Adversarial Localization Network

Lijie Fan, Shengjia Zhao, Stefano Ermon

flj14@mails.tsinghua.edu.cn, sjzhao@stanford.edu, ermon@stanford.edu

Introduction

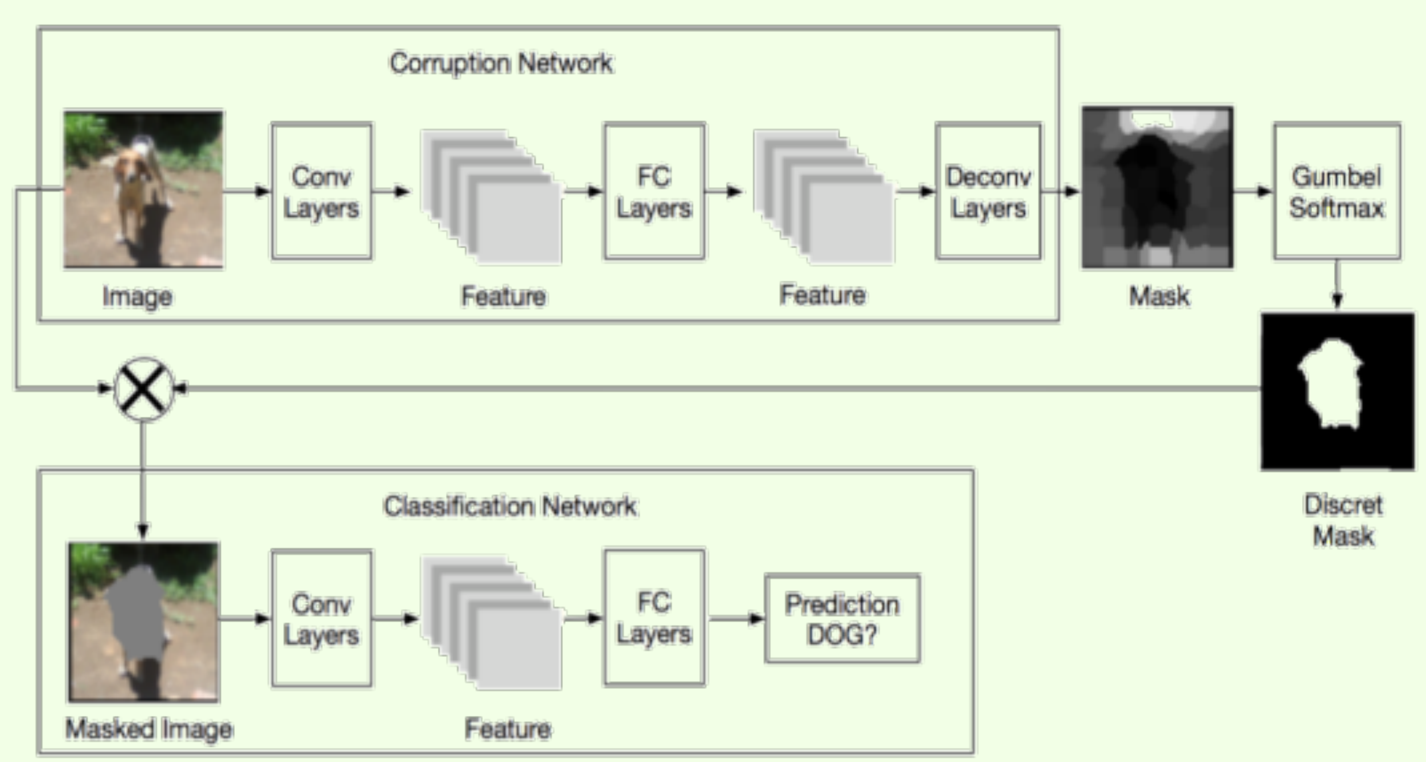
Problem
- Hard to obtain localization and segmentation annotations
- Develop weakly supervised approaches
- Localize object with image-level labels only

Previous Work
- Perturb image regions, eg. by occlusion
- Maximally decreases prediction confidence of a classifier
- Assume object region important to the decision

Limitations
- Classifier vulnerability to adversarial noise
- Partition the space into very coarse grids to alleviate
- Lead to coarse grained and inaccurate localizations

Our work
- Inspired by VAT, perform adversarial training
- Making classifier more robust against adversarial noise
- Utilize super-pixels, lead to better object boundary

Architecture



- **Corruption Network**
  Produce *saliency masks* from original image
- **Classification Network**
  Produce *classification scores* of masked image
- **Adversarial Training**
  Robust against adversarial noise

Approach

Corruption Network
- Capture both *global* and *local* information
- **Global Information**
  Deep Conv-Deconv neural network architecture
- **Local Information**
  Residual connections between intermediate layers

Probabilities to Masks
- **Super-pixel Representation**
  - object boundaries align with image edges
  - masking decision for each super-pixel as a whole
  - alleviates the vulnerability to adversarial perturbations
- **Probability Discretize**
  - Gumbel Trick
  - Sample: $u_1, u_2 \sim \mathrm{Uniform}(0,1)$, $z_i = -ln(-ln(-u_i)), i = 1, 2$
  - Draw sample from: $x = \mathbb{I}(p + z_1 > 1 - p + z_2)$
- **Back-propagate**
  - Gradients approximated by the straight-through estimator

