of MIL is that the detected bounding box may not cover a complete latent object. In order to alleviate this issue, method in [33] refines several branches of instance classifiers online based on the outputs of previous branches. According to the method, the proposals are produced on raw images, while the proposal refinement is undertaken on the feature maps instead. This inconsistency affects precision of the refined proposals. Apart from MIL, there are some proposal-free methods by mining on the salient regions from deep feature maps and class activation maps. In [38], an adversarial complementary learning method is proposed to discover new and complementary object regions by erasing the discovered regions from the deep feature maps gradually. However this method fails when more than one objects from the same category are in presence within an image. Method from [39] boosts the performance of object detection by mining on the class activation map, which roughly reflects the object region. In order to enhance the localization accuracy, method in [40] adopts both the re-training and re-localization steps during the training. Unfortunately, extra training set with the object level annotations is required.

Due to the latent issues of one way or another in the existing methods, aforementioned WSOD pipelines cannot be directly adopted for instance search task. In our design, two WSOD pipelines, namely proposal clustering learning (PCL) [33] and soft proposal network (SPN) [39] are tailored to fitting into our instance feature representation. Namely, PCL is adopted mainly to produce bounding boxes for the latent objects of known and unknown categories. Its performance is further enhanced by replacing Selective Search [41] with EdgeBoxes [42] for object proposal generation. In addition, in order to avoid background interference, SPN is introduced to assign object-aware weights on bounding boxes to produce a more discriminative instance feature representation.

3. The proposed method

In this section, we firstly introduce the overall framework of the proposed method in Section 3.1. Thereafter the instance localization which is based on weakly supervised learning method, is given in Section 3.2. With the object bounding boxes supplied by the localization pipeline, the CNN pipeline designed for feature map weighting and feature extraction is introduced in Section 3.3.

3.1. Overall framework

As discussed in Section 1, the existing instance level representations require either object level or pixel level annotations for the training set, which makes it hardly feasible in real scenarios. In contrast to these works, in our design instance level feature representation is produced via weakly supervised learning. It only requires image level class labels for the training set. In addition, since the bounding box (usually given as a rectangle) does not cover the exact shape of an object, direct feature extraction from within the region may introduce the noises from the background. To address this issue, a spatial weighting network is introduced to generate object-aware weights for features extracted from the refined proposals. The framework of our method basically consists

[1] They are not in any of the annotated classes.

