Scene-Based Priors for Bayesian Semantic Image Segmentation

Christopher Menart, Ohio State University, USA
James W. Davis, Ohio State University, USA
Muhammad N. Akbar, Ohio State University, USA
Roman Ilin, Air Force Research Laboratory, USA

ABSTRACT

Based on the observation that semantic segmentation errors are partially predictable, this study proposes a compact formula using the confusion statistics of a trained classifier to refine (re-estimate) the initial label hypotheses. The proposed strategy is contingent upon computing the classifier confusion probabilities for a given dataset and estimating a relevant prior on the object classes present in the image to be classified. This study provides a procedure to robustly estimate the confusion probabilities and explore multiple prior definitions. Experiments are shown comparing performances on multiple challenging datasets using different priors to improve a state-of-the-art semantic segmentation classifier. This study demonstrates the potential to significantly improve semantic labeling and motivates future work for reliable label prior estimation from images.

KEYWORDS
ADE20K, Air Force Research Laboratory, Bayesian, Confusion Probabilities, Context, NYUDv2, Ohio State University, PASCAL-Context, Probabilistic Refinement, Refinenet

INTRODUCTION

Semantic segmentation is a challenging computer vision problem wherein a class label is assigned to each pixel in an image. This provides much richer scene information than traditional image classification, and therefore it is inherently a more difficult task. While image classification requires only detection of the primary object, semantic segmentation requires precise, localized pixel-wise detections. Annotating training images for semantic segmentation in a supervised manner is more expensive and time-consuming, and as a result, most public datasets are relatively small in comparison to those for image classification (e.g., ImageNet) (Deng, Todorovic, & Latecki, 2015).

Current semantic segmentation methods show promising results on multiple datasets. However, consistent errors in their output can readily be found. For example, RefineNet (by Lin et al. (2017)) mis-classifies ground as sidewalk so often in the PASCAL-Context dataset (by Mottaghi et al. (2014)) that pixels classified as sidewalk are 60% likely to have the label ground in truth. In this example, pixel accuracy would actually be improved by simply relabeling all sidewalk pixels as ground. However, we would prefer not to lose all ability to classify sidewalks, but rather exploit our knowledge of this confusion to reason about the appropriate labels to assign and the corresponding confidence we should have.

DOI: 10.4018/IJSST.2019010101

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
We categorize labeling errors into two types. An “in-context” labeling error is when a pixel is assigned an incorrect label from the set of actual labels for a given image (e.g., a foreground object label is assigned to an incorrect location). An “out-of-context” labeling error is an invalid pixel label assignment that is outside of the actual label set for the image (e.g., an indoor object label is assigned to a pixel in an image of an outdoor scene). Examples of these errors are shown in Figure 1.

Given that a classifier will have similar performance characteristics on related images, an analysis of the classification errors and label confusions across a dataset should support a secondary (post-processing) refinement stage to re-estimate the output label likelihoods. These more informed classifications should greatly reduce the incidence of in-context errors and make out-of-context errors less likely. The proposed approach is based on these motivations.

Our method is derived from a direct marginalization of $p(l|d)$, the probability of label $l$ given input data $d$, resulting in a decomposition of the formulation into classifier output label probabilities and learned classifier/truth confusions. Applied together, the framework treats the confidence levels of the original classifier output as partial evidence and incorporates the confusion information to determine the final probability of witnessing each object label at each location in an image. In a sense, this confusion approach can be considered a form of context-aware re-estimation. We present a robust method to compute the confusion probabilities and also outline various label priors used to bias the refinement. Upper bound performances of the framework with various priors are reported for three challenging datasets to justify the approach.

The rest of this paper is organized as follows. Section II describes related work in semantic segmentation. Section III describes our framework for classifier refinement using label confusion probabilities and priors. Section IV presents experimental results showing the performance improvements of the approach on multiple datasets, followed by a conclusion in Section V.

RELATED WORK

Convolutional Neural Networks (CNNs) such as those in Lin et al. (2017), Long et al. (2015), Chen et al. (2016), and Lin et al. (2016), have achieved unprecedented results in semantic segmentation. However, they also introduce a tension between precise localization and the inclusion of broader image context (Garcia-Garcia, Ors-Escolano, Oprea, Villena-Martinez, & Garcia-Rodriguez, 2017). Many recent innovations in network architecture have been motivated by this issue.

Dilated convolutions are a modification to CNN filters which attempt to expand the inclusion of context with no change in localization accuracy (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2016). Multi-scale architectures attempt to build features at multiple scales, hopefully catching both localization information and context at different ones (Raj, Maturana, & Scherer, October 2015). Zoom-out features combine information extracted from several nested regions around a superpixel (Mostajabi, Yadollahpour, & Shakhnarovich, 2015). Recursive convolutional networks use recurrent connections to repeatedly entangle the activations at each location in a feature map (Visin, et al., 2016). Convolutional upsampling allows high-resolution information to be re-introduced to a network after contextual features are computed (Ronneberger, Fischer, & Brox, 2015).

Other techniques attempt to exploit context independent of or alongside neural architectures. K-nearest-neighbors methods are used to label images contextually by fetching images of visually similar environments and transferring labels from those environments (Tighe & Lazebnik, 2013). Conditional Random Fields are used to propagate semantic information across the entire image (Shotton, Winn, Rother, & Criminisi, 2006) (Krähenbühl & Koltun, 2011). Precision matrices, by Souly and Shah (2016), perform a similar function. These approaches remain relevant in part due to their flexibility. They operate well not only with deep learning but with various classification techniques. Research has continued to investigate estimating and employing co-occurrence statistics for Conditional Random Fields (CRFs) (Zhang, Zeng, Wang, & Xue, 2015). Some recent methods have attempted to integrate them directly into neural models (Zheng, et al., 2015; Arnab, Jayasumana,
Creep Rupture Forecasting: A Machine Learning Approach to Useful Life Estimation
[www.igi-global.com/article/creep-rupture-forecasting/123952?camid=4v1a](www.igi-global.com/article/creep-rupture-forecasting/123952?camid=4v1a)

Design and Early Simulations of Next Generation Intelligent Energy Systems
[www.igi-global.com/article/design-and-early-simulations-of-next-generation-intelligent-energy-systems/123955?camid=4v1a](www.igi-global.com/article/design-and-early-simulations-of-next-generation-intelligent-energy-systems/123955?camid=4v1a)