Adversarial Methods Improve Object Localization

Sima Behpour  Brian D. Ziebart
University of Illinois at Chicago
{sbehpo2,bziebart}@uic.edu

Abstract

We propose deep adversarial object localization, which approximates ground truth annotations of training images instead of approximating the loss function by posing object localization as an adversarial game between a loss-minimizing prediction player and a loss-maximizing adversarial player. We constrain the adversary to match specified properties of training data that are uncovered from a convolutional neural network’s feature representation. We demonstrate the efficiency and predictive performance of our approach on the ILSVRC2012 image dataset, showing significant improvements over other prediction methods.

1 Introduction

Performance in computer vision tasks is often assessed using specialized evaluation measures. For example, successful object detection is often defined by predicting a bounding box for an object that overlaps with more than some threshold (e.g., 50%, 70%) of the area of an object’s ground truth bounding box. Though deep learning architectures have drastically improved the proficiency of computer vision systems over recent years by constructing extremely rich and general feature representations for images [Szegedy et al., 2015, Simonyan and Zisserman, 2014a], there is still a significant mismatch between the objective functions they are designed to optimize and the application performance measures for which they will be applied. This degrades task performance in both theory [Liu, 2007] and in practice—as we show in this paper.

Adversarial methods have found success when used in generative formulations to train deep architectures and obtain useful feature representations [Goodfellow et al., 2014, Salimans et al., 2016, Chen et al., 2016, Mirza and Osindero, 2014, Denton et al., 2015] due to the additional robustness that considering an adversary introduces. We explore a similar hypothesis in this work: Does better aligning predictor training objectives with evaluation performance measures using adversarial prediction in conjunction with deep learning architectures provide advantages in object localization tasks?

2 Approach

Game formulation:  We view object localization using bounding box proposals (Figure 1) as a two-player game between a predictor player $\hat{Y}$ and an adversarial player $\hat{y}$ determining the evaluation distribution [Asif et al., 2015]. Player $\hat{Y}$ first chooses a predictive distribution of bounding boxes, $\hat{P}(\hat{y} | x)$, to minimize the expected loss, then player $\hat{Y}$ stochastically chooses an evaluation distribution, $\hat{P}(\hat{y} | x)$, that maximizes the expected loss while also (approximately) matching with a set of statistics, $\Phi(x, \hat{y})$, in expectation. We measure these statistics/features from labeled data and leverage the benefits of feature representations learned using a Convolutional Neural Network (CNN) [Vedaldi and Lenc, 2015] to define them in this paper.

Figure 1: Prediction game bounding boxes.

Workshop on Adversarial Training, NIPS 2016, Barcelona, Spain.
Definition 1. The Adversarial Object Localization (AOL) game is:
\[
\min_{\hat{P}} \max_{P} \mathbb{E}_{X \sim \hat{P}} \left[ \text{loss}(\hat{Y}, \hat{Y}) \right] \text{ such that: } \mathbb{E}_{X \sim \hat{P}} \left[ \phi(X, \hat{Y}) \right] = \mathbb{E}_{X, Y \sim \hat{P}} \left[ \phi(X, Y) \right]
\] (1)

\[
\iff \min_{\theta} \mathbb{E}_{X, Y \sim \hat{P}} \left[ \min_{P} \max_{\hat{P}} \mathbb{E}_{Y \mid X \sim \hat{P}} \left[ \text{loss}(\hat{Y}, Y) + \theta \cdot \phi(X, \hat{Y}) \right] X \right] - \theta \cdot \phi(X, Y). \right]
\] (2)

where \( \hat{P}(y \mid x) \) and \( P(y \mid x) \) are distributions over the \( |Y| \) predicted bounding boxes and \( \phi(\cdot, \cdot) \) are “features” characterizing relationships between the input pixels \( x \) and bounding box existence \( y \).

Due to strong Lagrangian duality, the dual solution (Equation (2)) is equivalent to the original formulation’s solution [Asif et al. 2015]. We consider two types of losses: the non-overlap, \( \text{loss}_{\text{non-overlap}}(\hat{y}, \hat{y}) = 1 - \text{area}(\hat{y} \cap \hat{y})/\text{area}(\hat{y} \cup \hat{y}) \); and the thresholded overlap, \( \text{loss}_{\text{overlap}}(\hat{y}, \hat{y}) = \text{area}(\hat{y} \cap \hat{y})/\text{area}(\hat{y} \cup \hat{y}) < \alpha \).

Table 1: The payoffs of the adversarial object localization game based on losses, \( \ell(\hat{y}, \hat{y}) \), between predicted \( (\hat{y}) \) and adversarial \( (\hat{y}) \) bounding boxes, and potential terms \( \psi(\hat{y}) = \theta \cdot \phi(\hat{y}) \).

<table>
<thead>
<tr>
<th>( \hat{y} )</th>
<th>( \hat{y} )</th>
<th>( \hat{y} )</th>
<th>( \hat{y} )</th>
<th>( \hat{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Efficient game solutions: We employ a constraint-generation method [McMahan et al. 2003; Wang et al. 2015] to more efficiently solve the AOL game. Its operation is detailed in Algorithm 1. It iteratively obtains a Nash equilibrium for a game defined over a subset of the bounding boxes, finds a player’s best response strategy (bounding box) to that equilibrium distribution, and then adds the best response to the set of strategies defining the game. When additional best responses no longer improve either player’s game value (Figure 2), the subgame equilibrium is guaranteed to be an equilibrium to the larger game [McMahan et al. 2003]. We employ the most probable strategy under this distribution at testing time.

