Searching for Objects in 20 Questions

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a new strategy for simultaneous object detection and segmentation. Instead of evaluating classifiers for all possible locations of objects in the image, we develop a divide-and-conquer approach by sequentially posing questions about the query and related context, like playing a "Twenty Questions" game, to decide where to search for the object. We formulate the problem as a Markov Decision Process (MDP) and learn a policy using imitation learning that can dynamically select questions based on the query, the scene and current observed responses given by object detectors and classifiers. The algorithm reduces over 30% of the object proposals evaluation while maintaining high average precision comparable to exhaustive search. Experiments show promising results compared with baselines of exhaustive search, searching for objects in random sequences and random locations.

1 Introduction

Object detection and segmentation in complex scenes is a central and challenging problem in computer vision and robotics. This problem is usually tackled by running multiple object detectors exhaustively on densely sampled sliding windows [9] or category-independent object proposals [7, 24, 2]. Such methods are time-consuming since they need to evaluate a large number of object hypotheses, and easily introduce false positives if purely considering local appearance cue.

Instead of checking all hypotheses indiscriminately and exhaustively, humans only look for a set of related objects in a given context [4, 15]. Context information is an effective cue for human to detect low-resolution or small objects in cluttered scenes [22]. Many contextual models have been proposed to capture relationships between objects at the semantic level to reduce ambiguities from unreliable detection results [13, 10, 17]. However, such methods still need to evaluate high order co-occurrence statistics and spatial relations with *all* other object classes appearing in the scene altogether, some of which may not be informative and even introduce distractions. By contrast, human vision is an active process that sequentially samples the optic array in an intelligent, task-specific way [21] **[Stephen: Cite Nature]**. Research in neuro-science have revealed that when humans search for a target, those objects that are associative to the query will reinforce attention with the query and weaken recognition of unrelated distractions [19]. For instance, in Figure **??**, knowing the top of the scene is sky is not very helpful to distinguish whether there is a car or a boat since both can be under the sky; while observing road instead of water in the lower part gives a strong indication of the existence of cars. And then cars are more likely to be found beside the buildings after observing the road. This motivates us to raise the question: can object detection algorithms decide where to look for the object more accurately and efficiently by exploring a few related context dynamically like human?

In this paper, we formulate this process as a Markov Decision Process (MDP), and use imitation learning to learn a context-driven policy that sequentially and dynamically selects the most informative context class to explore and refine the search area for the target. We show our framework in Figure 3. Specifically, like playing a Twenty Question game, at each step the policy makes a decision about which detector to run given the query and responses from previous contextual classifiers. After taking the action to run context classifiers/detectors and observing responses, it then refines the search area for the query object using spatial-aware contextual models. We also incorporate early rejection as an action to avoid further running detection if the policy determines there is a low chance of having the query object in the scene, and this decision can be made even before running any object detectors so it can reduce a large number of unnecessary detection. Finally, we run the query object detector in the output search area if the policy thinks enough contextual information has been gathered and decides to stop. Object detection experiments in images of complex scenes like Pascal VOC dataset show that our algorithm can produce a search space that's closer to the target object by checking its context, thus significantly reduces the number of object proposals and detector evaluation while maintaining comparable mean average precision (mAP) to exhaustive search. To the best of our knowledge, this is one of the first works to model the challenging task of simultaneous object detection and segmentation in complex real world scenes using a imitation learning policy fully driven by semantic context.

2 Related Work

Sequential Testing. The "20 question" approach to pattern recognition dates back to Blanchard and Geman [5], motivated by the large number of possible explanations in scene interpretation. They formally studied coarse-to-fine search in the theoretical framework of sequential hypothesis testing, and proposed optimal strategies considering both the cost and effectiveness of each test. Although they did not consider contextual information, their work provides a theoretical foundation for the design of sequential algorithms.

