ADVERSARIAL EXAMPLES: GENERATING AND EVALUATING AN IMAGE-INDEPENDENT PATCH

Matthew Baron
mcbaron@andrew.cmu.edu

James Gianoglio
jgianogl@andrew.cmu.edu

Anna Belova
abelova@andrew.cmu.edu

We plan to to generate and evaluate an image-independent patch—an adversarial sticker—that is robust to affine transformation and universal across selected image recognition models. To this end, we will train specific pixels of input images to represent the adversarial sticker pattern, such that each selected image recognition model produces a classification of our choice. We will then evaluate the effectiveness of this sticker across image classes and selected image recognition models. The experiments will be conducted using three models designed for the the ImageNet task: VGG, GoogLeNet, AlexNet.

1 INTRODUCTION

Deep learning concepts have generated applications in a wide variety of arenas ranging from image and speech recognition, consumer data analytics, autonomous transportation to natural language processing and health care. Despite this tremendous popularity, researchers increasingly recognize that deep learning systems are subject to vulnerabilities that can undermine public safety, security, and privacy (Papernot et al., 2016b). The generation and study of adversarial examples is designed to help identify threats, attacks, and defenses of systems built on deep learning principles. Work on adversarial examples is fundamental to exposing neural network vulnerabilities, addressing issues of security, and contributing to more robust approaches in neural network design. Adversarial examples are carefully constructed inputs that are designed to cause misclassifications by a neural network. In the area of image recognition, researchers have studied adversarial examples that only slightly perturb inputs of correctly classified examples (Papernot et al., 2016a) as well as adversarial examples that combine correctly classified images with the image-independent patches that bare little or no resemblance to correctly classified images but can disrupt a neural network (Brown et al., 2017). Research on adversarial examples that rely on image-independent patches is particularly important, given that an interference relying on an adversarial patch (e.g., in the form of a sticker) can target an arbitrary image and be agnostic about some aspects of the physical environment, such as lighting conditions, in which the attack may take place (Brown et al., 2017).

The aim of the proposed project is to generate and evaluate an image-independent patch—an adversarial sticker—that is robust to affine transformation and universal across image recognition models. To this end, we will train specific pixels of input images to represent the pattern of our adversarial sticker, such that each selected image recognition model produces a classification of our choice. We will then evaluate the effectiveness of this sticker across image classes and selected image recognition models.

2 INITIAL EXPERIMENTS

For this project, we will generate and evaluate an image sticker for three image recognition models designed for the ImageNet task (Deng et al., 2009). These models include: VGG (Simonyan & Zisserman, 2014), GoogLeNet (Szegedy et al., 2014), and AlexNet (Krizhevsky et al., 2012).

2.1 Adversarial Sticker Generation

We will use a computer vision algorithm to select pixels in input images that all lie roughly on a single plane. We will then create an input-buffer layer, in which these pixels will be masked. This layer will be prepended to the pretrained image model; the weights of the buffer layer will then be optimized to achieve our desired image classification. This approach will impose affine invariance on the adversarial sticker.

2.2 Adversarial Sticker Evaluation

For the initial experiments we will work with VGG as our primary pretrained model and carry out the following assessments:

1. Compute the VGG-based adversarial sticker effectiveness—the percent of images for which an image model generates a classification of our choice[1]—with respect to VGG. The set-aside test collection of images will be a stratified sample from the ImageNet collection, such that each category is represented equally well.[2]
2. Compute the effectiveness of naive approaches, such as random noise applied to the masked pixels or placing a scaled version of an example from the desired class. Statistically test incremental effectiveness of the adversarial patch relative to the naive baseline(s).
3. Compute the adversarial sticker effectiveness with respect to the other two pretrained models (i.e., GoogLeNet and AlexNet).

3 Proof of Concept

The full proof of concept will involve applying our adversarial sticker generation methods using pretrained GoogLeNet and pretrained AlexNet, and carrying out evaluation steps described in Section 2.2. Separately for each model and for the three models combined, we will report a confusion matrix with the types of classes that our sticker best obscures and the classes that our patch least obscures for each experiment.

4 Final Experiments

[TBD]

5 Final Goals & Evaluation

We have developed four related objectives for our project:

1. Implement adversarial sticker generation using the VGG model.
2. Evaluate incremental effectiveness of the adversarial sticker for VGG, relative to at least one naive sticker generation approach.
3. Evaluate the incremental effectiveness of the VGG-based adversarial sticker for GoogLeNet and AlexNet.
4. Implement the adversarial sticker generation and evaluation (including cross-model evaluation) for GoogLeNet and AlexNet.

Correspondingly, we propose the following evaluation criteria:

- Completion of Items 1 and 2 would mean satisfactory completion of this project;

[1] We will exclude images that contain our classification of choice as either true classification, or as predicted classification.

[2] Additional strata may be added, to address expected variations of the effectiveness of our method, such as whether the original VGG classification was a correct one.

- Completion of work under Item 3 (i.e., cross-model evaluation of the VGG-based sticker), in addition to Items 1 and 2 would mean project success.
- Completion of work under Item 4 (i.e., adversarial sticker generation and evaluation for the three selected models) is our stretch-objective.

6 Related Work

We draw on prior work in adversarial generation for image recognition networks (Goodfellow et al., 2014). Much of the theory behind adversarial generation is rooted in the fact that the volume of the space of images is massive, and incredibly sparse. For example, the simple task of digit recognition as defined in the MNIST dataset (LeCun & Cortes, 2010), yields an input space of 10^{1888} possible 28x28 pixel black and white input images. Since the input space is very large, image recognition algorithms must partition it in some manner. Adversarial image generation approaches take advantage of the way in which the partition has been defined in image recognition algorithms.

Yuan et al. (2017) survey a variety of methodologies for generating attacks, review the effectiveness of different taxonomies of attacks, and explore some countermeasures and challenges. Many of the attack vectors discussed target the representational capacities of the networks. As our approach aims to be network agnostic, we target the input of the network. Nguyen et al. (2015) generate human unrecognizable images that deep networks classify with very high confidence as belonging to a specific class. Since we seek to generate a perturbation of the input image, simply replacing the input will not suffice. Brown et al. (2017) employ a procedure to generate an adversarial patch on the image that lies within the plane of the pixels of the image. We seek a patch that is robust to affine transformation, meaning it will sit on a surface within the image.

7 Data & Technical Requirements

We will use ImageNet (Deng et al., 2009), given its universal use by the target classifiers. We anticipate using a combination of Tensorflow and Keras, as well as Pytorch and Inferno for various training and testing procedures. We'll use Tensorboard for visualizing results.

For training of the images, we will rely on AWS EC3 instances. We will also deploy Docker files, which we will make available for the reproducibility of this research.

References

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, USA, 2012. Curran Associates Inc. URL `http://dl.acm.org/citation.cfm?id=2999134.2999257`.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pp. 372–387. IEEE, 2016a.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016b.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL `http://arxiv.org/abs/1409.1556`.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. URL `http://arxiv.org/abs/1409.4842`.

Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *arXiv preprint arXiv:1712.07107*, 2017.