Adversarial Framing for Image and Video Classification

Michał Zając,[*1,2] **Konrad Żołna,**[*1,3] **Negar Rostamzadeh,**[3] **Pedro O. Pinheiro**[3]
[1]Jagiellonian University, Kraków, Poland[†]
[2]Nomagic, Warsaw, Poland
[3]Element AI, Montréal, Canada
emzajac@gmail.com, konrad.zolna@gmail.com, negar@elementai.com, pedro@opinheiro.com

Abstract

Neural networks are prone to adversarial attacks. 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. In this paper, we propose a method that keeps the image unchanged and only adds an *adversarial framing* on the border of the image. We show empirically that our method is able to successfully attack state-of-the-art methods on both image and video classification problems. Notably, the proposed method results in a universal attack which is very fast at test time. Source code can be found at `github.com/zajaczajac/adv_framing`.

Introduction

It is a well-known fact that one can change the output of a neural network-based classifier by applying a small perturbation to its input (Szegedy et al. 2014). Such perturbations are a base for adversarial attacks, which we divide into two categories: *fully-affecting* and *partially-affecting*.

- Fully-affecting attacks generate small pixel intensity modifications which are optimized to be hardly visible for humans. These attacks typically have their ℓ_2 or ℓ_∞ norm constrained (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Carlini and Wagner 2017) and hence affect the whole image.
- Partially-affecting attacks usually have their ℓ_0 norm constrained. They introduce perceptible but small occlusion to the image, such as a patch (Brown et al. 2017; Karmon, Zoran, and Goldberg 2018) or a single pixel (Su, Vargas, and Sakurai 2017).

The attacks mentioned above either slightly modify all the pixels of the image or occlude parts of it. However, the attackers may find this to be a serious limitation and seek for new types of attacks. For instance, consider a scenario where they upload videos containing forbidden content, such as violence or pornography. The goal is to bypass video-sharing website's filters. At the same time, the perturbations introduced should not be distracting and all information should be retained.

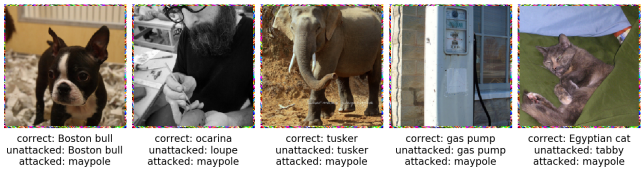

Figure 1: Examples from ImageNet with adversarial framing of width 3. Most of the images are wrongly classified as a maypole. We hypothesize that the colorfulness of that class makes it especially easy for **AF** to resemble it.

In this paper, we propose a new attack which is well-suited for the above-mentioned purposes. The method, dubbed adversarial framing (**AF**), consists of simply adding a thin border around the original image, keeping the whole content of the image unchanged (see Figure 1 for some qualitative results). Our attack is universal (Moosavi-Dezfooli et al. 2017), which means the same **AF** is applied to all inputs. Our method only requires substantial computing during the training procedure. At test time, the only extra computation required is the appending of the precomputed framing to the input.

In this work, we consider a white-box setting, in which an access to the architecture and weights of the trained classifier is given. Previous work has shown that if only black-box access is given, a surrogate model can be leveraged to obtain an attack that transfers well to the original model (Papernot, McDaniel, and Goodfellow 2016). Therefore, a white-box model is a realistic assumption and, in fact, is the most commonly considered paradigm in the literature.

Although extensive literature exists on attacks against image classifiers, we are aware of only a few works on video classifier attacks (Wei, Zhu, and Su 2018; Li et al. 2018; Rey-de Castro and Rabitz 2018). While resulting in successful attacks, these approaches are fully-affecting and hence introduce adversarial artifacts in the video. In contrast, output from our attack contains the original video and no information is lost. Moreover, the framing is constant over all video frames, removing any "flickering" effect that could potentially be distracting to viewers[1].

[*]Equal contribution
[†]ul. Łojasiewicza 6, 30-348 Kraków, Poland, +48 12 664 66 29

[1]See `youtu.be/PrU9R6eFNTs` for some video attacks.

	$W=1$	$W=2$	$W=3$	$W=4$
Unattacked	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%

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

	$W=1$	$W=2$	$W=3$	$W=4$
Unattacked	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%

Table 2: Accuracies of the UCF101 classifier (full validation set) for various values of the framing width W.

Method

Suppose we are given a differentiable classifier f for some image or video classification problem. The attack procedure is simple and consists in adding the precomputed **AF** around a given image. The framing's width W is a tunable hyperparameter.

Note that the input size is modified due to the addition of the framing. This does not pose any issue to the CNN-based classifier, as most modern architectures (such as ResNet or ResNeXt) accept various input sizes. If the classifier's input size is fixed, the proposed algorithm can be simply modified so that the image is resized before applying adversarial framing.

During training, we optimize **AF** to minimize the score assigned by f to the ground-true class (see Algorithm 1).

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

Experiments

ImageNet On ImageNet, we performed untargeted attacks against pretrained ResNet-50 from PyTorch Model Zoo.

We compare our **AF** to two simple baselines. One applies uniformly distributed random noise (**RF**) and another black pixels only (**BF**). Results are reported in Table 1.

UCF101 UCF101 is a dataset containing realistic videos. Each video contains a person performing some action, out of 101 possible classes.

We tested our method by performing an untargeted attack on a state-of-the-art method, ResNeXt-101 based spatio-temporal 3D CNN – we used model pretrained by (Hara, Kataoka, and Satoh 2018). This model takes clips as input, each containing 16 consecutive frames. Results of the experiments are reported in Table 2.

Conclusion

In this work, we present a simple method for attacking both image and video classifiers. The proposed attack is universal (*i.e.* the same adversarial framing can be applied in different images or videos), efficient and effective. Moreover, our method does not modify the original content of the input and only adds a small border to surround it.

Acknowledgments

Konrad Żołna is financially supported by National Science Centre, Poland (2017/27/N/ST6/00828). Michał Zając is co-financed by National Centre for Research and Development as a part of EU supported Smart Growth Operational Programme 2014-2020 (POIR.01.01.01-00-0392/17-00).

References

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *CoRR* abs/1712.09665.

Carlini, N., and Wagner, D. A. 2017. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy*.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*.

Karmon, D.; Zoran, D.; and Goldberg, Y. 2018. LaVAN: Localized and visible adversarial noise. In *ICML*.

Li, S.; Neupane, A.; Paul, S.; Song, C.; Krishnamurthy, S. V.; Roy-Chowdhury, A. K.; and Swami, A. 2018. Adversarial perturbations against real-time video classification systems. *CoRR* abs/1807.00458.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. *CVPR*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*.

Papernot, N.; McDaniel, P. D.; and Goodfellow, I. J. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR* abs/1605.07277.

Rey-de Castro, R., and Rabitz, H. 2018. Targeted nonlinear adversarial perturbations in images and videos. *CoRR* abs/1809.00958.

Su, J.; Vargas, D. V.; and Sakurai, K. 2017. One pixel attack for fooling deep neural networks. *CoRR* abs/1710.08864.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR*.

Wei, X.; Zhu, J.; and Su, H. 2018. Sparse adversarial perturbations for videos. *CoRR* abs/1803.02536.