Adversarial Training Procedure
- **Corruption Loss**
  Encourage classifier for same logit score for each class

$$\mathrm{Loss}_{\mathrm{corruption}}(x^i) = \sum_{j=1}^{K} y_j l_j - \frac{1}{K-1}(1-y_j) l_j$$

- **Classification Loss**
  Softmax cross-entropy loss

$$\mathrm{Loss}_{\mathrm{classifier}}(x^i) = \sum_{j=1}^{K} l_j log(y_j)$$

- **Penalization**
  Penalize the total area of the generated masks
  Force to generate the minimum confusing mask

Conclusion
- A novel weakly supervised approach for object localization
- Apply adversarial training to avoid vulnerability
- Utilize super-pixels, which lead to:
  1. Better object boundary
  2. Alleviate vulnerability to adversarial perturbations

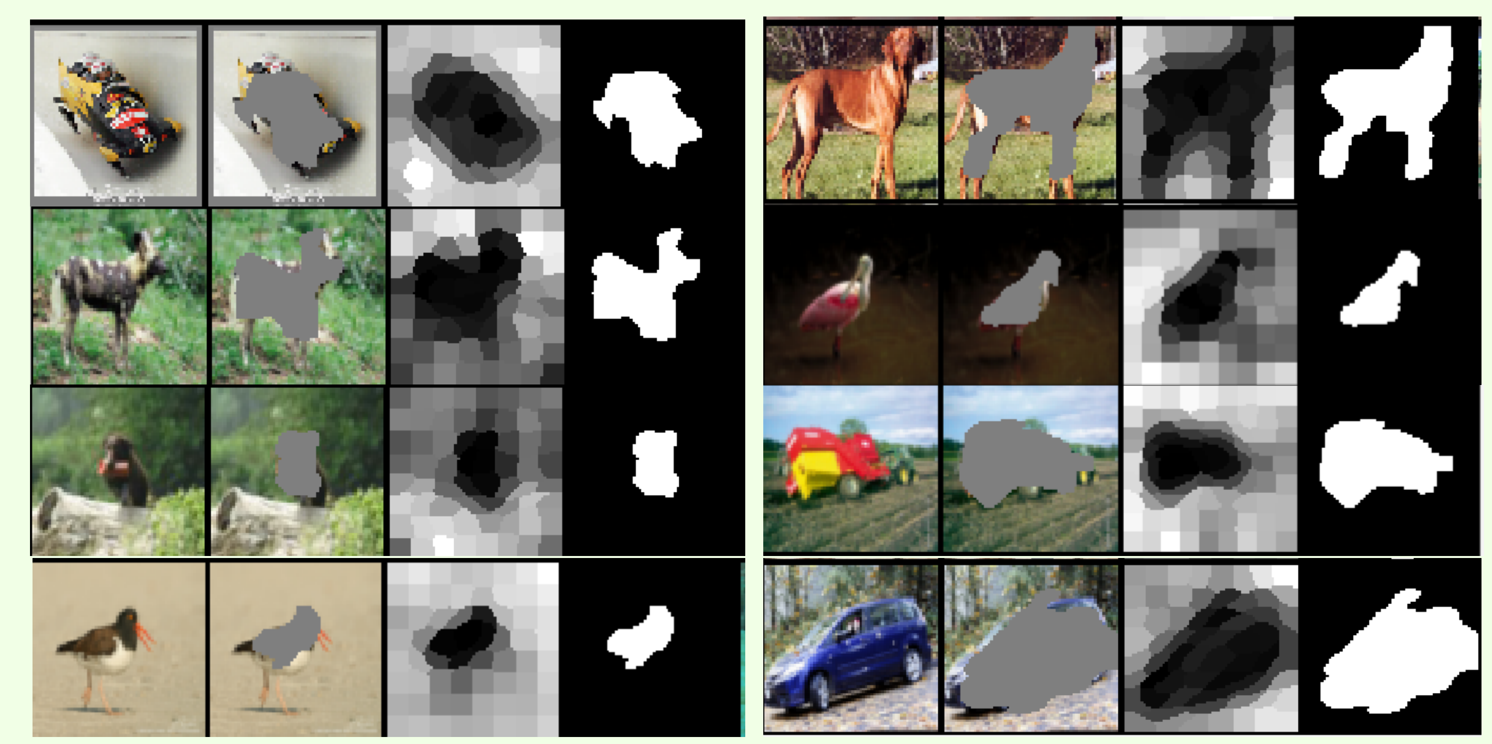
Experiments

Evaluation Metric
- **Metric 1**: Top-1 Accuracy
  - more than 50% IoU between prediction and ground truth
- **Metric 2**: Top-1 Accuracy with Classification
  - more than 50% IoU between prediction and ground truth
  - Image-level class predict correctly

Localization Performance
- Small 64×64 input image
- simple 4-layer convolutional network
- 0.093 seconds per image
- Compare with baseline models:

Methods	Metric 1	Metric 2
Max Image Box	41.0%	34.3%
Backprop	--	38.7%
Layer-wise Relevance Propagation	42.2%	--
Global Average Pooling	51.9%	43.6%
Adversarial Localization Network	**56.5%**	**45.5%**

- Sample Masks



Adversarial Necessity
- Get adversarial artifacts when freeze classifier
- Trainable layers increase, adversarial artifacts decrease
- Adversarial Examples: