Active Image Segmentation Propagation

Suyog Dutt Jain, Kristen Grauman
University of Texas at Austin

Abstract

We propose a semi-automatic method to obtain foreground object masks for a large set of related images. We develop a stagewise active approach to propagation: in each stage, we actively determine the images that appear most valuable for human annotation, then revise the foreground estimates in all unlabeled images accordingly. In order to identify images that, once annotated, will propagate well to other examples, we introduce an active selection procedure that operates on the joint segmentation graph over all images. It prioritizes human intervention for those images that are uncertain and influential in the graph, while also mutually diverse. We apply our method to obtain foreground masks for over 1 million images. Our method yields state-of-the-art accuracy on the ImageNet and MIT Object Discovery datasets, and it focuses human attention more effectively than existing propagation strategies.[1]

1 Introduction

Large-scale labeled image datasets have had a transformative impact on computer vision in recent years, most notably for image classification. However, image annotation remains a costly undertaking in terms of both time and money. In particular, gathering high quality *spatial annotations*—pixel-level foreground masks—is extremely challenging and expensive. The difficulty in generating foreground spatial annotations for image collections is problematic given their high potential utility e.g. for training region based object detectors, image retrieval, data driven image synthesis etc.

An attractive alternative with low manual effort is to take a pool of images known to contain the same object category (weak supervision), and exploit the repeated patterns to jointly segment out the foreground per image. Such collections can be easily downloaded from web, however using this weak supervision alone still results in largely imperfect segmentations.

In this work, we propose an intermediate solution. Rather than rely solely on human-provided segmentations (accurate but too expensive) or automatic segmentations (inexpensive but too inaccurate), we develop a *semi-automatic segmentation propagation* approach. The idea is to actively request human annotations for select images that, once labeled with their foreground, are most expected to help co-segment the remaining unlabeled images. The propagation engine proceeds in stages, each time (1) using the recently annotated images to revise foreground estimates in all unlabeled images, and (2) using those results to determine the next best batch of images to present to human annotators. In this way, we neither restrict ourselves to the saturation point of the fully automatic methods, nor do we get large volumes of data labeled by humans (see Figure 1).

To achieve this goal, we develop an active selection approach tailored to foreground propagation. It operates on a graph constructed over all images in the collection. Our active selection process favors choosing images that are *uncertain*—poorly explained by any images labeled so far, as well as *influential*—similar to many unlabeled images, making their foreground mask transferrable—and mutually *diverse*—so as to avoid redundant human effort. A critical part of our method design is its stagewise propagation, which permits both human-annotated *and* automatically annotated images to influence the system's view of what most needs human attention next.

Our framework differs in important ways from existing work on both active learning and segmentation propagation. Active learning methods for recognition aim to train a model that will make accurate category label predictions on unseen test images (e.g., [13, 15, 16]). In contrast, our goal is to get all available images spatially annotated by semi-automatic propagation (i.e., ours is a transductive setting). There is very limited prior work on segmentation propagation, and existing methods are either passive [5] or only select annotations to initialize the algorithm [11]. A key insight of our technical approach is to repeatedly analyze the current segmentations to actively decide on subsequent annotations. Applying our method to more than 1 million images, we show that intelligently focusing human effort leads to significantly better foreground extraction.

[1]This abstract summarizes our CVPR 2016 paper [6].

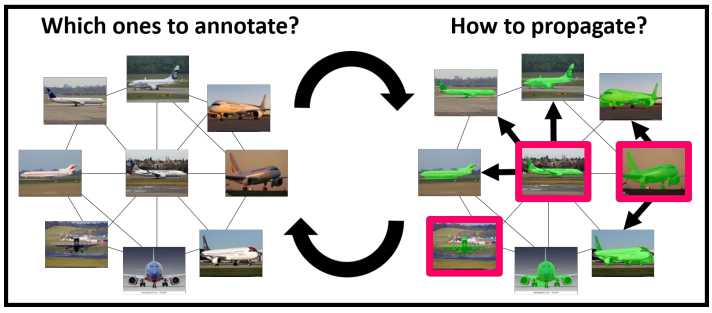

Figure 1: Our active image segmentation propagation method alternates between: (1) Actively choosing images which once annotated by humans will likely be most useful in propagating segmentations to other images and (2) Given human annotations on actively chosen images (marked in pink), propagating them (dark arrows) to generate segmentations for other unlabeled images.

2 Approach

Given a collection of weakly supervised images (denoted by $\mathcal{I}$), all of which contain instances of the same object category, our goal is to jointly segment these images, yielding a foreground object mask for each one. Our proposed approach iterates between the two main components: 1) A joint segmentation procedure to simultaneously solve for all foreground masks, given foreground annotations on only a subset of the images (see Sec. 2.1). 2) An active procedure for identifying the set of images that should be annotated next by human annotators (see Sec. 2.2).

2.1 Semi-automatic joint foreground segmentation

We define a Markov Random Field (MRF) joint segmentation graph $\mathcal{G} = (\mathcal{R}, \mathcal{E})$ over *region proposals* extracted from all images in the collection. Here, $\mathcal{R} = \{R_{ij}\}$ denotes the set of all region proposals in all N images, where R_{ij} denotes the j-th region for image $I_i \in \mathcal{I}$. Each region $R_{ij} \in \mathcal{R}$ forms a node and the edges $\mathcal{E}$ connect pairs of similar regions. The goal of our joint segmentation procedure is to identify the subset of region proposals that are good, and fuse them to obtain the final segmentation. Let $\mathcal{S} \subseteq \mathcal{I}$ denote the current subset of images labeled with foreground masks by human annotators. (We explain in Sec. 2.2 how the composition of this set is iteratively and actively defined.)

The unary terms in our proposed joint segmentation graph prefers to label as foreground those regions that are (1) *higly salient* and/or (2) form a good *match* with some human labeled foreground masks ($\mathcal{S}$). The pairwise terms encourages similar-looking regions to take the same label and enforces consistency in our joint selection of good region proposals.

The minimum energy solution to this MRF yields a set of good region proposals for each image in the collection. Note that we do not constrain only one proposal to be selected per image. We purposely allow selecting *multiple* good regions per image, for two reasons. First, an image can naturally have multiple good region proposals (e.g. covering different object parts). Second, it allows us to efficiently and exactly minimize our energy function using graph-cuts [2]. Finally, our fusion step combines these multiple partial good proposals in each image to a single accurate segmentation.

2.2 Active selection for propagation

We now describe our stagewise algorithm to actively select images for annotation. The active selection procedure takes as input the image collection $\mathcal{I}$, an annotation budget k specifying the number of images to get labeled per stage, and the number of total annotation stages T. In each stage t, we solicit

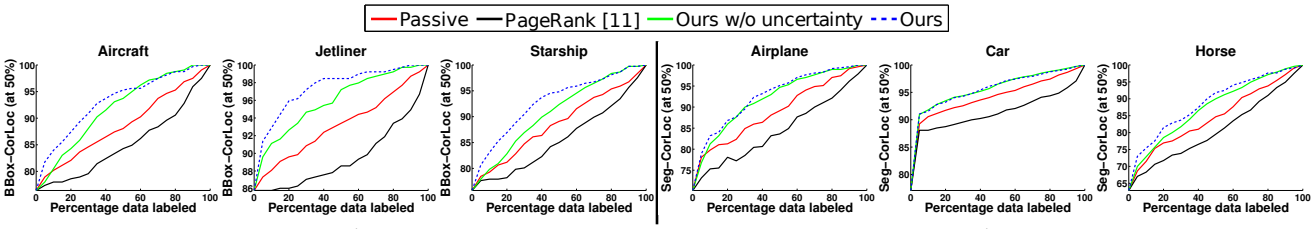

Figure 2: Active propagation for varying amounts of human annotation on a subset of the 3,624 ImageNet total synsets and MIT dataset. We show bounding-box localization (BBox-CorLoc) accuracy for ImageNet dataset and segmentation localization (Seg-CorLoc) accuracy for MIT dataset (see [6] for additional results).

annotations for the actively chosen batch $\mathcal{S}_t$, augment $\mathcal{S}$ with that newly labeled data, and propagate the segmentation as described above. The output after T rounds is the resulting propagated masks on all images. Note that throughout the stages, each unlabeled mask is continually refined, and its intermediate results affect subsequent stages' active selections.

Our active selection algorithm accounts for three criteria—influence, diversity, and uncertainty. The former two criteria account for relationships between images that are relevant to propagation, while the latter accounts for the inherent difficulty of individual images. An image *influential* for propagation is similar to many other images in the collection. Intuitively, labeling such a "hub" image can directly improve the mask quality of the related images, particularly given our match-based unaries and localized image neighborhoods. A batch of images that are *diverse* ensures broad coverage over the entire collection. Selecting images which are influential but also very similar would not lead to a large information gain. Hence, we enforce diversity by adding a penalty for selecting mutually similar images. An image that is *uncertain*—inherently difficult to segment automatically—is also a good candidate for human supervision. We quantify the uncertainty of an image by predicting if its badly segmented using a learned regressor over descriptors suggestive of segmentation quality on an external dataset.

At each stage, we would like to identify the set maximizing all three criteria simultaneously. This is a combinatorial problem over all subsets $\mathcal{S}_t \subseteq \mathcal{I}$ and impractical to solve optimally. We instead employ a greedy approach similar to the maximization of monotone submodular functions to approximately solve this optimization problem.

3 Results

Datasets: We evaluate our approach on ImageNet [12] ($\sim$1M images, 3,624 classes) and the MIT Object Discovery [10] (2488 images, 3 classes) dataset.

Baselines: Apart from an ablated version of our method (i.e., w/o uncertainty), we compare with these baselines:

- Passive: At every stage, randomly pick k images to be labeled by humans.

- PageRank [11]: Only existing active propagation method, uses PageRank importance ranking and clustering to pick k good images at each stage.

- Semantic Propagation [5]: An existing propagation method that promotes propagation between semantically related classes.

- Weakly supervised: We also compare the special case of our method (no human annotation) with several existing approaches [3, 7, 8, 9, 10, 14].

