Class-agnostic Object Detection

Ayush Jaiswal, Yue Wu, Pradeep Natarajan, Premkumar Natarajan
Amazon Alexa Natural Understanding
Manhattan Beach, CA, USA
{ayujaisw, wuayue, natarap, premknat}@amazon.com

Proposed Class-agnostic Detectors
Convention Class-aware Detectors
Object Proposal Methods
BING EdgeBox Selective Search Faster R-CNN Faster R-CNN SSD

Figure 1: Class-agnostic object detection using object proposal methods (OPMs), conventional class-aware detectors, and the proposed adversarially trained class-agnostic detectors. Green and magenta boxes indicate ground-truth and detection outputs, respectively. OPMs generate regions of interest that could contain objects without precisely locating any object. Class-aware detectors detect objects from a set of known object-types. The proposed class-agnostic detection aims to localize all objects irrespective of their types including those of unknown classes. Results show that the proposed adversarially trained class-agnostic models detect novel objects for which no annotations are available. Best viewed digitally and zoomed in.

Abstract

Object detection models perform well at localizing and classifying objects that they are shown during training. However, due to the difficulty and cost associated with creating and annotating detection datasets, trained models detect a limited number of object types with unknown objects treated as background content. This hinders the adoption of conventional detectors in real-world applications like large-scale object matching, visual grounding, visual relation prediction, obstacle detection (where it is more important to determine the presence and location of objects than to find specific types), etc. We propose class-agnostic object detection as a new problem that focuses on detecting objects irrespective of their object-classes. Specifically, the goal is to predict bounding boxes for all objects in an image but not their object-classes. The predicted boxes can then be consumed by another system to perform application-specific classification, retrieval, etc. We propose training and evaluation protocols for benchmarking class-agnostic detectors to advance future research in this domain. Finally, we propose (1) baseline methods and (2) a new adversarial learning framework for class-agnostic detection that forces the model to exclude class-specific information from features used for predictions. Experimental results show that adversarial learning improves class-agnostic detection efficacy.

1. Introduction

Human visual scene understanding relies on the ability to detect and recognize objects in one’s field of view. This naturally carries over to machine scene understanding through computer vision techniques. Hence, the field of object detection has garnered tremendous research interest [50] and large advances have been made in improving detection systems, especially since the adoption of deep learning.

The object detection task is formulated as the prediction of bounding boxes and classes for objects in a given image.
This requires densely labeled data that contains annotations for all objects in training images. However, creating such datasets is extremely challenging and expensive. Therefore, conventional object detection focuses instead on the reduced task of locating and recognizing “objects of interest” corresponding to limited types of objects that are labeled in the training data, with objects of unknown types treated as background content. Notably, detection of unknown object types is explicitly penalized in the widely adopted mean average precision metric used for benchmarking. This hinders the adoption of trained detectors in real-world applications due to the added cost of retraining them for application-specific object types. Furthermore, such detectors cannot be used in applications like obstacle detection, where it is more important to determine the location of all objects present in the scene than to find specific kinds of objects.

In order to address the aforementioned limitations of conventional class-aware detection, we propose class-agnostic object detection as a new research task with the goal of predicting bounding boxes for all objects present in an image irrespective of their object-types. Intuitively, this task additionally seeks to detect objects of unknown types, which are not present or annotated in the training data. This challenging yet high-reward goal of generalization to unseen object types would benefit downstream applications (e.g., application-specific object classification, object retrieval from large databases, etc.) that can consume such class-agnostic detections.

Besides the problem formulation, we propose training and evaluation protocols for class-agnostic detection with generalization and downstream utility as primary goals. Generalization is evaluated on PASCAL VOC [9], MS COCO [24], and Open Images [20] in the form of recall of unseen classes. Specifically, models trained on VOC are evaluated on unseen classes of COCO and those trained on COCO are tested on non-overlapping classes of Open Images. Furthermore, the VOC dataset is split into seen and unseen classes in order to measure recall of seen and unseen classes within the same dataset. Utility of trained detectors is evaluated as the accuracy of pretrained models used in applications like obstacle detection, where it is more important to determine the location of all objects present in the scene than to find specific kinds of objects.