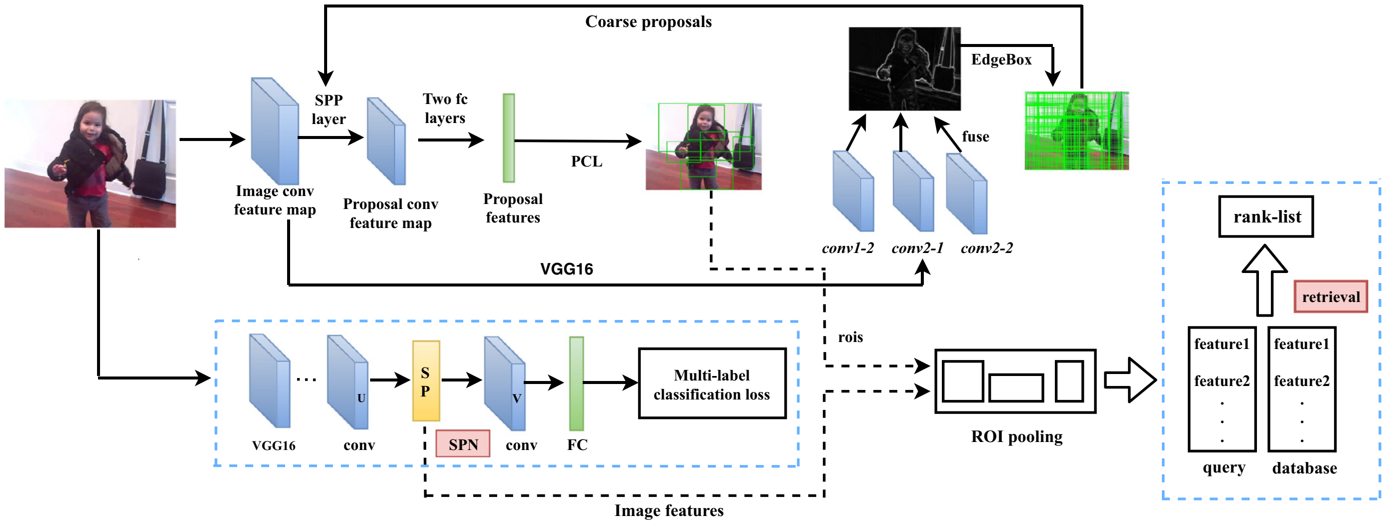


Fig. 1. The framework of weakly supervised instance level feature learning. The upper flow shows the instance localization procedure, where coarse proposals are generated with edge information from shallow convolution layers of VGG-16. The image and its coarse proposals are therefore fed into a PCL [33] stream to obtain refined proposals. The flow on the lower side is the pipeline for assigning object-aware weights on the feature maps of convolutional layers. The first part is nothing more than a soft proposal network [39] (inside the blue dashed box). Then feature for each instance is produced by ROI pooling from the proposals supplied by the upper flow on the weighted feature map of SPN.

of two CNN pipelines. One is trained to produce bounding boxes for all the latent instances in an image, while another is trained to learn the object-aware weights for features. The instance level feature is finally produced by region of interest (ROI) pooling [17] on the weighted feature maps with the bounding boxes provided from the first pipeline. The overall framework is shown in Fig. 1. The first pipeline is originally designed for object detection [33], while the second pipeline is originally adopted to discover more discriminative visual evidence in images [39]. The backbone network for both is VGG-16 [15]. Considerable modifications have been made on both to fit them into our task, which are detailed in the following sections.

3.2. Instance localization

In order to build instance level feature representation, it is critical to localize each latent instance by a bounding box. In our framework, the feature extraction for an individual instance mainly relies on its bounding box. The bounding box is therefore expected to cover the discovered instance as precisely as possible. In our design, edge information from shallow convolutional (conv) layers is exploited to generate hundreds of object-like proposals by EdgeBoxes. However, it demands high computational cost to extract all the proposal features of one image to match with the query, which is unaffordable for large scale datasets. To this end, a proposal clustering learning method which helps to cluster the proposals related to one instance together, is used in our instance localization pipeline.

For the convenience of following discussion, the feature map from a conv layer is represented as $F \in \mathbb{R}^{C \times W \times H}$, where C, W and H are the channel number, width and height of the feature map respectively. The response map of the feature map is represented as $R \in \mathbb{R}^{W \times H}$. For each layer, the response map is derived from its corresponding feature map by taking average over the channel dimension, which is given by Eq. (1). r_{ij} and f_{cij} in Eq. (1) are elements in R and F, respectively. Edgeboxes procedure produces proposals based on the input edge response map which is represented as $R_{edge} \in \mathbb{R}^{W \times H}$. After further resizing the response maps to the size of original image, the edge response map is generated by taking average again over these response maps. In our design, response maps from *conv1-2*, *conv2-1* and *conv2-2* of VGG-16 are averaged to produce the edge response map (shown in Eqn. 2).

$$r_{ij} = \frac{1}{C}\sum_{c=1}^{C} f_{cij}\ (i = 1, \ldots, W \ and \ j = 1, \ldots, H) \tag{1}$$

$$R_{edge} = \frac{1}{3}(R_{conv1\text{-}2} + R_{conv2\text{-}1} + R_{conv2\text{-}2}) \tag{2}$$

The parameters of the first four conv layers are from ImageNet pre-trained model and their gradients will no longer be updated in the later training procedure. Since the categories of training dataset will not affect the previous four conv layers, they are more likely to be sensitive to unknown categories. Note that *conv1-1* and deeper layers are not chosen since they have either high response to almost all the image regions or have response only to instance regions of known categories.

Considering the high computational cost on hundreds of proposals for each image, a proposal clustering learning (PCL) [33] method is introduced to refine the generated proposals. It consists of a basic MIL stream and two instance classifier refinement streams. The whole pipeline is optimized by minimizing the image classification errors. For each stream, proposal clusters are generated according to the proposal classification scores. The current stream provides proposal classification scores as supervisions for the next stream. After the PCL processing, the object related proposals are refined while the objects belonging to background are filtered. The examples of proposals produced by our instance localization pipeline are showed as Fig. 2. As shown in the figure, the proposed method is able to cover most of the objects in an image of both known and unknown categories. Since it only relies on feature maps of shallow layers, the localization is also very efficient.

3.3. Feature extraction

Although the bounding boxes that are produced by the above instance localization pipeline cover the instances (of both known and unknown categories) well, the feature maps from the first

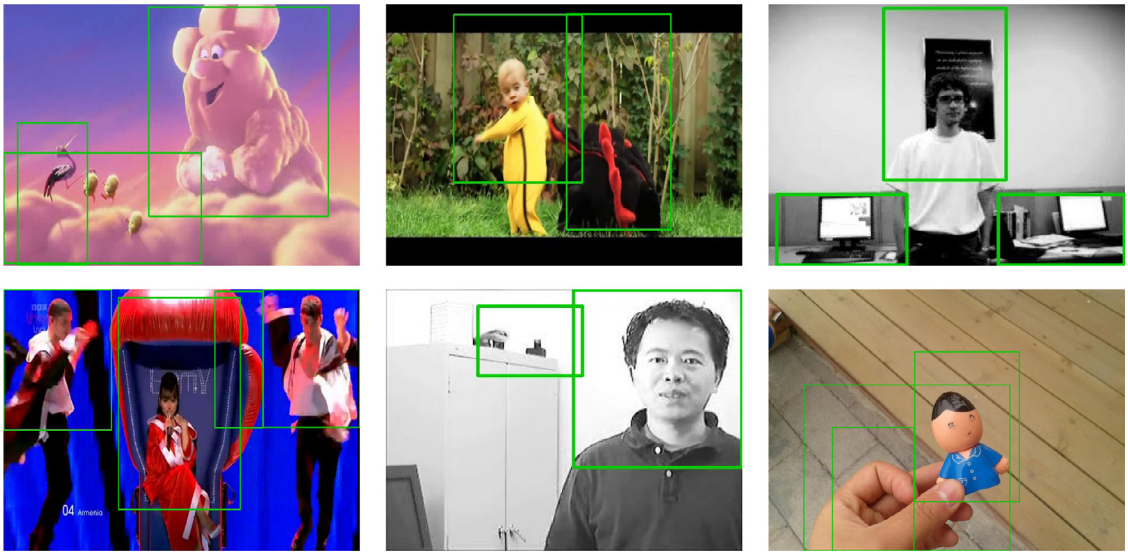


Fig. 2. The examples of proposals obtained by our instance localization pipeline. As shown in the figure, there are many objects that are never seen in the training categories.

pipeline are not suitable for instance feature representation. According to our observation, since the features from the first pipeline are trained to be on the semantic level, the differences between instances of the same category are largely lost. In addition, one could not expect the bounding boxes produced from weakly supervised method are as precise as those from fully supervised one. As the consequence, features extracted directly from regions detected by the first pipeline may be mixed with noises from the background or neighboring objects.

To address this issue, a spatial weighting network named soft proposal network (SPN) [39] is adopted to generate object-aware weights for instance level features separately. SPN is basically a modification over VGG-16, in which the Soft Proposal (SP) is plugged into the last conv layer. The SP module learns an objectness map which reflects the dissimilarity of each object region from their surroundings by a graph propagation algorithm. Therefore, in the objectness map, the object region will be highlighted while the background one will be suppressed. In the forward propagation, Hadamard product is performed to combine the objectness map with the feature map from the next conv layer. Then in the back-propagation procedure, the gradient is apportioned by the parameters of the objectness map. As the results, all the conv layers of SPN are enhanced by the object-aware weights from the objectness map. Most importantly, SPN only requires image level annotations for training. In our design, the feature maps from SPN are ROI pooled based on the bounding boxes produced by the first pipeline. This finally leads to a compact feature representation of equal size for instance from each bounding box. In the Section 4.3, ablation analysis is conducted to show the layer and the combination of layers of SPN from which the features are the best suitable for instance search. The resulting features are l_2-normalized and *Cosine* distance is adopted in the comparison.

4. Experiments

In this section, the performance of instance search based on the proposed feature representation is studied on two instance search datasets. The brief on the evaluation dataset and the experiment setups are introduced in Section 4.1. The ablation analysis about our method is presented in Section 4.2. The feature selection test about our method is presented in Section 4.3. The performance comparisons to several state-of-the-art methods on two evaluation benchmarks are presented in Section 4.4.

4.1. Datasets and experiment setup

The performance of the proposed method is studied on two challenging datasets, namely Instance-160 [29] and INSTRE [43]. Instance-160 is derived from *160* video sequences originally used for visual tracking evaluation. There are *160* individual instance queries. There are *12,045* images in the reference set. INSTRE is built by collecting images of various categories. The *200* instances range from objects such as buildings, common objects, sculptures to logos. There are *28,543* images in total. Following the evaluation protocol in [44], *1250* images in the dataset are treated as queries, and the rest are given as *27,293* reference images. For both datasets, the bounding boxes of the query instances are provided in advance.

Performance is evaluated with mean Average Precision (mAP). Representative feature representations of both conventional and CNN-based are considered in our comparison. BoVW [11] and BoVW with Hamming embedding (BoVW+HE) [45] are based on SIFT. The considered image-level deep features are bags of local convolutional features (BLCF) [27], BLCF with saliency weighting (BLCF-SalGAN) [27], regional of maximum activation of convolutions (R-MAC) [23], cross dimensional weighting scheme (CroW) [25] and features from weighted Class Activation Map (CAM) [26]. Their features are extracted from pre-trained CNNs. Our method is also compared to two fully-supervised methods, both of which produce instance-level features. They are Deepvision [24] and FCIS with deformable convolution and ResNeXt-101 (FCIS+XD) [29]. Deepvision extracts instance-level features from proposals produced by Faster-RCNN [18]. Features for FCIS+XD are extracted from a fully convolutional neural network [21] that is augmented for both instance segmentation and instance search.

The networks in our framework are implemented by PyTorch. All of our experiments are pulled out on an Nvidia Titan X GPU. Our networks are trained with image level annotations. They are pre-trained on ImageNet and fine-tuned on Microsoft COCO 2014 dataset [46].

4.2. Configuration test

In the first experiment, ablation analysis is conducted with five different runs. We mainly study the effectiveness of the instance localization, the contribution of edge information to the localization and the impact of weight assignment on our instance

Table 1
Performance evaluation of different enhancement schemes in our method on Instance-160 and INSTRE, the feature dimension is fixed to *512*.

Dataset	Method	Top-10	Top-20	Top-50	Top-100	All
Instance-160	SPN	0.142	0.223	0.340	0.380	0.422
	PCL	0.167	0.271	0.412	0.460	0.509
	PCL*	0.169	0.276	0.429	0.485	0.539
	PCL+SPN	**0.183**	**0.303**	0.474	0.537	0.596
	PCL*+SPN	0.177	0.297	**0.476**	**0.541**	**0.603**
INSTRE	SPN	–	–	–	–	0.077
	PCL	–	–	–	–	0.243
	PCL*	–	–	–	–	0.328
	PCL+SPN	–	–	–	–	0.256
	PCL*+SPN	–	–	–	–	**0.415**

Table 2
Performance comparison on Instance-160.

Method	Dim.	Top-10	Top-20	Top-50	Top-100	All
BoVW [11]	65,536	0.106	0.165	0.248	0.281	0.314
BoVW+HE [45]	65,536	0.148	0.236	0.355	0.403	0.438
BLCF [27]	336	0.046	0.076	0.126	0.167	0.227
BLCF-SalGAN [27]	336	0.063	0.105	0.175	0.214	0.278
R-MAC [23]	512	0.101	0.164	0.268	0.307	0.358
CroW [25]	512	0.073	0.130	0.239	0.284	0.338
Deepvision [24]	512	0.194	0.328	0.541	**0.666**	**0.731**
FCIS+XD [29]	1536	**0.211**	0.356	0.575	0.659	0.724
Ours	1024	**0.211**	**0.358**	**0.578**	0.660	0.722

Table 3
Performance comparison on 40 queries of Instance-160 where heavy background changes are in present.

Method	Dim.	Top-10	Top-20	Top-50	Top-100	All
Deepvision [24]	512	0.193	0.298	0.464	0.559	0.589
FCIS+XD [29]	1536	**0.262**	**0.430**	**0.647**	**0.698**	**0.737**
Ours	1024	0.239	0.363	0.517	0.569	0.610

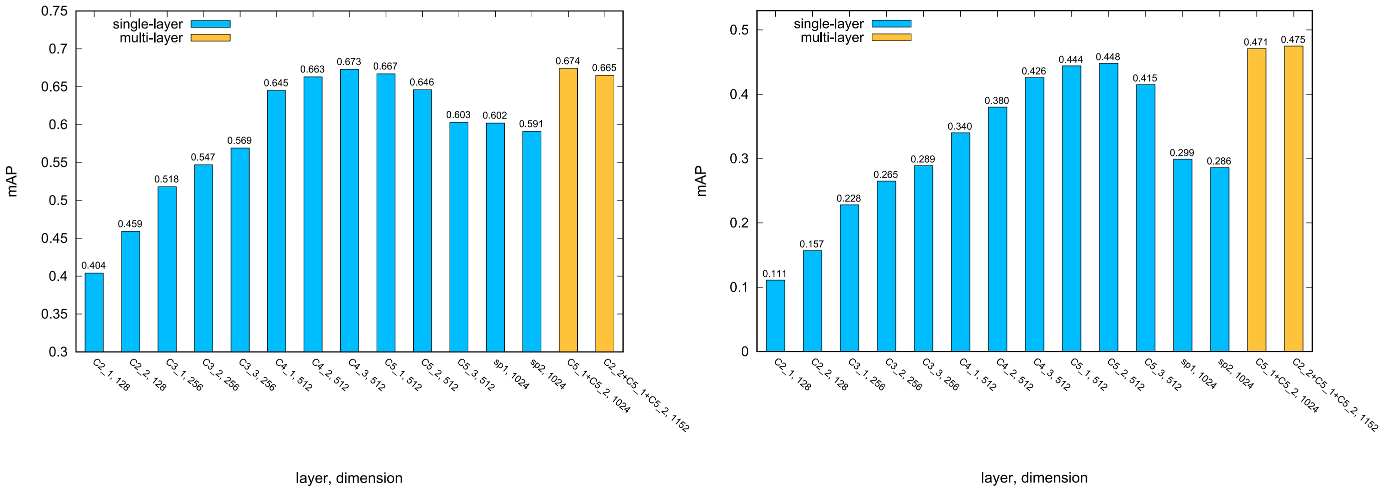
search task. The features produced by the spatial weighting network named SPN are treated as the comparison baseline. These SPN features are produced by ROI pooling with the proposals produced by SPN itself. Features extracted from proposals produced by original PCL and the one from enhanced PCL (given as PCL*) are also studied. In addition, the features from SPN feature maps while being ROI pooled with proposals produced by PCL are studied, which is given as "PCL+SPN". Our features are ROI pooled from the weighted feature maps from SPN with the proposals produced by PCL*, which is given as "PCL*+SPN". Features for above five configurations are extracted from the *conv5-3* layer of VGG-16. The PCA whitening is not adopted in any of the above configurations.

The evaluation is in line with the protocol of each benchmark. Following [29], the performance on Instance-160 is measured by mAP at top-*k*, where *k* varies from *10* to *100*. As shown on Table 1, PCL+SPN outperforms SPN by *0.174* and *0.179* on Instance-160 and INSTRE respectively. This basically indicates the instance localization plays an important role for feature representation. Furthermore, considerable improvement is observed from PCL* over PCL. Similar trend is observed when comparing PCL*+SPN to PCL+SPN. This confirms our choice of shallow conv layers as the input to EdgeBoxes for bounding box estimation. It also shows the ability of discovering unknown category is critical to boost the search performance. The superiorities of PCL+SPN over PCL and PCL*+SPN over PCL* demonstrate the weight assignment on the feature maps is helpful. Due to the superior performance, features from PCL*+SPN are adopted as the standard configuration for our method in the rest of our experiments.

4.3. Feature selection

Theoretically speaking, the feature maps from each conv layer could be used to derive the instance level features under our framework. However, as witnessed by many studies, the performance varies considerably for the features derived from different layers. In this ablation analysis, we study the performance of features derived from different conv layers. Such that we try to seek the best representation for the detected instances. Namely, features derived from *conv2-1* to *conv5-3* layers of VGG-16 and two layers of SP module are tested. The experiment results on Instance-160 and INSTRE are shown in Fig. 3.

As seen from the figure, features from intermediate layers show relatively good performance, which shares similar observation as [29]. In order to enhance the performance, the combinations of features from different layers are also tried. The single-layer features we select to concatenate is based on the consideration of both their distinctiveness and dimensionalities.



(a) (b)

Fig. 3. Performances of deep features extracted from different conv layers and different feature concatenation ways, where the (a) shows the performance on Instance-160, the (b) shows the performance on INSTRE. The previous 13 bars represent single-layer features and the later 2 bars represent multi-layer features. C2_1, 128 donates the first conv layer in second conv group of VGG-16 with the feature dimension is *128*. sp1 and sp2 donate the two layers in SP module.

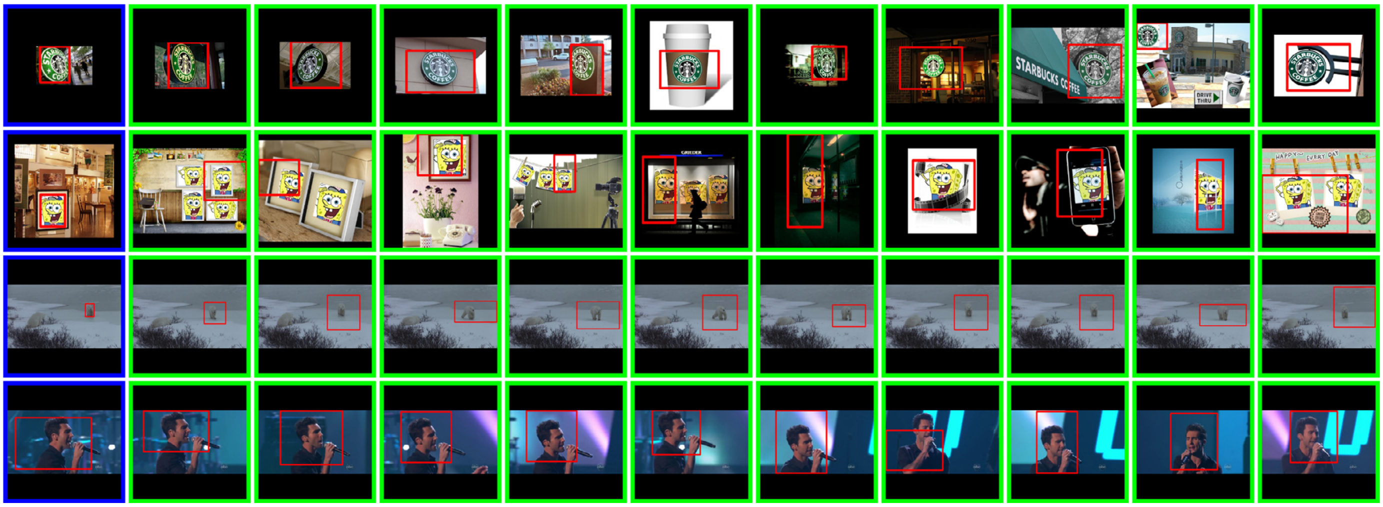


Fig. 4. Top-10 results of four sample queries from INSTRE and Instance-160. The first two rows are from INSTRE, the third and fourth rows are from Instance-160. Images of the first column are the queries. The rest images in one row are the returned results (ranked from left to right). The detected bounding boxes are also shown.

Table 4
Performance comparison on INSTRE.

Method	Off-the-shelf	Dim.	All
Deepvision [24]	No.	512	0.197
FCIS+XD [29]	No.	1536	0.067
CroW [25]	Yes	512	0.416
CAM [26]	Yes	512	0.325
R-MAC [23]	Yes	512	0.523
BLCF [27]	Yes	336	0.636
BLCF-SalGAN [27]	Yes	336	**0.698**
Ours	No.	1024	0.575

According to the results on the two benchmarks, we find the best combination comes from *conv5-1* and *conv5-2*. Moreover, according to our off-the-shelf test, concatenating features from more layers is not helpful or the improvement is minor. Therefore our feature is produced from the combination of features from *conv5-1* and *conv5-2* in the rest of experiments.

4.4. Comparison to state-of-the-arts

In this section, the performance of our method is studied in comparison to several representative works such as BoVW [11], BoVW+HE [45], R-MAC [23], BLCF [27], CAM [26], CroW [25], Deepvision [24] and FCIS+XD [29]. The evaluation is conducted on Instance-160 and INSTRE. The evaluation is in line with the protocol of each benchmark. Following several other works [24,27], features extracted from our method are l_2-normalized, followed by PCA whitening and a second round of l_2-normalization. For those methods that cannot return bounding boxes, their mAPs are measured on image level.

The performance on Instance-160 is shown on Table 2. As seen from the table, Deepvision, FCIS+XD and our method show relatively superior performance. Among them, FCIS+XD demonstrates similar performance trend as ours. In contrast, conventional handcrafted features (e.g., BoVW and BoVW+HE) and global deep features (e.g., BLCF, R-MAC and Crow) show considerably lower performance. Even though FCIS+XD and Deepvision are competitive with ours, the training conditions for them are demanding. While our method only requires image-level class labels for the training set.

Another disadvantage of Deepvision is that it still relies on global deep feature. It actually undertakes two-stage search. On the first stage, the deep global features[2] are adopted to search over the entire image set. On the second round, the features from query instance are compared to the instance level features from top-ranked candidates of the first stage search. As a result, instances with considerable background variations may be missed on the first stage search.

Another experiment on a subset of Instance-160 confirms our observation. In this subset, instances whose backgrounds are under severe variations are selected. On this subset, it is easy to see the advantage of instance level representation over those features derived from the whole image. As shown on Table 3, the performance of Deepvison drops a lot compared to previous result on the entire Instance-160 dataset. This basically indicates that instance-level features show better distinctiveness over global features. Although FCIS+XD achieves the best performance on this dataset, the training conditions are too demanding to be undertaken for large-scale tasks. Compared to existing solutions, our method achieves a good trade-off between the performance and the training cost.

The performance on INSTRE is shown on Table 4. Our method is only next to BLCF and BLCF-SalGAN. Notice that these two methods produce features on image level and are unable to localize instances from the retrieved images. Both Deepvision and FCIS+XD show considerable performance degradation on INSTRE, in contrast to their high performance on Instance-160. The performance degradation is mainly due to their insensitivity to unknown categories. On the contrary, BLCF, BLCF-SalGAN and R-MAC show competitive performance on INSTRE while poor performance is observed on Instance-160. This basically indicates that their feature representations are not robust to the scenarios where instances are embedded in the complex background. Overall, our method shows stable performance across two challenging evaluation benchmarks. On the one hand, it shows that our method is capable of identifying unknown categories. On the other hand, it also indicates our feature representation remains distinctive despite of the severe variations in instance appearance or the interference from complex backgrounds.

Fig. 4 further shows search samples from our method. Interestingly, we find that our method has the ability to identify unknown categories. For instance, STARBUCK logo and the cartoon character "Spongebob", which are not in our training categories have been successfully identified. In addition, the feature description is suf-

[2] Features are extracted from whole image for both query and reference images.

ficiently robust that gets the query instance well-matched to the candidate instances even under severe geometric variations. Our method also shows encouraging performance on non-rigid objects as demonstrated on the last two rows.

5. Conclusion

In this paper, a feature representation scheme that is designed for instance search has been presented. Different from many existing instance search schemes, the feature is built genuinely on instance level and object-aware weights are assigned on the regions of the objects residing in. This leads to the distinctive feature representation as well as precise instance localization. Moreover, this feature is trained based on a weakly supervised object detection network, which only requires image level annotations and is less reliant on known category labels. It therefore turns out to be sensitive to latent instances of known and unknown categories in an image. Stable and superior performance has been observed on two challenging evaluation benchmarks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China under grants 61572408 and grants of Xiamen University 0630-ZK1083.

References

[1] W.-L. Zhao, H. Jegou, G. Guillaume, Sim-min-hash: an efficient matching technique for linking large image collections, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2013, pp. 577–580.

[2] A. Gordo, J. Almazan, J. Revaud, D. Larlus, End-to-end learning of deep visual representations for image retrieval, Int. J. Comput. Vis. 124 (2) (2017) 237–254.

[3] A. Babenko, V. Lempitsky, Aggregating local deep features for image retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1269–1277.

[4] M. Tzelepi, A. Tefas, Deep convolutional learning for content based image retrieval, Neurocomputing 275 (2018) 2467–2478.

[5] C. Bai, L. Huang, X. Pan, J. Zheng, S. Chen, Optimization of deep convolutional neural network for large scale image retrieval, Neurocomputing 303 (2018) 60–67.

[6] Y. Li, X. Kong, H. Fu, Q. Tian, Aggregating hierarchical binary activations for image retrieval, Neurocomputing 314 (2018) 65–77.

[7] Y. Li, Z. Miao, J. Wang, Y. Zhang, Nonlinear embedding neural codes for visual instance retrieval, Neurocomputing 275 (2018) 1275–1281.

[8] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[9] H. Bay, T. Tuytelaars, L. Van Gool, SURF: Speeded up robust features, in: Proceedings of the European Conference on Computer Vision, Springer, 2006, pp. 404–417.