Evaluation metrics: We use: (1) **Jaccard Score:** Standard intersection-over-union (IoU) metric between predicted and ground truth segmentation masks (for MIT) and between bounding boxes (for ImageNet), and (2) **CorLoc Score:** Percentage of images correctly localized according to PASCAL criterion (i.e IoU > 0.5) used in [4, 14]. For MIT we use the segmentation masks (Seg-CorLoc) and for ImageNet we use bounding boxes (BBox-CorLoc) since it lacks ground truth masks.

3.1 Active segmentation propagation

First we present results for active selection. In this setting we iteratively request annotators to provide true segmentations for a subset of images. We then use these labeled images to improve the joint segmentation of other unlabeled images in the collection.

Figures 2 shows sample results on both ImageNet and MIT datasets. On the extreme left on x-axis, we have the performance of the purely weakly supervised setting (no human input) and on the extreme right, annotators provide ground-truth segmentations for all images in the collection. In between we see the trade-off between actively allocating human effort versus other baselines. For all metrics and datasets, the proposed approach outperforms all baselines. While all methods naturally improve with more labeled data, the slope of our improvement curve is substantially sharper using minimal human effort—sometimes dramatically so (e.g., Jetliner on ImageNet or Airplane on MIT). It is important to note that all methods are using identical CNN features and the same propagation algorithm, hence our gains exactly show the impact of making wiser annotation choices.

Surprisingly, we find that the Passive baseline outperforms the active PageRank method [11]. We believe this is because PageRank emphasizes the influence property more, and, despite its clustering component, fails to select sufficiently diverse examples (in [11] no comparison with a passive baseline is shown). On the other hand, our method takes into account influence, diversity, and uncertainty to choose good candidates for annotation. This leads to better annotation choices and in turn better propagation. We also see that omitting uncertainty from our approach decreases accuracy, showing the value of this segmentation-specific active selection component.

We also see that our gains are much higher for larger collections ($>$ 100 images). Larger collections exhibit both greater redundancy and multiple modes within the data. Our method successfully exploits these patterns while making annotation choices. For e.g., in MIT "Airplanes", we correctly localize 90% of the images with only 30% of the data labeled by annotators. In contrast, the Passive and active PageRank baselines require significantly more annotations (55% and 70%, resp.) to achieve the same accuracy.

We also compare with the state of the art segmentation propagation approach from Guillaumin et al. [5]. We consider all images which are common between our experimental setup and that of [5]. For the same amount of labeled data our active segmentation propagation approach achieves a Jaccard score of 65% as opposed to 62.63% by [5]. More importantly, reducing the supervision budget for our method, we achieve the same accuracy as this (passive) state of the art propagation method [5] when using 26% *less* human-annotated data. This large savings in human effort shows the clear value of actively determining where human guidance is most needed.

3.2 Weakly supervised foreground segmentation

Next we test our method in a purely weakly supervised setting against several existing methods. In this special case, weak supervision (i.e., all images have an object from the same category) is the only information available. No additional human annotation is requested. Here we briefly describe our results, please see [6] for details.

On MIT dataset we outperform several existing methods [3, 7, 8, 9, 10] in majority of the classes. On ImageNet, our method outperforms the state of the art [14] by a considerable margin (4.44%), which again highlights the strengths of our joint segmentation graph. With nearly 1M images, a gain of 4.44% means that we correctly localize 41,715 more images than [14].

Figure 3 shows qualitative results. Our method is able to segment objects well in spite of large intra-class variations. Because of the joint segmentation graph, our method can successfully segment some challenging instances where the object is not easily separable from the background but matches well with similar regions in easier images.

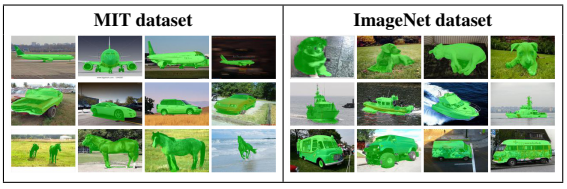

Figure 3: Qualitative results for weakly supervised joint segmentation.

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an Object? In *CVPR*, 2010.

[2] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(9):1124–1137, September 2004. ISSN 0162-8828. doi: 10.1109/TPAMI.2004.60. URL http://dx.doi.org/10.1109/TPAMI.2004.60.

[3] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Enriching Visual Knowledge Bases via Object Discovery and Segmentation. In *CVPR*, 2014.

[4] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100(3):275–293, September 2012.

[5] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. ImageNet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014.

[6] S. Jain and K. Grauman. Active image segmentation propagation. In *CVPR*, 2016.

[7] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

[8] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012.

[9] G.H. Kim, E.P. Xing, L. Fei Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011.

[10] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[11] Michael Rubinstein, Ce Liu, and William T. Freeman. Annotation propagation in large image databases via dense image correspondence. In *ECCV*, 2012.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[13] B. Siddiquie and A. Gupta. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Cl ass Active Learning. In *CVPR*, 2010.

[14] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014.

[15] Alexander Vezhnevets, Joachim M Buhmann, and Vittorio Ferrari. Active learning for semantic segmentation with expected change. In *CVPR*, 2012.

[16] S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, 2010.