2. Related Work

Zou et al. [50] provide a comprehensive survey of advances in object detection in the past 20 years. Some conventional detection works [16, 21, 32, 42, 46] describe the task as “generic” object detection but they use it to signify their focus on common objects of known types, which is distinct from our work on detecting objects of all types.

Two-stage object detectors like Faster R-CNN [35] employ a region proposal module that identifies a set of regions in a given image that could contain objects, which are then used by a downstream detector module to localize and identify objects. A number of works [5, 7, 12, 19, 29, 31, 39, 48] have been proposed recently that aim to improve the quality of such object proposals and reduce their number in order to speed up the final detection task. A majority of proposal methods are trained end-to-end with the detector module, which biases them to objects of known types. Methods also exist for generating [6, 28, 40, 49] or filtering [26, 27, 34] object proposals based on edge-related features and object-ness metrics. However, as defined, these proposals do not correspond directly to final detections and must be fed to a detector module to infer final detections. In contrast, the proposed task of class-agnostic detection is aimed at predicting final bounding boxes for all objects in an image. Two-stage methods for class-agnostic detection could employ these works to generate intermediate object proposals.

Some works [14, 45] focus on generating pixel-level objectness scores in order to segment object regions out of background content. These works produce objectness heatmaps which could sometimes be used to detect individual objects if they do not overlap and are strikingly distinct. A few works [8, 13] have shown the efficacy of convolutional neural networks like AlexNet [17] for localizing objects irrespective of their classes. In this work, we take this idea further and formally define the task of class-agnostic object detection along with training and evaluation protocols for benchmarking current and future research.

Kuen et al. [18] show how weights from object recognition networks can be transferred to detection models to scale-up the number of object-types for detection. This line of research is also related to few-shot [10, 15, 30, 41, 43, 44, 47] and zero-shot [1, 22, 33] object detection, which are targeted towards detecting objects of novel types given a few reference samples or descriptions, respectively.
3. Class-agnostic Object Detection

In this section, we describe the proposed task of class-agnostic object detection and contrast it with conventional class-aware object detection. We then discuss a few intuitive baseline models for the proposed task.

3.1. From Class-aware to Class-agnostic Detection

The conventional formulation of object detection can be described as the task of locating and classifying objects in a given image. The goal of models developed to solve this task is to predict bounding box coordinates and class labels for objects present in the image. Given this formulation, a plethora of models have been proposed that train on images with object annotations. However, due to the difficulty and cost associated with collecting and labeling large datasets, annotations are typically collected for a limited number of object categories. This has inadvertently reduced the original formulation to the task of detecting “objects of interest”, corresponding to object types that are annotated in training datasets. Although reduced in form, the conventional formulation remains a challenging problem, with direct application in use-cases with a fixed set of known object types.

In this work, we propose a sibling task of class-agnostic object detection that aims to detect all objects in a given image irrespective of their object-types. More specifically, the goal of this task is to predict bounding boxes for all objects present in an image but not their category. Furthermore, given that most available training datasets do not contain dense annotations for all kinds of objects, an additional implicit goal for models developed for this task is to generalize to objects of unknown types, i.e., those for which annotations are not available in the training dataset. This is in direct contrast to conventional class-aware object detection, which treats unknown objects as background content. As compared to conventional detectors, class-agnostic models can be more easily adopted in complex real-life applications like object retrieval from a large and diverse database, recognition of application-specific object-types through training a downstream object recognition model (instead of retraining the full detection network), etc.

3.2. Baseline Models

We identify three intuitive and straightforward baseline models for solving the class-agnostic detection task, selected due to their ease of implementation and natural curiosity about their performance. The baseline models are:

- region proposal network of a two-stage detector,
- class-aware detector trained for object-type classification and bounding box regression,
- pretrained class-aware model finetuned end-to-end for object-or-not binary classification instead of object-type, along with bounding box regression,
- detection model trained from scratch for object-or-not binary classification and bounding box regression.

