Adversarial Framing for Image and Video Classification

Background

In general, such attacks deteriorate the quality of the input by either slightly modifying most of its pixels, or by occluding it with a patch

Goal

- Keeping the whole content unchanged
- Simply add a thin border around the original input
- Untargeted attacks
- Targeted attacks

Method

- For fixed-size images and videos:

1.Every example is surrounded with the same framing

2.The classification loss is backpropagated

3.The framing is modified using its gradients to maximize the loss.

(Consider a white-box setting)

$$Lossfunction = \frac{1}{|D|}\sum_i \log(f_{y_i}(\hat{x}_i))$$

$D - dataset$

$f - classification$

Algorithm

Algorithm 1 Training of the adversarial framing

input: Dataset $\mathcal{D} = \{(x_i, y_i)\}$, $x_i \in [0,1]^{h \times w \times 3}$, classifier f, framing's width W
output: Universal adversarial framing θ
Initialize $\hat{\theta} \sim \mathcal{N}(0,1)$, of size $2W(h + w + 2W)$
repeat
    for each datapoint $(x_i, y_i) \in \mathcal{D}$ **do**
        $\hat{x}_i \leftarrow x_i$ surrounded by $\theta := Sigmoid(\hat{\theta})$
    end for
    update $\hat{\theta}$ to minimize $\frac{1}{|\mathcal{D}|} \sum_i \log(f_{y_i}(\hat{x}_i))$
until convergence

untargeted attacks

- Goal:Train the AF adversarial framework to make the classifier misclassify

- Method:Backpropagation for the frame gradient direction

untargeted attacks

- Goal:Train the AF confrontation framework to mislead the classifier to the target class

- Method:Maximize the score of the randomly selected target class and report the success rate of misclassification into the target class

Result analysis

Figure 1: Examples from ImageNet with adversarial framing of width 3. Most of the images are wrongly classified as a maypole. We hypothesize that the colorfulness of that class makes it especially easy for **AF** to resemble it.

Result analysis

1.Introducing uniformly distributed random noise (RF) and black pixels (BF) and AF confrontation framework for horizontal comparison. Both RF and BF can make the classification accuracy of the classifier slightly lower, but AF can significantly reduce the classification accuracy of the classifier

2.For the AF confrontation frame, the larger the W (width), the better the effect.

3.When W=4, the success rate of misleading the target category is extremely high, with pictures up to 99.98% and videos up to 99.78%

Attack	$W=1$	$W=2$	$W=3$	$W=4$
None	76.13%			
RF	70.13%	67.63%	68.36%	67.25%
BF	72.99%	72.9%	72.39%	72.34%
AF	10.53%	0.44%	0.11%	0.1%

(a) ImageNet dataset

Attack	$W=1$	$W=2$	$W=3$	$W=4$
None	85.95%			
RF	82.57%	80.53%	81.11%	79.74%
BF	84.94%	84.73%	84.75%	84.59%
AF	65.77%	22.12%	9.45%	2.05%

(b) UCF101 dataset

Table 1: Accuracies of the classifiers (full validation set) for various values of the framing width W.

	min	**avg**	**max**
AF, $W=4$	99.15%	99.66%	99.98%

(a) ImageNet dataset

	min	**avg**	**max**
AF, $W=4$	73.04%	89.63%	99.78%

(b) UCF101 dataset

Table 2: Success rate of targeted attacks (the higher the better) with adversarial framing of width 4. Minimum, average and maximum values are taken across 8 different targets.

Grad-CAM

- GradCAM(Selvarajuetal.2017)is a method for producing visual explanations for a convolutional neuralnet work's predictions
- For a given classifier f,input x and a class c,it computes a heatmap visualizing how much particular regions of x contribute to a score of the class c output by f.

