Adversarially Optimizing Intersection over Union for Object Localization Tasks

Sima Behpour  
Department of Computer Science  
University of Illinois at Chicago  
sbehpo2@uic.edu

Kris M. Kitani  
Department of Computer Science  
Carnegie Mellon University  
kkitani@cs.cmu.edu

Brian D. Ziebart  
Department of Computer Science  
University of Illinois at Chicago  
bziebart@uic.edu

Abstract

An implicit uncertainty exists in the annotations of computer visions datasets due to annotator disagreement and the high-dimensional space that annotations must be selected from. Rather than attempting to remove all annotation uncertainty, which we view as hopeless, or ignoring it, which can be detrimental, we choose to embrace uncertainty in the design of our learning approach. Specifically, we address uncertainty adversarially by approximating provided datasets annotations within a game-theoretic formulation of prediction tasks. The adversarial approximator is constrained to resemble the training data annotations according to a set of specified features. This induces a learned feature-based potential function that we then apply to new test cases. We demonstrate the efficiency and predictive performance of our approach on the ILSVRC2012 image dataset, showing significant improvements over existing methods.

Introduction

Annotations for computer vision datasets are often treated as being completely certain, despite significant annotator disagreement in the annotation process (Welinder et al. 2010; Nowak and Rüger 2010) that is not fully resolved by consensus-forming methods (Upchurch et al. 2016; Saragih, Lucey, and Cohn 2011; Zhu and Ramanan 2012; Sagonas et al. 2013). For example, in the construction of the ILSVRC2012 image dataset (Russakovsky et al. 2015), proposed bounding boxes were rejected by other annotators 37.8% of the time (Su, Deng, and Fei-Fei 2012), and recent evaluations have been performed on “a subset of the validation set which omit[s] about 1700 blacklisted entities [of 50000 total] due to poor bounding boxes” (Szegedy, Ioffe, and Vanhoucke 2016). Dataset augmentation methods that supplement ground truth annotations of images with perturbed annotations help to address this inherent uncertainty and have proven beneficial in practice. However, choosing perturbations that improve performance on specific evaluation measures (e.g., thresholded intersection over union for object localization) is an error-prone process. While larger perturbations may effectively increase the amount of training data, they can also cause the learned predictor to drift away from learning the intended concept in a manner that is sensitive to the specific performance measure.

We propose to integrate annotation perturbations directly into the learning process in a manner that is sensitive to the performance measure of interest by introducing an adversary who approximates the ground truth annotations in the worst ways possible. We constrain the adversary to match feature moments of the ground truth annotations. This forces the adversary’s approximations of the annotations to resemble the ground truth annotations of the training data. Thus, our approach takes the form of a zero-sum game between a predictor player seeking to maximize performance and a constrained adversarial approximator seeking to minimize expected performance (Grünwald and Dawid 2004; Asif et al. 2015; Wang et al. 2015). We transform this game, which is effectively over the entire training set, into a set of independent games that share only a learned feature-based potential function that motivates the adversary to satisfy the feature-matching constraint.

Though our approach is general and can be applied to a range of annotation types, we focus on object localization tasks that are evaluated using intersection over union (IoU) between a predicted bounding box and the annotated bounding box. Successful object detection is often defined by an IoU exceeding 50% (Cinbis, Verbeek, and Schmid 2014; Siva et al. 2013; Tang et al. 2014) or 70% (Felzenszwalb et al. 2010; Dai and Hoiem 2012). We consider both of these thresholded IoUs and the average IoU as performance measures of interest. We investigate both the guarantees of our approach—in terms robustly bounding the performance—and its practical benefits. For the latter, we employ our
method as the final convolutional layer of a pre-existing deep architecture and compare against other common final layer predictions. We show that our adversarial approach: (1) provides better task performance in general; and (2) is more adaptive to the performance measure for which it is trained.

Our contribution is distinct from adversarial methods for generating images, such as generative adversarial networks (GANs) (Goodfellow et al. 2014; Salimans et al. 2016; Chen et al. 2016; Mirza and Osindero 2014; Denton et al. 2015), which approximately solve games between a generative model seeking to confound a discriminator trying to distinguish between real and synthetic samples. Instead, the games of our approach are played only over annotations of the images and not in the generation of the images themselves. This provides a distinct advantage in that the resulting games can be solved exactly, but limits the robustness of the approach to perturbations solely for the annotations and not in the training images. However, like GANs, our approach leverages the benefits of constructing a predictor against an adversary that is able to adapt to the predictor’s behavior. We believe that this has significant implications for how deep learning methods should best be discriminatively leveraged.