4. Adversarial Learning Framework for Class-agnostic Object Detection

4.1. General Framework

Conventional class-aware detection focuses on detecting “objects of interest”, which inherently requires models to be able to distinguish between types of objects in a closed known set. Intuitively, models achieve this by encoding features that are discriminative of object-types. However, for class-agnostic detection and for models to be able to detect previously unseen types of objects, detectors should encode features that more effectively distinguish objects from background content and individual objects from other objects in the image, without discriminating between object-types.

Naively training conventional object detectors for the binary classification task of object-or-not along with bounding box regression is not sufficient to ensure that models fo-
focus on class-agnostic features and more importantly, ignore

type-distinguishing features so that they can better general-
ize to unseen object-types. In order to overcome this prob-
lem, we propose to train class-agnostic object detectors in
an adversarial fashion such that models are penalized for
encoding features that contain object-type information.

We begin with observing a popular two-part pattern in
the model design of both one-stage and two-stage conven-
tional object detectors. The first upstream part of a detec-
tion model learns a set of convolutional features from en-
tire images (one-stage) or regions of interest (two-stage).
The second downstream part consumes these features and
passes them through classification and regression branches
for object-type and bounding box prediction, respectively.
This two-part setup allows for external control on the infor-

data output by the first part and consumed by the second.

We propose to augment class-agnostic detectors with
adversarial discriminator branches that attempt to classify
object-types (annotated in the training data) from the fea-
tures output by the upstream part of detection networks,
and penalize the model training if they are successful. The
models are trained in an alternating manner such that the
discriminators are frozen when the rest of the model is up-
dated and vice versa. While updating the discriminators, we
use the standard categorical cross-entropy loss with object-
types used as prediction targets. On the other hand, while
training the rest of the model, we minimize (a) the cross-
entropy loss for object-or-not classification, (b) smooth L1
loss for bounding box regression, and (c) the negative en-
tropy of discriminator predictions. This entropy maximiza-
tion forces the upstream part of detection models to exclude
object-type information from the features it outputs. The
discriminator is updated five times for every update to the
rest of the model and the negative entropy is weighted with
a multiplier $\alpha$ (tuned on $\{0.1, 1\}$) in the overall objective.
Figure 2 summarizes the complete framework.

During test-time inference, the discriminators are de-
tached from the model, giving back the original network
with the standard layers and parameter-count. Thus, our
framework does not cause performance delays.

### 4.2. Model Instantiation

We demonstrate the applicability of the proposed adver-
sarial framework to both Faster R-CNN (FRCNN), a two-
stage detector, and SSD, a one-stage detector. We use the
publicly available MMDetection [4] framework for imple-
menting the models and running experiments. We train two
versions of the adversarial models – one is trained from
scratch and the other is finetuned from a pretrained base-
line class-aware model. Further details are as follows.

**Faster R-CNN.** The FRCNN model first generates re-
gions of interest in the input image, which is followed by ex-
tracting features from and making object-type and bounding

box predictions for each region. We create a class-agnostic
adversarial version of FRCNN by (1) replacing the multi-
class object-type classification layer with a binary object-
or-not layer, and (2) attaching an adversarial discriminator
on top of the feature extraction layer that provides inputs to
the classification and regression heads. Thus, during train-
ing, this feature layer serves three prediction heads instead
of the standard two. In our experiments, we use the standard
FRCNN model available in MMDetection, which includes a
ResNet-50 backbone and a Feature Pyramid Network [23].