There are several works [11] on classifying objects by running classifiers sequentially in an active order. [6] proposed an information gain based approach to iteratively pose questions for users and incorporate human responses and computer vision detector results for fine-grained classification. [16] formulated object classification as a Markov decision process, where actions are the detector to deploy next. The model maintains a belief of object classes and keeps updating it based on new observations. They used reinforcement learning to train the detector selection policy, which becomes expensive when the number of classes and data size is large due to exploration. However, these approaches only focus on classifying objects. They have not addressed the challenging problem of simultaneous segmentation and localization of objects in a multi-class scene as we do in this chapter, and did not exploint inter-object context.

[1] applied a sequential decision making framework to window selection. The next window is selected based on votes of previously evaluated windows. However, the voting process needs to look up nearest neighbors in hundreds of thousands of exemplar window pairs in the training set because their context is at the exemplar/instance level, which is highly inefficient. In contrast, our context modeling is semantically aware so we do not compute nearest neighbors over hundreds of thousands of windows in a high dimensional descriptor space to retrieve the voters, we only need votes from a few regions within the search space of context class instead of sampling hundreds of windows in [1]. Our context model achieves good accuracy while greatly reducing computational complexity.

Object Detection. A common approach to object detection is based on applying gradient based features over densely sampled sliding windows [9].Such methods achieve good results on classes like human and vehicles, but they are very inefficient since they evaluate thousands of windows in an image, and false positve detections arise. To reduce the number of windows evaluated. [18] proposed a subwindow search based on a branch-and-bound scheme and only evaluates the high scoring windows. Recently, category independent object proposals [7, 24, 2] have been proposed to generate a small number of high quality regions or windows that are likely to be objects. These approaches dramatically reduce the number of candidates and reduce false positive detections. Using these object proposals [12, 14] train and apply deep neural network models on large datasets to learn the feature extractor and classifiers, and achieve state-of-the-art performance on the Pascal VOC detection challenge.

Object Recognition using Context. Context has been shown to improve object recognition and detection. Model-based approaches learn the appearance of semantic categories and relations among them using a parametric model. In [13, 10, 20, 23, 17], CRF models are used to combine unary potentials based on visual features extracted from superpixels with neighborhood constraints and low level context. Inter-object context in the scene has also been shown to improve recognition [10, 8]. Most of these context models are used as post-detection smoothing after all classifers are run as unary potentials, and then they are jointly incorporated in inference regardless of their importance to different kinds of objects and scenes. Our framework, in contrast, evaluates the informativeness of context in an active loop before classifications of all objects are made, and goes beyond simple co-occurence statistics.

3 Approach

Our framework is shown in Figure 3. Our goal is to learn a policy $\pi(s)$ that can dynamically determine a sequence of contextual classes to detect to help narrow down the search area of the query or make an early rejection, under certain budget constraint. We model it in a reinforcement learning framework by introducing the Markov Decision Process (MDP), which defines a single *episode* of selecting actions for the input image X and target query class c_q.

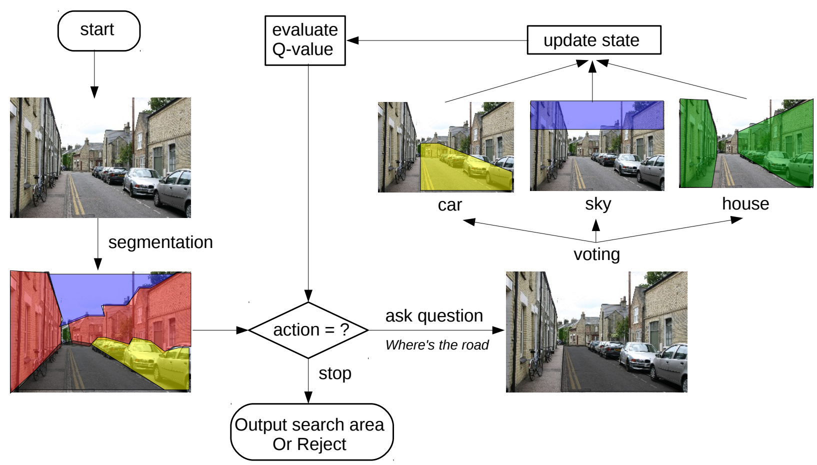

Figure 1: Framework of our context driven object searching. We first generate region hypotheses using object proposal algorithms, then the policy evaluates the current state and iteratively selects the action maximizing the Q-value function. Afterwards, the possible search locations are updated and the posterior probabilities of each category are evaluated for the next state.

3.1 Context Driven Object Search as Markov Decision Process (MDP)

Specifically, given an image X and a query object class c_q in classes $1, .., C$, we define the MDP as follows.

Definition 1. *The* ***detection action selection MDP*** *consists of the tuple* $(\mathcal{S}, \mathcal{A}, T(.), R(.), \gamma)$*:*

- ***State*** $s(t) = (X, R^t) \in \mathcal{S}$ *that includes the image* X *and the observations* $R^t = \{r_1, r_2, ..., r_t\}$ *over time till* t*.*
- *The set of* ***actions*** $\mathcal{A} = \{a_1, ..., a_C, Stop, Reject\}$*, where* a_i *is to detect class* c_i*, Reject to reject query class in the image, and Stop to output the search area and run query detector.*
- ***State transition*** *function* $T(s'|s, a, X)$

- *The **reward** function $R(s, a, s') \to \mathbb{R}$.*
- *The **discount** constant γ defines a tradeoff between taking the action by greedily maximizing the immediate reward or the considering the long term expected reward.*

During test time, our search policy $\pi(s) : \mathcal{S} \to \mathcal{A}$ iteratively selects an action a^t in the action space $\mathcal{A}$. Then the policy obtains response r_t at time step t given by the detection or classification results of action a_t.

Formally, since we would like to select the actions dynamically, we want to learn a value function taking action a in state s under the policy π, denoted $Q^\pi(s, a) : S \times A \to \mathbb{R}$, where S is the space of all possible states, to assign a value to a potential action $a \in A$ given the current state. We can then define policy π to take the action that maximizes the expected value:

$$\pi(s) = \arg \max_{a_i \in A \setminus R} Q(s, a_i) \tag{1}$$

3.2 Reward Function

We define the *immediate reward* R as the immediate gain in intersection/union of the search space after conducting action a_i at time step t under state s as:

$$R(s^t, a_i) = \frac{X_i^{t+1} \cap X_q}{X_i^{t+1} \cup X_q} - \frac{X_i^t \cap X_q}{X_i^t \cup X_q} \tag{2}$$

where X_i^{t+1} is the updated search area after executing action a_i in state s_t, determined by the context models described in Section 3.4. X_q is the groundtruth mask of the query object instances in the image.

3.3 Learning Context Driven Search Policy

We propose a policy to use maximum expected reward to select a_t. To learn such a policy, we adopt a standard Q-learning algorithm [3], where the action-value function is estimated by using the Bellman equation iteratively:

$$Q^*(s, a) = \mathbb{E}_{s' \sim \xi}\big[R + \gamma \max_{a'} Q^*(s', a') | s, a\big] \tag{3}$$

where a' and $s' \in \xi$ is the possible action-state pairs at the next time steps $t + 1$.

Since the space of possible states S is intractable, we use a linear function of the features of the states to approximate the Q-values:

$$Q^\pi(s, a) = \theta_\pi^T \phi(s, a) \tag{4}$$

where $\phi(s, a) = \phi((X, R^t), a) = \phi(X^t, a)$ is the feature of the undetermined area X^t at time t after observing detector responses of $a_1, ..., a_t$.

The parameters θ_π are learned by *policy iteration*. We collect (s, a, r, s') samples by running episodes starting from a random or empty state, then we search and prune in the tree of states and collect states samples and corresponding features. An SVM regression is trained for each class to predict the Q-values given current states.