Bayes Optimality for Specialized Measures

Given an uncertain annotation, Bayes optimal predictions are the best that a learning algorithm can hope to provide. For input $x$, these minimize the expected loss under the annotation distribution $P(y|x)$:

$$f(x) = \arg\min_{\hat{y}} \sum_{y \in \mathcal{Y}} P(y|x) \text{loss}(\hat{y}, y).$$

Unlike having annotations with complete certainty, Bayes optimal predictions depend on the specific loss function (negation of a performance measure) considered. We focus on measures based on the intersection over union (IoU), $IoU(A,B) = \frac{\text{area}(A \cap B)}{\text{area}(A \cup B)}$, used in the Visual Object Classes (VOC) challenges (Cho et al. 2015; Everingham et al. 2005; Siva et al. 2013; Blaschko and Lampert 2008).

Here, $A$ and $B$ are the predicted and evaluation bounding boxes. Specifically, we focus on losses defined in terms of the amount of non-overlap:

$$\text{loss}_{1-IoU}(A,B) = 1 - IoU(A,B),$$

which equal to one when $A$ and $B$ are disjoint, zero when they are identical, and smoothly transitions in between those extremes; and thresholded overlap losses:

$$\text{loss}_{IoU<\alpha}(A,B) = \begin{cases} 1 & IoU(A,B) < \alpha \\ 0 & IoU(A,B) \geq \alpha, \end{cases}$$

which assigns zero or one loss to bounding boxes with or without sufficient overlap.

Classification and Empirical Risk Minimization

We consider predicting image annotations, like bounding boxes, as a supervised machine learning task. The problem domain considers examples from an input space $x \in \mathcal{X}$, each with a an annotation $y \in \mathcal{Y}$. A prediction, $\hat{y}$, is needed for input $x$ based on a set of training data—pairs $(x,y)$ that are distributed according to sample distribution $P(x,y)$. We let $\phi(y,x)$ denote a vector feature function representing some relationships between the input values and that annotations.

Existing classification methods are predominantly based on empirical risk minimization (ERM). This approach seeks the parameters for a parametric classifier, $f_\theta(x)$, that optimally minimizes expected risk:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{x,y \sim P} [\text{loss}(Y, f_\theta(X))].$$

Unfortunately, for losses that are non-convex (or even non-continuous), such as Eqs. (2) and (3), solving this problem is NP-hard (Steinwart and Christmann 2008).

Instead, surrogate loss functions are minimized in place of non-convex/non-continuous performance measures. One of the most widely used surrogates for learning potential functions, $\psi_\theta(y,x) = \theta \cdot \phi(y,x)$, which are maximized to make predictions, $f(x) = \arg\max_{y \in \mathcal{Y}} \psi_\theta(y,x)$, is the hinge loss (Wu and Liu 2012):

$$\text{hinge}_{\text{loss}}(\psi_\theta(\cdot), y) = \max_{\hat{y} \in \mathcal{Y}} \text{loss}(\hat{y}, y) - (\psi(y,x) - \psi(\hat{y}, x)).$$

Bayesian Categorization

However, disagreement between annotators is a widely-known concern on measures based on the intersection over union (IoU), $IoU(A,B) = \frac{\text{area}(A \cap B)}{\text{area}(A \cup B)}$, used in the Visual Object Classes (VOC) challenges (Cho et al. 2015; Everingham et al. 2005; Siva et al. 2013; Blaschko and Lampert 2008).

Here, $A$ and $B$ are the predicted and evaluation bounding boxes. Specifically, we focus on losses defined in terms of the amount of non-overlap:

$$\text{loss}_{1-IoU}(A,B) = 1 - IoU(A,B),$$

which equal to one when $A$ and $B$ are disjoint, zero when they are identical, and smoothly transitions in between those extremes; and thresholded overlap losses:

$$\text{loss}_{IoU<\alpha}(A,B) = \begin{cases} 1 & IoU(A,B) < \alpha \\ 0 & IoU(A,B) \geq \alpha, \end{cases}$$

which assigns zero or one loss to bounding boxes with or without sufficient overlap.

Classification and Empirical Risk Minimization

We consider predicting image annotations, like bounding boxes, as a supervised machine learning task. The problem domain considers examples from an input space $x \in \mathcal{X}$, each with a an annotation $y \in \mathcal{Y}$. A prediction, $\hat{y}$, is needed for input $x$ based on a set of training data—pairs $(x,y)$ that are distributed according to sample distribution $P(x,y)$. We let $\phi(y,x)$ denote a vector feature function representing some relationships between the input values and that annotations.