**SSD.** An SSD model utilizes features from several layers
of its backbone network to detect objects of different scales,
corresponding to the depth-levels of the backbone layers.
Specifically, SSD models contain classification and regres-
sion layers for making predictions at each depth-level. In
order to create a class-agnostic adversarial version of the
SSD model, we (1) replace each object-type classification
layer with a binary object-or-not layer, and (2) attach an
adversarial discriminator at each depth-level where predic-
tions are made. Thus, during training, each prediction level
in the resulting model has three prediction heads instead of
the conventional two. We use the standard SSD-300 with
VGG-16 pretrained on ImageNet as the backbone.

### 5. Training and Evaluation Protocols

We propose two kinds of experiments for evaluating class-agnostic detection, geared towards measuring (1) gen-
eralization to unseen object types, and (2) the downstream utility of trained models. While there are several ways to
design experiments for (1) and (2), we propose first steps with potential for refinement in future work.

#### 5.1. Generalization to Unseen Object-types.

Class-agnostic detectors should, by definition, not be limited to object types seen during training. Hence, it is im-
portant to evaluate their efficacy at identifying unseen types of objects. We measure this performance as average recall
(AR; also known as detection rate [49]) at various number
$k \in \{3, 5, 10, 20, 30, 100, 300, 1000\}$ of allowed bound-
ing box predictions, i.e., AR@$k$ (IoU $\geq 0.5$). The goal
here is to achieve high AR@$k$ at all $k$ levels. We design
two sets of generalization experiments based on this setup.

**Experiment I.** We split the VOC dataset into 17 seen
classes and three unseen classes. Seen classes in the VOC
07+12 training set are used for learning the models while
both seen and unseen classes in the VOC 07 validation
set are used for evaluation as AR@$k$ of seen and unseen
classes. The harmonic mean of seen and unseen AR@$k$; for
each $k$, is reported as a measure of overall performance.

In order to select the three unseen classes, we compute the confusion matrix\(^1\) of a 300-layer SSD model trained on

\(^1\)https://github.com/kaanakan/object_detection_confusion_matrix
Figure 3: Generalization results for FRCNN models trained on the seen VOC dataset. The top row shows macro-level AR@$k$ for seen and unseen classes in VOC and their harmonic mean (AR-HM). FRCNN-agnostic-adv performs the best overall. The second row shows micro-level results for the easy, medium, and hard unseen classes. FRCNN-agnostic-adv performs the best on the hard and easy classes with recall drop for the medium class. The last row provides results of evaluation on the COCO data of 60 unseen classes. FRCNN-agnostic-adv achieves the best AR@$k$ for objects of all sizes.

The standard 20-class VOC 07+12 dataset, and use $F_1$ scores to determine easy, medium, and hard classes. Specifically, the class with the highest $F_1$ is the hardest for generalization as it is the most distinguishable from the other classes, making it the most difficult to generalize to if it were not seen during training. Similarly, the class with the lowest $F_1$ is picked as the easy class. The medium class is the one with the median $F_1$. Thus, the unseen set comprises cow (easy), boat (medium) and tvmonitor (hard) classes, while the rest are considered seen. In addition to the macro AR@$k$ scores, micro AR@$k$ are reported for each unseen class. Besides the evaluation on VOC, the (VOC) models are tested for generalization on the 60 non-VOC classes in the COCO 2017 validation set using the same AR@$k$ metric. AR@1000 is further reported for small, medium, and large sized objects belonging to the 60 classes.

Experiment II. In this experiment, models are trained on the COCO 2017 training set and evaluated on non-overlapping classes in the Open Images V6 test set. We first normalize the class names in both COCO and Open Images to find exact matches between classes in the two datasets. We identify a few classes that have slightly different names in the two datasets and normalize them manually. We then build a semantic tree using the publicly available Open Images class hierarchy and exclude all classes that either match exactly with a COCO class or have a COCO class as a predecessor or successor. The remaining classes are used for testing. AR@$k$ is used as the evaluation metric.