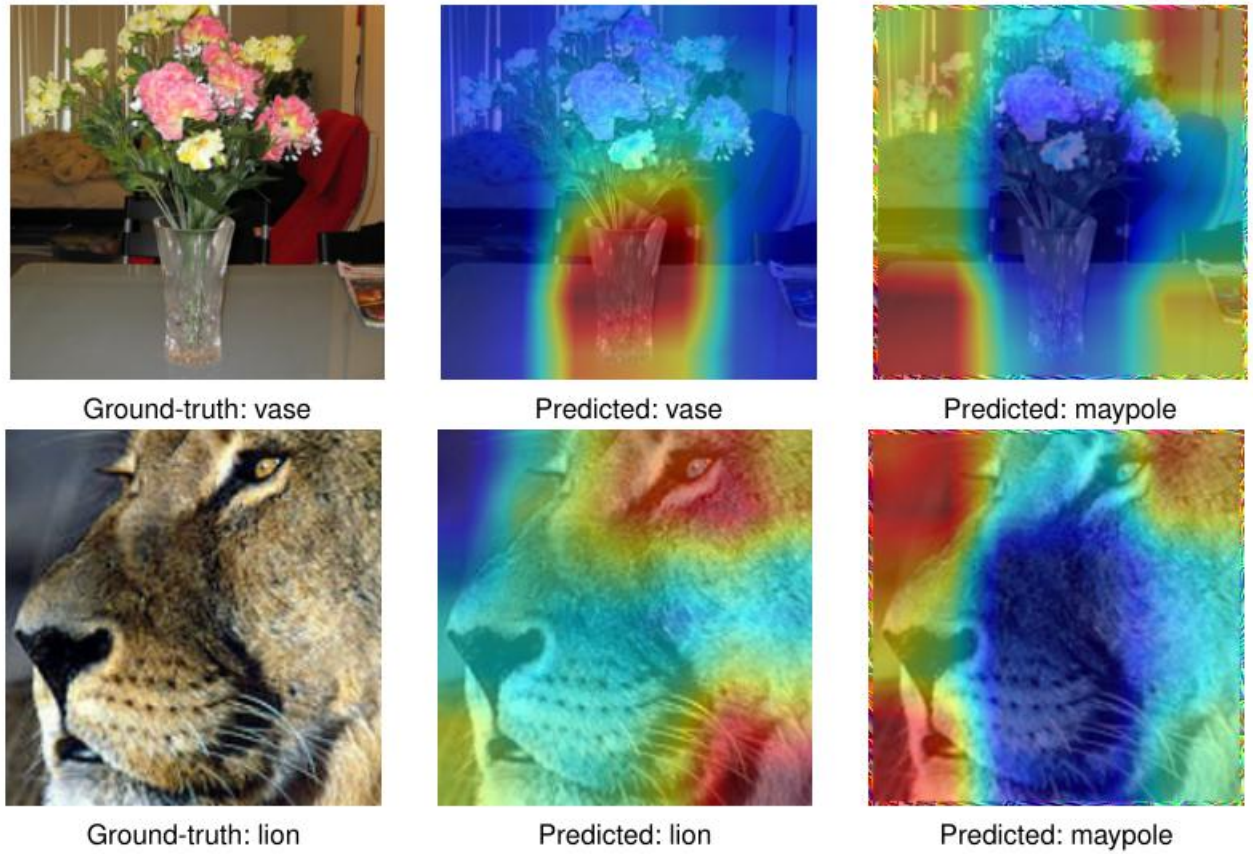

Figure 2: Grad-CAM for inputs from ImageNet. For each example, first the original image is shown, and then the visualizations for unattacked and attacked image. While the network correctly identifies key objects for classification in unattacked images, it concentrates on the image borders when given adversarial input.

Conclusion

Advantage

- The method does not modify the original content of the input
- only adds a small border to surround it
- The proposed attack is universal (i.e. the same adversarial framing can be applied in different images or videos)
- While the network correctly identifies key objects for classification in unattacked images, it concentrates on the image borders when given adversarial input

Disadvantage

- Although the increase in the border has an obvious effect on the classifier, it will be recognized if there is manual sampling recognition

Application scenario:

- Audio and video transmission that avoids sensitive detection (fidelity)
- Human machine recognition