Existing classification methods are predominantly based on empirical risk minimization (ERM). This approach seeks the parameters for a parametric classifier, $f_\theta(x)$, that optimally minimizes expected risk:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{x,y \sim P} [\text{loss}(Y, f_\theta(X))].$$

Unfortunately, for losses that are non-convex (or even non-continuous), such as Eqs. (2) and (3), solving this problem is NP-hard (Steinwart and Christmann 2008).

Instead, surrogate loss functions are minimized in place of non-convex/non-continuous performance measures. One of the most widely used surrogates for learning potential functions, $\psi_\theta(y,x) = \theta \cdot \phi(y,x)$, which are maximized to make predictions, $f(x) = \arg\max_{y \in \mathcal{Y}} \psi_\theta(y,x)$, is the hinge loss (Wu and Liu 2012):

$$\text{hinge}_{\text{loss}}(\psi_\theta(\cdot), y) = \max_{\hat{y} \in \mathcal{Y}} \text{loss}(\hat{y}, y) - (\psi(y,x) - \psi(\hat{y}, x)),$$
Adversarial Object Localization

of bounding boxes (Figure 2) to minimize this payoff. Then, we formulate object localization against adaptive annotation approximations as a two-player game between a predictor and a player (adversary) that is defined jointly over the adversary’s bounding boxes from all input images \( x \). This provides the adversary with the useful option of coordinating annotations across different examples to satisfy the constraints. Because of this, the resulting game has \(|Y|^n\) strategies (one for the combination of bounding boxes across each image). Therefore, solving this primal optimization problem directly is impractical when the number of training examples is large. We efficiently address this challenge in later sections.

**Dual Optimization**

We consider the dual optimization problem, which often allows optimization over a much smaller number of variables. The dual AOL game (Boyd and Vandenberghe 2004) is a set of independent zero-sum games for each training example tied together only by Lagrange multiplier vector \( \theta \).

**Definition 2.** The dual Adversarial Object Localization (AOL\(d\)) game is:

\[
\min_{\theta} \mathbb{E}_{X,Y \sim \tilde{P}} \left[ \min_{\tilde{P}} \mathbb{E}_{\tilde{Y} \sim \tilde{P}} \left[ \min_{\tilde{Y} \sim \tilde{P}} \mathbb{E}_{\tilde{Y} \sim \tilde{P}} \left[ \text{loss}(\tilde{Y}, \tilde{Y}) \right. \right. \right. \\
+ \theta \cdot \phi_{\text{CNN}}(X, \tilde{Y}) \bigg| X \bigg] - \theta \cdot \phi_{\text{CNN}}(X, Y) \bigg].
\]

Due to strong Lagrangian duality (Boyd and Vandenberghe 2004), the solution to this problem is equivalent to the primal game formulation (Definition 1) (Asif et al. 2015).\(^1\)

This dual form for solving the AOL game provides a Lagrangian-augmented payoff matrix (Table 1) with potential terms, \( \psi = \theta \cdot \phi_{\text{CNN}}(X, \tilde{y}) \), that additionally motivate the adversary to select or avoid selecting bounding boxes that are more similar to or different from training data feature representations.

\(^1\)Small amounts of slack in the primal constraints introduce regularization of \( \theta \) in the dual (Dudik and Schapire 2006)
statistics. Conceptually, the Nash equilibrium for this zero-sum game can be obtained efficiently using a pair of linear programs:

$$
\min_v v \text{ such that: } p^T C \leq v 1 \text{ and } p^T 1 = 1; \quad \text{and} \quad (9)
$$

$$
\max_v v \text{ such that: } C p \geq v^T 1 \text{ and } p^T 1 = 1,
$$

where $v$ is the value of the game and $C$ is the game matrix (Table 1) with values: $C_{\hat{y}, \hat{y}} = \text{loss}(\hat{y}, \hat{y}) + \psi(\hat{y})$. However, forming and solving such a game over a very large set of bounding boxes for each image is computationally burdensome.

### Efficient Game Solutions

Given the computational costs of forming the full payoff matrix for the Lagrangian-augmented game (Definition 2) and solving a linear program (Equation 9) using it, we employ a constraint-generation method (McMahan, Gordon, and Blum 2003; Wang et al. 2015) to more efficiently solve ALO. Its operation is depicted in Figure 3 and further detailed in Algorithm 1.