5.2. Downstream Utility

A major motivation for research in the domain of class-agnostic object detection is the potential for widespread downstream utilization of class agnostic bounding boxes for extracting objects in images and using them for various visual reasoning tasks. In this work, we propose evaluation of downstream utility in terms of object recognition on images in the ObjectNet [2] dataset. Detection models are used to predict $M$ bounding boxes for each image, which are then
Table 1: Generalization results for experiments I and II. AR@1000-Unseen shows results for unseen VOC classes, AR@1000-COCO for unseen COCO classes, and AR@1000-OI for unseen Open Images classes. “Ovr”, “Med”, “Sml”, and “Lrg” stand for overall, medium, small, and large, respectively. “-aw-” and “-ag-” in the model name indicate whether the model is class-aware or -agnostic, “-ft-” tells whether the model was finetuned from the class-aware baseline, and “-ad” represents models trained adversarially. FRCNN-aw-prop refers to the evaluation of the FRCNN-aw proposals from its first stage Region Proposal Network. Class-agnostic models generalize better than class-aware models with those trained adversarially from scratch performing the best overall.

<table>
<thead>
<tr>
<th>Model</th>
<th>I: Training on seen VOC</th>
<th>II: Training on COCO</th>
<th>AR@1000-OI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ovr</td>
<td>Easy</td>
<td>Med</td>
</tr>
<tr>
<td>FRCNN-aw-prop</td>
<td>0.374</td>
<td>0.388</td>
<td>0.383</td>
</tr>
<tr>
<td>FRCNN-aw</td>
<td>0.545</td>
<td>0.822</td>
<td>0.501</td>
</tr>
<tr>
<td>FRCNN-ft-ag</td>
<td>0.382</td>
<td>0.733</td>
<td>0.302</td>
</tr>
<tr>
<td>FRCNN-ag</td>
<td>0.596</td>
<td>0.814</td>
<td>0.550</td>
</tr>
<tr>
<td>FRCNN-ft-ag-ad</td>
<td>0.528</td>
<td>0.745</td>
<td>0.447</td>
</tr>
<tr>
<td>FRCNN-ag-ad</td>
<td>0.601</td>
<td>0.823</td>
<td>0.537</td>
</tr>
<tr>
<td>SSD-aw</td>
<td>0.663</td>
<td>0.831</td>
<td>0.594</td>
</tr>
<tr>
<td>SSD-ft-ag</td>
<td>0.671</td>
<td>0.819</td>
<td>0.668</td>
</tr>
<tr>
<td>SSD-ag</td>
<td>0.702</td>
<td>0.810</td>
<td>0.669</td>
</tr>
<tr>
<td>SSD-ft-ag-ad</td>
<td>0.683</td>
<td>0.809</td>
<td>0.651</td>
</tr>
<tr>
<td>SSD-ag-ad</td>
<td>0.722</td>
<td>0.833</td>
<td>0.689</td>
</tr>
</tbody>
</table>

use to crop the image into \( M \) versions. Pretrained\(^2\) ImageNet classifiers are then used to predict the object class from the cropped images. Here we use ResNet-152 [11], MobileNet-v2 [37], and Inception-v3 [38] models.

Two metrics are used for evaluation – Accuracy@\( M \) for \( M \in \{1, 5, 10\} \), and Best-overlap (BO) accuracy. For Accuracy@\( M \), the classifier’s prediction on at least one of the \( M \) crops needs to be correct for the image to be considered as successfully classified. BO-accuracy is calculated using bounding boxes that have the highest intersection over union (IoU) with the ground-truth [3] bounding boxes.

6. Results

6.1. Generalization to Unseen Object-types

Experiment I. We train the baseline models and the adversarial models on the VOC seen set and evaluate them on both the VOC set and the COCO unseen classes. Results of this experiment with FRCNN as the base model are presented in Figure 3. Class-agnostic models outperform standard class-aware models in overall performance, especially for unseen VOC classes. Breakdown of recall for unseen classes reveals that the margin in recall becomes larger as the difficulty of generalizability is increased, with the largest performance gains on the hard unseen class. Results on the COCO unseen classes show that the adversarial model performs the best overall and for objects of all sizes.