3.4 Context Modeling

Since our task is not only to detect the object but also refine the search space of the query in the image as accurately as possible, conventional modeling of context as simple co-occurrence statistics is inadequate. Instead we present a data-driven location aware approach to represent the spatial correlation between the objects and the scene.

Here we formulate the context $p(c_t|c, X)$ as a posterior of the probabilistic vote map $p(c|c_t, X_s)$ defined on each pixel $(x_i, x_j) \in X$ over the image, and the responses of class c_t after action a_t:

$$p(c_t|c, X) = \sum_{s \in X^t} p(c|c_t, X_s)p(c_t|X_s) \tag{5}$$

Given a refined search space $X^t \in X$ of a context class c_t at time t, we formalize $p(c|c_t, X)$ as a weighted vote from the cooccurring region pairs of class c_t and c in training scenes. Let $(s^i_{c_t}, s^i_c)$ be the i-th pair of co-occurring regions of class c_t and c, and $b^i_{c_t}$ and b^i_c be the corresponding bounding boxes. We can now define the probabilistic vote map $p(c|c_t, X)$ as:

$$p(c|c_t, X_s)_{s \in X^t} = \frac{1}{Z_c} \sum_i W(s^i_{c_t}, s; \theta^W).T(b^i_{c_t}, b^i_c) \tag{6}$$

where $s \in X^t$ is a region within the search space of the context class c_t. Z_c is the normalization function. $W(.)$ is a kernel measuring similarity of region s with a training region s_i. $T(b^i_{c_t}, b^i_c)$ models the transformation from $b^i_{c_t}$ to b^i_c, including translation and scaling. Figure 3 shows a few examples of the vote maps. We can see that with the exemplar based and semantically aware voting, the resulted vote maps give more accurate search area of the query objects.

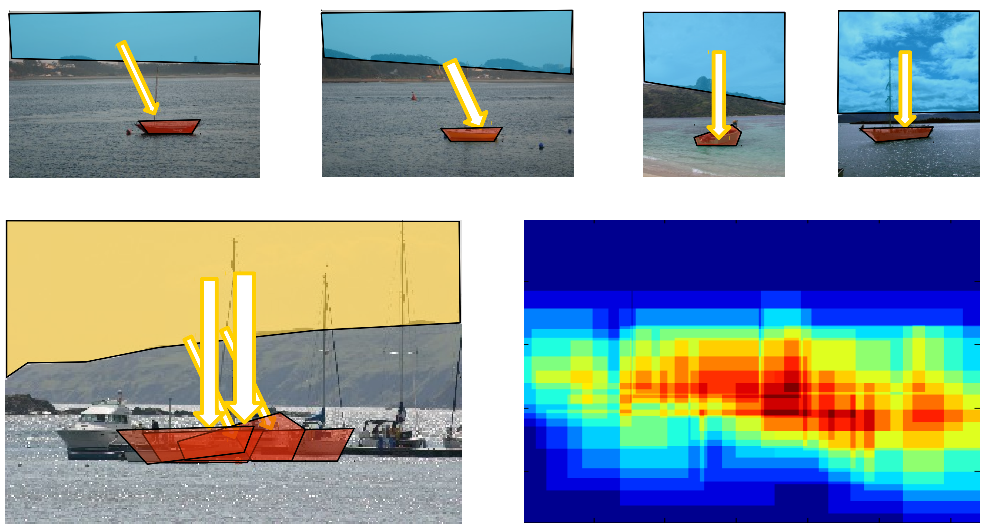

Figure 2: Examples of our weighted vote map for the context from sky to boat. The first rows are the training sample pairs of sky and boat and the second row is the test image and the resulted weighted voting map. The widths of the arrows denote the weighted similarity $W(s^i_{c_t}, s; \theta^W)$ between the test segment of sky (highlighted in yellow) and a training instance of sky segment (in light blue)

The final context probabilistic vote map is given by

$$p(c_t|c, X) = \sum_{s \in X^t} p(c_t|X_s) \sum_i W(s^i_{c_t}, s; \theta^W).T(b^i_{c_t}, b^i_c) \tag{7}$$

where $p(c_t|X_s)$ is the probabilities of s as class c_t after taking the action a_t to run classification at time t.