The approach works by iteratively obtaining a Nash equilibrium for a game defined over a subset of the bounding boxes, finding a player’s best response strategy (bounding box)—either the Bayesian loss minimizer (1) or maximizer—to that equilibrium distribution, and then adding the best response to the set of strategies defining the game. When additional best responses no longer improve either player’s game value, the subgame equilibrium is guaranteed to be an equilibrium to the larger game (McMahan, Gordon, and Blum 2003). In practice, we accelerate this algorithm by adding not just the single best bounding box, but up to ten bounding boxes with best response values that are greater than the previous equilibrium game value.

Table 1: The payoffs of the adversarial object localization game for a predicted bounding box ($\hat{y}$) determined by the row of the table and adversarial bounding box (y) determined by the column of the table. The payoff value for the combination of bounding boxes $\hat{y}$ and $y$ is the combination of the loss, $\ell(\hat{y}, y)$ (e.g., $1 - IoU(\hat{y}, y)$, $IoU(\hat{y}, y) < \alpha$) and the Lagrangian potential of the adversary’s bounding box, $\psi(\hat{y}) = \theta \cdot \phi_{CNN}(x, \hat{y})$. Under a mixed equilibrium, which exists when potentials of the bounding boxes with the largest potentials do not differ greatly, each player chooses a distribution over the rows or columns of this table. Algorithm 1 incrementally expands this table (as outlined in Figure 3) by adding each player’s best additional bounding box until no improvement results (meaning the equilibrium for the larger game is recovered).

| $\hat{y} = \text{red}$ | $\hat{y} = \text{green}$ | $\hat{y} = \text{blue}$ | $\hat{y} = \text{purple}$ | $\cdots$ |
|------------------------|------------------------|------------------------|------------------------|--------|
| $y = \text{red}$       | $\ell(\text{red}, \text{red}) + \psi(\text{red})$ | $\ell(\text{red}, \text{green}) + \psi(\text{green})$ | $\ell(\text{red}, \text{blue}) + \psi(\text{blue})$ | $\ell(\text{red}, \text{purple}) + \psi(\text{purple})$ | $\cdots$ |
| $y = \text{green}$     | $\ell(\text{green}, \text{red}) + \psi(\text{red})$ | $\ell(\text{green}, \text{green}) + \psi(\text{green})$ | $\ell(\text{green}, \text{blue}) + \psi(\text{blue})$ | $\ell(\text{green}, \text{purple}) + \psi(\text{purple})$ | $\cdots$ |
| $y = \text{blue}$      | $\ell(\text{blue}, \text{red}) + \psi(\text{red})$ | $\ell(\text{blue}, \text{green}) + \psi(\text{green})$ | $\ell(\text{blue}, \text{blue}) + \psi(\text{blue})$ | $\ell(\text{blue}, \text{purple}) + \psi(\text{purple})$ | $\cdots$ |
| $y = \text{purple}$    | $\ell(\text{purple}, \text{red}) + \psi(\text{red})$ | $\ell(\text{purple}, \text{green}) + \psi(\text{green})$ | $\ell(\text{purple}, \text{blue}) + \psi(\text{blue})$ | $\ell(\text{purple}, \text{purple}) + \psi(\text{purple})$ | $\cdots$ |
| $\vdots$               | $\vdots$               | $\vdots$               | $\vdots$               | $\vdots$   | $\vdots$   |

Algorithm 1 ALO Equilibrium Computation

**Input:** Image $img$; Parameters $\theta$

**Output:** Nash equilibrium, $(\hat{P}, \hat{P})$

1. $BoxProposals \leftarrow EdgeBox(img)$
2. $\Phi = \text{VggNet.LastCNNLayer}(img, BoxProposals)$
3. $\psi \leftarrow \theta \cdot \Phi$
4. $\hat{S} \leftarrow \hat{S} \leftarrow \arg\max_y \psi(\hat{y})$
5. repeat
6. $(\hat{P}, \hat{P}, \hat{v}) \leftarrow \text{solveGame}(\psi(\hat{S}), \text{loss}(\hat{S}, \hat{S}))$
7. $(\hat{y}_{\text{new}}, v_{\text{max}}) \leftarrow \max_y E_{\hat{Y}, \hat{P}}[\text{loss}(\hat{y}, \hat{Y}) + \psi(\hat{y})]$
8. $\text{if } (\hat{v} \neq v_{\text{max}}) \text{ then}$
9. $\hat{S} \leftarrow \hat{S} \cup \hat{y}_{\text{new}}$
10. end if
11. $(\hat{P}, \hat{P}, \hat{v}) \leftarrow \text{solveGame}(\psi(\hat{S}), \text{loss}(\hat{S}, \hat{S}))$
12. $(\hat{y}_{\text{new}}, v_{\text{min}}) \leftarrow \min_y E_{\hat{Y}, \hat{P}}[\text{loss}(\hat{Y}, \hat{y})]$
13. $\text{if } (\hat{v} \neq v_{\text{min}}) \text{ then}$
14. $\hat{S} \leftarrow \hat{S} \cup \hat{y}_{\text{new}}$
15. end if
16. until $\hat{v} = v_{\text{max}} = \hat{v} = v_{\text{min}}$
17. return $(\hat{P}, \hat{P})$