**Algorithm 1** Adversarial Object Localization equilibrium computation

**Input:** Image \( \text{img} \); Parameters \( \theta \)

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

\begin{align*}
\text{BoxProposals} & \leftarrow \text{EdgeBox(img)} \\
\Phi & \leftarrow \text{YggNetLastLayer(img, BoxProposals)} \\
\psi & \leftarrow \theta \cdot \Phi \\
S & \leftarrow S \leftarrow \text{argmax } \psi
\end{align*}

**repeat**

\begin{align*}
(P, \hat{P}, Val) & \leftarrow \text{solveGame}(S, \hat{S}, \psi(S), \text{loss}(\hat{S}, \hat{S})) \\
(S_{\text{new}}, \text{maxV}) & \leftarrow \text{argmax } \mathbb{E}_{\hat{P}(S)} [\text{loss}(\hat{S}, \hat{S}) + \psi(\hat{S})]
\end{align*}

**if** \( (\text{Val} \neq \text{maxV}) \) **then**

\begin{align*}
\hat{S} & \leftarrow \hat{S} \cup S_{\text{new}}
\end{align*}

**end if**

\begin{align*}
(P, \hat{P}, Val) & \leftarrow \text{solveGame}(S, \hat{S}, \psi(\hat{S}), \text{loss}(\hat{S}, \hat{S})) \\
(S_{\text{new}}, \text{minV}) & \leftarrow \text{argmin } \mathbb{E}_{\hat{P}(S)} [\text{loss}(\hat{S}, \hat{S})]
\end{align*}

**if** \( (\text{Val} \neq \text{minV}) \) **then**

\begin{align*}
\hat{S} & \leftarrow \hat{S} \cup S_{\text{new}}
\end{align*}

**end if**

**until** \( \text{Val} = \text{maxV} = \text{Val} = \text{minV} \)

**return** \( (P, \hat{P}) \)

![Figure 2: Final strategies for predictor (black/blue bounding boxes) and adversary (black/red bounding boxes) with non-zero probability. Black boxes are the highest probability strategies.](image)

2
3 Experimental Evaluation

Experimental setup: We evaluate the effectiveness of our method in solving object localization tasks using 10 classes (Table 2) from the Imagenet2015 dataset [Russakovsky et al., 2015]. Since the images vary greatly in size, we first re-size all of the images to 1360 by 800 pixels, then apply EdgeBox [Larry Zitnick, 2014] to generate a relatively small set of candidate bounding boxes (up to 250) that cover the objects in the image. We represent bounding boxes for images using sets of 1000 CNN descriptors [Simonyan and Zisserman, 2014b, Vedaldi and Lenc, 2015] for each bounding box proposals provided by EdgeBox.

To show the relative performance of AOL, we benchmark it against a multiclass support vector machine (SVM) trained to incorporate the overlap into its hinge loss function, and multiclass logistic regression. We use an existing SSVM implementation [Vedaldi, 2011] for the former to learn and produce predictions. It employs constraint generation and 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, loss1−α, and two additional variants (denoted SSVM50 and SSVM70) using thresholded losses, loss<50% and loss<70%. For logistic regression, we estimate a distribution over all proposed bounding boxes that maximizes the conditional likelihood of proposed bounding boxes with an overlap of at least 50% with the example’s ground truth bounding box annotation. We produce bounding box predictions from the Bayesian optimal decision from the estimated conditional bounding box distribution for the non-overlap loss (LR) and for the thresholded overlap losses (LR50, LR70). Similarly, we train our AOL method for the non-overlap (AOL) and for the thresholded overlap (AOL50, AOL70) loss functions.

Evaluation results: We evaluate the performance of each approach using the overlap between predicted bounding box and ground truth bounding box. Figure 3 shows the amount of predicted bounding box overlap with the object’s ground truth bounding box across the entire set of examples for each method and object class. We note a few general trends. First, the two AOL localizers are either the best or competitive with the best for all of the datasets. Second, we note that the AOL localizer, the AOL50 localizer, or the AOL70 localizer provides the best performance for the amount of overlap and the thresholded overlaps (50% and 70%), with the exception of Sofa 50% and Sofa 70%.

Of more significance is how each method adapts to loss function modification. The adaptation between AOL and AOL50 and AOL70 is visually evident for many of these datasets: the black line

<table>
<thead>
<tr>
<th>#Class</th>
<th>#Training</th>
<th>#Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airplane</td>
<td>400</td>
<td>50</td>
</tr>
<tr>
<td>Bird</td>
<td>1600</td>
<td>200</td>
</tr>
<tr>
<td>Bus</td>
<td>330</td>
<td>50</td>
</tr>
<tr>
<td>Car</td>
<td>565</td>
<td>100</td>
</tr>
<tr>
<td>Cow</td>
<td>325</td>
<td>55</td>
</tr>
<tr>
<td>Dog</td>
<td>500</td>
<td>100</td>
</tr>
<tr>
<td>Horse</td>
<td>520</td>
<td>50</td>
</tr>
<tr>
<td>Monitor/TV</td>
<td>385</td>
<td>50</td>
</tr>
<tr>
<td>Sofa</td>
<td>380</td>
<td>50</td>
</tr>
</tbody>
</table>

Figure 3: The number of test examples (x axis) having at most a specified amount of overlap with the ground truth bounding box (y axis) for three methods: Adversarial object localization (AOL), logistic regression (LR), and structured support vector machines (SSVM) for eight classes from the Imagenet 2012 dataset. Larger values are better.
An unexpected result of these experiments is that in optimizing for the 50% and 70% threshold, work: that by better aligning localizer training objectives with performance measures of interest, (AOL)

This research was supported as part of the Future of Life Institute (futureoflife.org) FLI-RFP-AI1

References