[10] S.S.C.-Z. Zhu, H. JĘgou, NII team: query-adaptive asymmetrical dissimilarities for instance search, in: Proceedings of the TRECVID 2013 workshop, Gaithersburg, USA, 2013, pp. 1705–1712.

[11] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2003, pp. 1470–1477.

[12] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3304–3311.

[13] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 143–156.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 (2014).

[16] S. Yu, S. Jia, C. Xu, Convolutional neural networks for hyperspectral image classification, Neurocomputing 219 (2017) 88–98.

[17] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[18] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[19] T. Zhang, L.-Y. Hao, G. Guo, A feature enriching object detection framework with weak segmentation loss, Neurocomputing 335 (2019) 72–80.

[20] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[21] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei, Fully convolutional instance-aware semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2359–2367.

[22] A.S. Razavian, J. Sullivan, S. Carlsson, A. Maki, Visual instance retrieval with deep convolutional networks, ITE Trans. Media Technol. Appl. 4 (3) (2016) 251–258.

[23] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of CNN activations, arXiv:1511.05879 (2015).

[24] A. Salvador, X. Giró-i Nieto, F. Marqués, S. Satoh, Faster R-CNN features for instance search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 9–16.

[25] Y. Kalantidis, C. Mellina, S. Osindero, Cross-dimensional weighting for aggregated deep convolutional features, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 685–701.

[26] A. Jimenez, J.M. Alvarez, X. Giro-i Nieto, Class-weighted convolutional features for visual instance search, arXiv:1707.02581 (2017).

[27] E. Mohedano, K. McGuinness, X. Giró-i Nieto, N.E. O'Connor, Saliency weighted convolutional features for instance search, in: Proceedings of the International Conference on Content-Based Multimedia Indexing (CBMI), IEEE, 2018, pp. 1–6.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[29] Y. Zhan, W.-L. Zhao, Instance search via instance level segmentation and feature representation, arXiv:1806.03576 (2018).

[30] W. Ren, K. Huang, D. Tao, T. Tan, Weakly supervised large scale object localization with multiple instance learning and bag splitting, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2015) 405–416.

[31] R.G. Cinbis, J. Verbeek, C. Schmid, Weakly supervised object localization with multi-fold multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (1) (2016) 189–203.

[32] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2846–2854.

[33] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, A.L. Yuille, PCL: proposal cluster learning for weakly supervised object detection, IEEE Trans. Pattern Anal. Mach. Intell. (2018).

[34] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[35] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot multibox detector, in: Proceedings of the European conference on computer vision, Springer, 2016, pp. 21–37.

[37] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, in: Proceedings of the Advances in Neural Information Processing Systems, 1998, pp. 570–576.

[38] X. Zhang, Y. Wei, J. Feng, Y. Yang, T.S. Huang, Adversarial complementary learning for weakly supervised object localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1325–1334.

[39] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, J. Jiao, Soft proposal networks for weakly supervised object localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1841–1850.

[40] M. Shi, V. Ferrari, Weakly supervised object localization using size estimates, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 105–121.

[41] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.

[42] C.L. Zitnick, P. Dollár, Edge boxes: locating object proposals from edges, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 391–405.

[43] S. Wang, S. Jiang, INSTRE: a new benchmark for instance-level object retrieval and recognition, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 11 (3) (2015) 37.

[44] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, O. Chum, Efficient diffusion on region manifolds: recovering small objects with compact CNN representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2077–2086.

[45] H. Jegou, M. Douze, C. Schmid, Hamming embedding and weak geometric consistency for large scale image search, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 304–317.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 740–755.

Jie Lin received her Bachelor degree of Computer Science from Fuzhou University, China in 2017. She is currently a graduate student at Department of Computer Science, Xiamen University. Her research interest is content-based image retrieval and instance search.

Yu Zhan received his Master degree of Computer Science from Xiamen University, China in 2019. He is currently an Algorithm Engineer in Aibee Group. His research interest is visual object detection, image retrieval and machine learning.

Wan-Lei Zhao received his Ph.D degree from City University of Hong Kong in 2010. He received M.Eng. and B.Eng. degrees in Department of Computer Science and Engineering from Yunnan University in 2006 and 2002 respectively. He currently works with Xiamen University as an associate professor, China. Before joining Xiamen University, he was a Postdoctoral Scholar in INRIA, France. His research interests include multimedia information retrieval and video processing.