Targeted Universal Adversarial Patch Attacks on Convolutional Neural Networks

Nishanth Adithya Chandramouli
Department of Robotics Engineering
Worcester Polytechnic Institute
nchandramouli@wpi.edu

***Abstract*—Convolutional Neural Networks (CNNs) have demonstrated remarkable performance on image classification tasks; however, they are known to be vulnerable to adversarial attacks. In this work, we investigate targeted universal adversarial patch attacks against a ResNet-18 classifier trained on the CIFAR-10 dataset. Unlike pixel-level perturbations, adversarial patches are spatially localized and more realistic for physical-world deployment. Using Expectation Over Transformation (EoT), we train a single universal patch that successfully forces the classifier to predict a chosen target class across diverse inputs and transformations. Experimental results show a targeted attack success rate of 63.55% while reducing classification accuracy from 86.23% to 35.95%, demonstrating the effectiveness and robustness of the proposed attack.**

***Index Terms*—Adversarial Examples, Adversarial Patch, Deep Learning Security, Computer Vision, CNN Robustness**

I. Introduction

Deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved state-of-the-art performance in computer vision tasks. Despite their success, these models are highly susceptible to adversarial attacks—small, carefully crafted perturbations that cause incorrect predictions.

Adversarial patch attacks represent a particularly dangerous threat model, as they are spatially localized, input-agnostic, and can be physically realized as stickers or patterns. In this project, we study targeted universal adversarial patches that cause a classifier to consistently predict a chosen target class regardless of the true input.

II. Related Work

Early work on adversarial examples focused on imperceptible pixel-level perturbations. Brown et al. introduced adversarial patches as a universal, localized attack robust to transformations and real-world conditions. Expectation Over Transformation (EoT) was proposed to ensure robustness against spatial and photometric variations.

III. Mathematical Formulation

A. Classifier Model

Let

$$f_\theta : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^K$$

denote a convolutional neural network with parameters θ, where $K = 10$ for CIFAR-10. The network outputs logits

$$\mathbf{z} = f_\theta(x),$$

which are converted into class probabilities using the softmax function:

$$p_\theta(y = k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}.$$

B. Adversarial Patch

The adversarial patch is a learnable tensor

$$P \in [0,1]^{3\times s\times s},$$

where s is the patch size. The pixel constraint is enforced by projection:

$$P \leftarrow \Pi_{[0,1]}(P).$$

C. Patch Application Operator

Given an image $x \in [0,1]^{3\times H\times W}$ and a patch location $\ell = (u, v)$, the patched image is defined as:

$$A(x, P, \ell)_{:,i,j} = \begin{cases} P_{:,i-u,j-v}, & (i,j) \in \Omega_\ell \\ x_{:,i,j}, & \text{otherwise,} \end{cases}$$

where Ω_ℓ denotes the spatial region covered by the patch.

D. Targeted Loss

For a target class y_t, the targeted cross-entropy loss is:

$$\mathcal{L}_{\text{CE}}(x, P, \ell) = -\log p_\theta\left(y_t \mid A(x, P, \ell)\right).$$

E. Expectation Over Transformation

To ensure robustness, the patch is optimized under random transformations $t \sim \mathcal{T}$ (rotation, scaling, brightness). The optimization objective becomes:

$$\min_P \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{t\sim\mathcal{T}}\mathbb{E}_{\ell\sim\mathcal{L}}\left[\mathcal{L}_{\text{CE}}(x, t(P), \ell)\right].$$

F. Total Variation Regularization

To encourage spatial smoothness, a Total Variation (TV) regularizer is added:

$$\text{TV}(P) = \sum_{c,i,j} |P_{c,i,j} - P_{c,i,j+1}| + |P_{c,i,j} - P_{c,i+1,j}|.$$

The final objective is:

$$\min_P \mathbb{E}_{x,t,\ell}\left[\mathcal{L}_{\text{CE}}\right] + \lambda\,\text{TV}(P).$$

Algorithm 1 Targeted Universal Adversarial Patch Optimization

Require: Classifier f_θ, training dataset $\mathcal{D} = \{(x_i, y_i)\}$, target class y_t, learning rate η, number of epochs E
Ensure: Targeted universal adversarial patch P

Initialize patch $P \sim \mathcal{U}(0, 1)$
for $e = 1$ to E **do**
  for each mini-batch $\{x_b\}_{b=1}^{B}$ **do**
    Sample transformations $\{t_b\}_{b=1}^{B} \sim \mathcal{T}$
    Sample patch locations $\{\ell_b\}_{b=1}^{B} \sim \mathcal{L}$
    for $b = 1$ to B **do**
      $\tilde{P}_b \leftarrow t_b(P)$ ▷ Apply EoT transformation
      $x_b^{\text{patch}} \leftarrow A(x_b, \tilde{P}_b, \ell_b)$
    end for
    Compute loss:

$$\mathcal{L} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \ell\big(f_\theta(x_b^{\text{patch}}), y_t\big)$$

    Update patch via gradient descent:

$$P \leftarrow P - \eta \nabla_P \mathcal{L}$$

    Project patch values:

$$P \leftarrow \Pi_{[0,1]}(P)$$

  end for
end for
return P

IV. Optimization Algorithm

V. Experimental Setup

A. Dataset

All experiments are conducted on the CIFAR-10 dataset, which consists of 60,000 color images of size 32×32 across 10 object classes. The dataset is split into 50,000 training images and 10,000 test images. Standard normalization is applied using dataset-specific mean and variance statistics.