solution. First, it is sparse, with only a very small number of strategies having non-zero probability. This provides a number of computational benefits, particularly when incorporated into a larger system and used for backpropagation, for example. Second, it is mixed, incorporating a (non-degenerate) distribution over multiple bounding boxes. This is unlike support vector machines, in which only the bounding box with highest loss-augmented potential and the ground truth bounding box are used to construct the gradient. This combination of a mixed, yet sparse inference solution provides a more useful gradient for updating our model’s parameters that balances computational efficiency with sensitivity to the application performance measure.

During testing time, we employ a similar inference pro-
Experimental Evaluation

We seek to investigate two main questions about our developed approach:

1. Does treating bounding box annotations as being uncertain using adversarially approximating provide benefits compared to treating the annotations as certain?

2. Does the annotation approximation provided by the game-theoretic formulation of our approach provide benefits versus annotation perturbation methods that are orthogonal to the learning method?

We consider two object localization tasks that are evaluated using IoU: localizing an object of known class; and detecting (localizing and labeling) an object of unknown class. We conduct both localization and detection experiments to investigate our first question and a localization experiment to investigate our second question.
To show the relative performance of AOL for localization, we benchmark it against a multiclass support vector machine (SVM) trained to incorporate the overlap into its loss function using a structured support vector machine (SSVM) formulation (Vedaldi 2011) and multiclass logistic regression. For each of these models, we train three versions based on three different training objectives.

For the SVM-based method, we use an existing SSVM implementation (Vedaldi 2011) to produce predictions. It employs constraint generation and, similar to our inference procedure, uses a technique to accelerate the learning process by adding multiple diverse constraints at each pass through the bounding boxes. We train one variant of the SSVM model (denoted SSVM) using non-overlap as the loss function, 

\[
\text{loss}_{1-\text{IoU}} \quad \text{(Equation (2))}, \quad \text{a second and third variant (denoted SSVM}_{50} \quad \text{and SSVM}_{70}) \quad \text{using thresholded loss, loss}_{\text{IoU}<50\%} \quad \text{and loss}_{\text{IoU}<70\%} \quad \text{(Equation (3))}. \quad \text{For the logistic regression method (i.e., softmax), we estimate a distribution over all proposed bounding boxes that maximizes the conditional likelihood of proposed bounding boxes with an IoU with the ground truth bounding box annotation exceeding 50\%}. \quad \text{We produce bounding box predictions from the Bayesian optimal decision for the estimated conditional bounding box distribution using Equation (1) and the specified loss function (LR for 1-IoU or LR}_{50} \quad \text{and LR}_{70} \quad \text{for thresholded IoU).}
\]

Similarly, we train our AOL method for the non-overlap (AOL) and for the thresholded IoU (AOL}_{50} \quad \text{and AOL}_{70}) \quad \text{loss functions.}

For object detection (localization and class labeling), we focus on the AOL}_{70}, SSVM}_{70}, \quad \text{and LR}_{70} \quad \text{predictors. Correctness requires a bounding box prediction with at least 70\% IoU and a correct label prediction.}

For this task, the data from all ten classes are pooled together and used as one single training or testing dataset. Each method learns a class-specific vector of parameters that must also be discriminative between classes, in addition to discriminating bounding boxes with high IoU exceeding 70\% from those without sufficiently large IoU.

For both tasks, all methods are provided with the same CNN features for each bounding box proposal. Though further optimization of the VGG16 network for this specific set of tasks is possible with all methods, we refrain from doing so, as not to confound the benefits of AOL on improving feature representation learning with its benefits for leveraging already-learned features. We leave further refinement of the deep learning architecture using AOL for future investigation.