Experiment II. In this experiment, the baseline and the adversarial models are trained on the COCO dataset and evaluated on the non-overlapping classes of the Open Images, as outlined in Section 5.1. Figure 5 summarizes the results for both FRCNN and SSD models. In both cases, the class-agnostic models generalize better than the class-aware models overall, with the proposed class-agnostic models trained from scratch achieving the best AR scores.

Table 1 presents AR@1000 generalization scores for experiments I and II, showing that the adversarially trained (from scratch) models perform the best in both settings.

We find that the recall of models that are finetuned from the pretrained class-aware baselines are worse than those that are trained from scratch. We attribute this to the dif-
Figure 4: Generalization results for SSD models trained on the seen VOC dataset. The top row shows macro-level AR@k for seen and unseen classes in VOC as well as their harmonic mean (AR-HM). SSD-agnostic-adv performs the best on AR-Unseen and AR-HM, with a drop in AR-Seen, but the models that outperform SSD-agnostic-adv on AR-Seen do significantly worse on AR-Unseen and AR-HM. The second row shows micro-level results for the easy, medium, and hard unseen classes. SSD-agnostic-adv performs the best in all categories. The last row provides results of evaluation on the COCO data of 60 unseen classes. SSD-agnostic-adv achieves the best AR@k with a slight reduction for small-sized objects.

6.2. Downstream Object Recognition

We evaluate the object recognition performance for images cropped using bounding boxes predicted by the class-aware and class-agnostic models trained for the generalization experiments in Section 5.1. Table 2 summarizes the results of these experiments. Results show that the adversarially trained class-agnostic models perform better than the baselines in general on both Accuracy@M and Best-overlap accuracy (as described in Section 5.2), irrespective of the ImageNet classifier used. These results correlate with the generalization results presented in Section 6.1, indicating that detection models that generalize better to unseen object-types have higher downstream utility.

7. Conclusion and Future Work

Conventional class-aware objection detection models perform well at detecting “objects of interest” that belong to known objects types seen during training. However, these models do not generalize well to unseen object-types, limiting their incorporation in real-world applications. In this work, we have formulated a novel task of class-agnostic object detection, where the goal is to detect objects of known and unknown types irrespective of their category. Furthermore, we have presented training and evaluation protocols for benchmarking models and advancing future research in this direction. These include two sets of experiments – (1) one for generalization to new object-types and (2) another for downstream utility in terms of object recognition. Finally, we have presented a few intuitive baselines and proposed a new adversarially trained model that penalizes the objective if the learned representations encode object-type.
Figure 5: Generalization results of (a) FRCNN and (b) SSD models trained on the COCO dataset and evaluated on the non-overlapping classes in the Open Images dataset. Results show that class-agnostic models generalize better than the class-aware models, with the SSD-agnostic-adv and FRCNN-agnostic-adv models achieving the best recall.

Table 2: Downstream ObjectNet recognition results. Images are cropped using boxes predicted by the detectors and pretrained ImageNet models are used for classification from the cropped images. Averages of accuracies across classifiers are also reported. BO-acc is the accuracy when the predicted box with the highest IoU with the ground-truth box is used. Accuracies are also presented for top $M$ proposals as Acc@$M$. Results for uncropped images and those cropped using ground-truth boxes are provided for reference. “-aw-” and “-ag-” in the model name indicate whether the model is class-aware or -agnostic; “-ad” represents models trained adversarially. Results of other baseline detectors are provided in the supplementary material. Adversarially trained class-agnostic models achieve the best results.

Results show that the proposed adversarial class-agnostic model outperforms the baselines on both generalization and downstream-utility experiments, for both one-stage (SSD) and two-stage (Faster R-CNN) detectors.

This work serves to establish a new direction of research in object detection. As such, the possibilities for future research are endless, including but not limited to — (1) further refinement of the problem formulation, (2) improved as well as additional training and evaluation protocols, and (3) novel methods for class-agnostic object detection.
References