References

[1] B. Alexe, N. Heess, Y. W. Teh, and V. Ferrari. Searching for objects driven by context. *NIPS*, 2012.

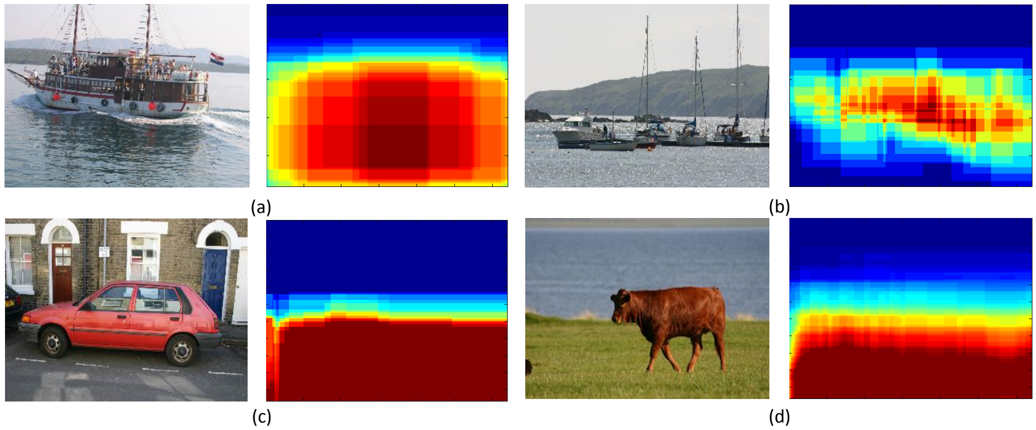

Figure 3: Examples of our context vote maps. Each pair of images corresponds to the original image and the vote-based probability map of object location from observed context. From (a) - (d) are the vote maps from water to boat, sky to boat, road to car and grass to cow, respectively. Best viewed in color.

[2] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. CVPR, 2014.

[3] A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[4] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

[5] G. Blanchard and D. Geman. Hierarchical testing designs for pattern recognition. *Annals of Statistics*, pages 1155–1202, 2005.

[6] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *Computer Vision–ECCV 2010*, pages 438–451. Springer, 2010.

[7] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1312–1328, 2012.

[8] X. Chen, A. Jain, A. Gupta, and L. S. Davis. Piecing together the segmentation jigsaw using context. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2001–2008. IEEE, 2011.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[10] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010.

[11] T. Gao and D. Koller. Active classification based on value of classifier. *NIPS*, 2011.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[13] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1–8. IEEE, 2009.

[14] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014.

[15] H. S. Hock, G. P. Gordon, and R. Whitehurst. Contextual relations: the influence of familiarity, physical plausibility, and belongingness. *Perception & Psychophysics*, 16(1):4–8, 1974.

[16] S. Karayev, T. Baumgartner, M. Fritz, and T. Darrell. Timely object recognition. *NIPS*, 2012.

[17] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In *Computer Vision–ECCV 2010*, pages 239–253. Springer, 2010.

[18] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2129–2142, 2009.

[19] E. Moores, L. Laiti, and L. Chelazzi. Associative knowledge controls deployment of visual selective attention. *Nature neuroscience*, 6(2):182–189, 2003.

[20] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild.

[21] J. Najemnik and W. S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005.

[22] D. Parikh, C. L. Zitnick, and T. Chen. Exploring tiny images: the roles of appearance and contextual information for machine and human object recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1978–1991, 2012.

[23] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Computer Vision–ECCV 2006*, pages 1–15. Springer, 2006.

[24] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1879–1886. IEEE, 2011.