**Evaluation Results for Localization**

We first consider the task of localizing a single object with known label in an image. We evaluate the performance of each approach using the IoU between predicted bounding box and ground truth bounding box thresholded at 50\% (Table 3) and 70\% (Table 4).

| Method | Object class | Airplane | Bird | Bus | Car | Cat | Cow | Dog | Horse | Monitor | Sofa | mAP |
|--------|--------------|---------|-----|-----|-----|-----|-----|-----|-------|---------|------|-----|
| AOL_{50} | 82.0 | 93.5 | 92.0 | 100.0 | 98.1 | 100.0 | 93.0 | 96.4 | 96.0 | 90.0 | 94.2 |
| LR_{50} | 84.0 | 86.5 | 84.0 | 87.0 | 79.7 | 83.0 | 62.0 | 72.7 | 72.0 | 80.0 | 77.7 |
| SSVM | 90.0 | 92.5 | 92.0 | 83.0 | 77.7 | 87.5 | 72.7 | 90.0 | 78.0 | 77.7 |
| AOL_{70} | 94.0 | 95.4 | 96.0 | 100.0 | 96.4 | 97.2 | 98.0 | 96.3 | 95.7 | 95.7 | 96.7 |
| LR_{70} | 91.9 | 94.1 | 90.0 | 96.0 | 83.6 | 80.0 | 82.8 | 90.9 | 90.0 | 94.0 | 89.3 |
| SSVM_{70} | 90.0 | 90.4 | 88.0 | 94.0 | 85.4 | 92.5 | 85.8 | 92.7 | 95.8 | 100.0 | 89.4 |
| AOL | 70.3 | 68.1 | 68.2 | 79.2 | 60.0 | 75.5 | 64.7 | 63.9 | 70.0 | 68.3 | 69.0 |
| LR | 64.9 | 64.0 | 61.3 | 67.1 | 54.9 | 58.1 | 54.6 | 57.0 | 59.0 | 60.2 | 60.1 |
| SSVM | 69.4 | 68.1 | 63.1 | 68.6 | 50.8 | 68.2 | 58.9 | 61.2 | 67.2 | 65.8 | 64.1 |

| Method | Object class | Airplane | Bird | Bus | Car | Cat | Cow | Dog | Horse | Monitor | Sofa | mAP |
|--------|--------------|---------|-----|-----|-----|-----|-----|-----|-------|---------|------|-----|
| AOL_{50} | 52.0 | 45.0 | 44.0 | 23.0 | 18.6 | 16.7 | 32.0 | 29.7 | 25.2 | 18.0 | 44.8 |
| LR_{50} | 38.0 | 33.0 | 20.0 | 45.0 | 12.7 | 12.5 | 14.0 | 99.1 | 14.0 | 22.0 | 22.0 |
| SSVM | 40.0 | 45.0 | 32.0 | 45.0 | 20.4 | 19.4 | 15.1 | 90.7 | 12.0 | 30.0 | 30.9 |
| AOL_{70} | 58.0 | 61.5 | 64.0 | 91.0 | 30.9 | 74.7 | 58.0 | 58.2 | 61.6 | 61.9 | 62.5 |
| LR_{70} | 47.6 | 45.7 | 40.0 | 62.8 | 20.0 | 42.5 | 25.1 | 25.4 | 31.4 | 44.2 | 38.5 |
| SSVM_{70} | 51.8 | 55.5 | 44.0 | 61.7 | 21.8 | 54.7 | 31.6 | 43.6 | 56.0 | 57.3 | 47.8 |

The relative benefits of our AOL approach are more pronounced at the 70\% IoU evaluation threshold (Table 4). We see specifically that AOL_{70} provides the most accurate localizations across all classes and outperforms the nearest competitor, SSVM_{70}, by a substantial margin, on average. Of more interest is that the difference in relative degradation of AOL_{70} (from 96.7% to 62.3\%) and AOL_{50} (from 94.2% to 44.8\%) between evaluation using the 50\% IoU threshold (Table 3) and the 70\% IoU threshold. This strongly suggests that AOL_{70} has been highly optimized towards this specific performance measure, while AOL_{50} produces a large number of bounding boxes with IoU between 50\% and 70\%, as it was designed to accomplish.

A general observation we make based on both sets of results is that training each of these methods on the more challenging objective measures of IoU> 70\% tends to improve the performance for both of these evaluation measures. One explanation is that the thresholded loss function creates a
sharper “strategic” game between predictor and adversary in the AOL formulation and more influential support vectors for SSVM. Though we also conducted additional unreported experiments with larger overlap thresholds, this relationship did not hold beyond the 70% threshold. We believe this is because the game becomes “too easy” for the adversary at these threshold levels. Namely, the adversary can distribute its annotations over sets of bounding boxes that do have IOU with one another above the threshold amount, but still effectively match the ground truth feature constraints.

Benefits of annotation perturbations: For our last set of experiments, we investigate our second question of whether game-theoretic approximation of labels is better than a decoupled annotation perturbation approach. To study this, we compare the benefits of adding perturbed training examples to the dataset when training SSVM to assess the improvements this yields for localization. We do so by supplementing training example annotations with all bounding boxes generated from EdgeBox that exceed an overlap threshold with the “ground truth” bounding box. Note that AOL already automatically incorporates non-ground truth bounding boxes as its approximate annotation distribution (from the adversary) and does not require this augmentation.

Table 5: Testing performance (IOU > 70%) of SSVM trained with a dataset augmented by bounding boxes with overlap exceeding the specified threshold.

| Method | Object class | Airplane | Bird | Bus | Car | Cat | Cow | Dog | Horse | Monitor | Sofa | Avg |
|--------|--------------|----------|------|-----|-----|-----|-----|-----|-------|--------|------|-----|
| SSVM30 |              | 51.8     | 57.9 | 49.9 | 64.0 | 22.6 | 55.2 | 57.3 | 45.5   | 56.7   | 57.8 | 50.3 |
| SSVM60 |              | 54.7     | 58.9 | 52.7 | 67.7 | 23.7 | 64.9 | 42.0 | 48.6   | 57.3   | 58.4 | 52.9 |
| SSVM80 |              | 56.4     | 61.6 | 56.8 | 70.8 | 25.4 | 67.3 | 49.1 | 51.9   | 58.6   | 58.8 | 55.7 |
| SSVM30 |              | 52.6     | 61.0 | 51.7 | 64.4 | 20.2 | 61.2 | 42.6 | 44.0   | 57.3   | 56.0 | 51.1 |
| SSVM60 |              | 49.8     | 52.0 | 44.9 | 60.3 | 20.2 | 55.8 | 33.1 | 41.4   | 55.8   | 52.7 | 46.6 |

The results of this experiment are shown in Table 5. We note that augmenting SSVM in this manner does improve test performance beyond that realized without adding perturbed annotations (Table 4) and that this gain is maximized around the 70% overlap threshold in particular. However, we find that with the exception of the Bird class, the performance gains do not reach the performance of our AOL approach.

Evaluation Results for Detection

We next investigate the more challenging object detection task of locating and labeling an unknown object that is one of the ten object classes. For these experiments, we restrict our consideration to models trained based on the 70% IOU threshold level due to the superior results shown for the localization task. Table 6 shows the object detection performance when evaluated at the 70% IOU threshold for correctness.

We again find strong support for our adversarial approach to dealing with uncertainty. Specifically, we find that AOL provides the best performance for all object classes. Though the relative performance advantage differs by object type, for classes like Dog, the improvement over the other approaches is nearly double. On average, AOL provides a significant performance improvement on this task over both SSVM and LR.

Table 6: Object detection evaluated based on: overlap > 70% and correct class.

| Method | Airplane | Bird | Bus | Car | Cat | Cow | Dog | Horse | Monitor | Sofa |
|--------|----------|------|-----|-----|-----|-----|-----|-------|---------|------|
| SSVM30 | 46.0     | 55.5 | 60.0 | 66.0 | 25.3 | 50.0 | 47.0 | 52.0   | 60.0    | 48.0 |
| SSVM60 | 40.0     | 42.5 | 42.0 | 55.0 | 16.4 | 52.5 | 16.0 | 29.1   | 22.0    | 34.0 |
| SSVM80 | 42.0     | 60.0 | 38.0 | 53.0 | 16.4 | 52.5 | 25.0 | 36.4   | 42.0    | 42.0 |

Conclusions and Future Work

In this paper, we have developed an adversarial formulation for object localization and detection that treats uncertainty adversarially. This has the advantage of learning from a number of adaptively-selected training example annotations that must approximate the ground truth annotation. We demonstrated the benefits for object localization and detection using experiments over ten different object classes, showing significant improvements for our approach when tuned to the thresholded IOU evaluation measure.

In our experiments to date, we have used a pre-trained deep architecture to inform our predictions. Further optimization of the deep architecture based on our adversarial formulation for object detection and localization is important future work. Since our approach produces useful gradients despite being based on non-convex underlying performance measures (e.g., bounding box overlap), this can be accomplished using backpropagation. We also plan to more broadly apply our adversarial prediction game formulation to other object recognition and localization tasks.
References

Asif, K.; Xing, W.; Behpour, S.; and Ziebart, B. D. 2015. Adversarial cost-sensitive classification. In Proceedings of the Conference on UAI.

Blaschko, M. B., and Lampert, C. H. 2008. Learning to localize objects with structured output regression. 2–15. Springer.

Blaschko, M. B., and Lampert, C. H. 2008. Learning to localize on UAI.

Boyd, S., and Vandenberghe, L. 2004. Convex optimization. Cambridge university press.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. CoRR abs/1606.03657.

Cho, M.; Kwak, S.; Schmid, C.; and Ponce, J. 2015. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In Proceedings of the IEEE CVPR, 1201–1210.

Cinbis, R. G.; Verbeek, J.; and Schmid, C. 2014. Multi-fold mil training for weakly supervised object localization. In 2014 IEEE CVPR, 2409–2416. IEEE.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. Machine learning 20(3):273–297.

Crammer, K., and Singer, Y. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of machine learning research 2(Dec):265–292.

Dai, Q., and Hoiem, D. 2012. Learning to localize detected objects. In CVPR, 2012 IEEE Conference on, 3322–3329. IEEE.

Denton, E. L.; Chintala, S.; Szlam, A.; and Fergus, R. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. CoRR abs/1506.05751.

Dudík, M., and Schapire, R. E. 2006. Maximum entropy distribution estimation with generalized regularization. In Proc. Computational Learning Theory, 123–138.

Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; and Zisserman, A. 2005. Pascal visual object classes challenge results. Available from www.pascal-network.org.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence 32(9):1627–1645.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Advances in Neural Information Processing Systems, 2672–2680.

Grünewald, P. D., and Dawid, A. P. 2004. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. Annals of Statistics 32:1367–1433.

Kovashka, A.; Russakovsky, O.; Fei-Fei, L.; and Grauman, K. 2016. Crowdsourcing in Computer Vision. Foundation and Trends in Computer Graphics and Vision 10(3):177–243.

Larry Zitnick, P. D. 2014. Edge boxes: Locating object proposals from edges. In ECCV. European Conference on Computer Vision. McMahan, H. B.; Gordon, G. J.; and Blum, A. 2003. Planning in the presence of cost functions controlled by an adversary. In Proceedings of the International Conference on Machine Learning, 536–543.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Nowak, S., and Rüger, S. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In Proceedings of the international conference on Multimedia information retrieval, 557–566. ACM.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. IJCV 115(3):211–252.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013. A semi-automatic methodology for facial landmark annotation. In The IEEE Conference on CVPR Workshops.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. arXiv preprint arXiv:1606.03498.

Saragih, J. M.; Lucey, S.; and Cohn, J. F. 2011. Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision 91(2):200–215.

Simoniano, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Siva, P.; Russell, C.; Xiang, T.; and Agapito, L. 2013. Looking beyond the image: Unsupervised learning for object saliency and detection. In The IEEE CVPR.

Steinwart, I., and Christmann, A. 2008. Support vector machines. Springer Science & Business Media.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.

Szegedy, C.; Ioffe, S.; and Vanhoucke, V. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07663.

Tang, K.; Joulin, A.; Li, L.-J.; and Fei-Fei, L. 2014. Co-localization in real-world images. In 2014 IEEE CVPR, 1464–1471. IEEE.

Upchurch, P.; Sedra, D.; Mullen, A.; Hirsh, H.; and Bala, K. 2016. Interactive consensus agreement games for labeling images.

Vedaldi, A., and Lenc, K. 2015. Matconvnet: Convolutional neural networks for matlab. In Proceedings of the 23rd ACM international conference on Multimedia, 689–692. ACM.

Vedaldi, A. 2011. A MATLAB wrapper of SVMstruct. http://www.vilfeat.org/~vedaldi/code/svm-struct-matlab.

Wang, H.; Xing, W.; Asif, K.; and Ziebart, B. 2015. Adversarial prediction games for multivariate losses. In Advances in Neural Information Processing Systems, 2710–2718.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. J. 2010. The multidimensional wisdom of crowds. In Advances in neural information processing systems, 2424–2432.

Wu, Y., and Liu, Y. 2012. Robust truncated hinge loss support vector machines. Journal of the American Statistical Association.

Zhu, X., and Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. In CVPR, 2012 IEEE Conference on, 2879–2886. IEEE.