B. Classifier Training

The baseline classifier is a ResNet-18 architecture adapted for CIFAR-10 by replacing the initial 7×7 convolution with a 3×3 kernel and removing max-pooling. The model is trained using cross-entropy loss and the Adam optimizer for 10 epochs, achieving a test accuracy of 86.23%.

C. Patch Training Configuration

The adversarial patch is trained using mini-batch stochastic gradient descent with the Adam optimizer. The patch size is fixed at 16×16, corresponding to 25% of the image area. During training, the patch is randomly placed at different spatial locations and subjected to random transformations using Expectation Over Transformation (EoT). The training process consists of 12 epochs with 400 optimization steps per epoch.

VI. Evaluation Metrics

To comprehensively evaluate the effectiveness of the adversarial patch, we report multiple metrics capturing both targeted and untargeted attack behavior.

A. Clean Accuracy

Clean accuracy measures the classification accuracy of the model on unmodified test images:

$$\text{Acc}_{\text{clean}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big(\arg\max f_\theta(x_i) = y_i\big).$$

B. Patched Accuracy

Patched accuracy quantifies the degradation in model performance when the adversarial patch is applied:

$$\text{Acc}_{\text{patched}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big(\arg\max f_\theta(A(x_i, P)) = y_i\big).$$

C. Targeted Attack Success Rate

The targeted success rate (TSR) measures how often the classifier predicts the target class:

$$\text{TSR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big(\arg\max f_\theta(A(x_i, P)) = y_t\big).$$

VII. Ablation Discussion

A. Effect of Patch Size

Larger patches increase attack success but reduce visual stealth. The chosen 16×16 patch balances effectiveness and realism, aligning with prior work on physically realizable attacks.

B. Role of Expectation Over Transformation

Without EoT, patches overfit to specific placements and transformations, leading to poor generalization. Incorporating EoT significantly improves robustness under translation, rotation, and brightness variation.

C. Impact of Regularization

Total Variation regularization stabilizes training and reduces high-frequency noise in the patch. Empirically, this improves convergence and visual smoothness without significantly affecting attack success.

VIII. Threat Model and Security Implications

The attack assumes white-box access to the classifier during patch optimization. However, once trained, the patch operates in a black-box setting, requiring no further access to model internals. This makes adversarial patches particularly concerning from a security standpoint.

In real-world scenarios, adversarial patches could be deployed as stickers or printed patterns, potentially compromising vision-based systems such as surveillance cameras, autonomous vehicles, and robotics platforms.

TABLE I: Classification Performance

Metric	Value (%)
Clean Accuracy	86.23
Accuracy with Patch	35.95
Targeted Attack Success	63.55

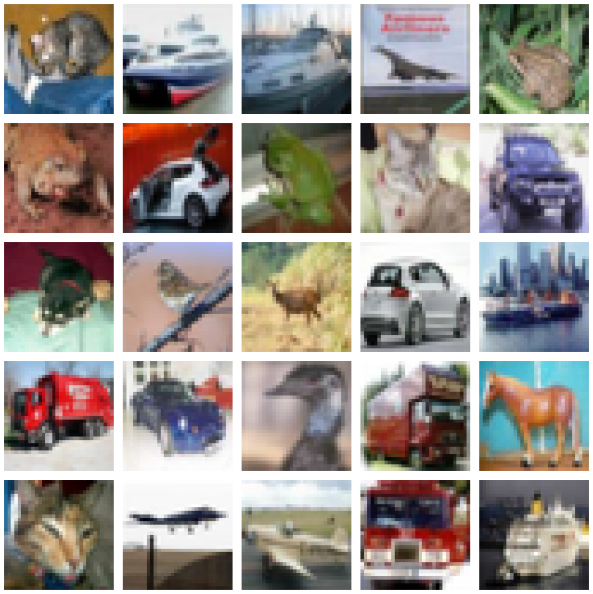

Fig. 1: Sample clean CIFAR-10 images.

IX. Results

A. Quantitative Evaluation

B. Qualitative Results

C. Confusion Matrix Analysis

The patched confusion matrix shows a collapse toward the target class, indicating successful targeted manipulation.

D. Attack Dynamics

X. Discussion

The results demonstrate that a small universal adversarial patch can reliably dominate CNN predictions. EoT significantly improves robustness, while the moderate patch size maintains realism. Although not achieving perfect success, the attack generalizes well across classes and transformations.

XI. Conclusion

This work demonstrates the vulnerability of modern CNNs to targeted universal adversarial patches. A single learned patch reduces classification accuracy by over 50% and forces targeted misclassification in over 63% of test samples. These findings highlight the importance of developing robust defenses against patch-based attacks.

Future work includes physical-world evaluation and defense mechanisms such as adversarial training and patch detection.

Fig. 2: Same images with adversarial patch applied.

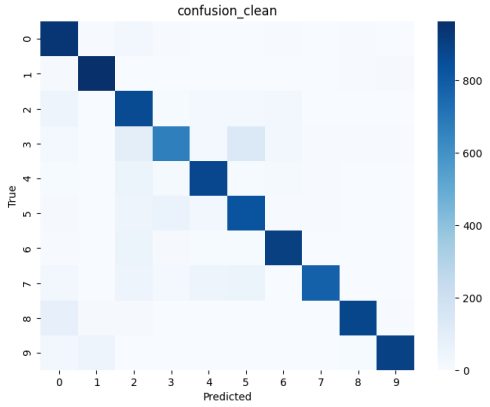

Fig. 3: Confusion matrix for clean classifier.

References

[1] T. Brown et al., "Adversarial Patch," *arXiv preprint arXiv:1712.09665*, 2017.

[2] A. Athalye et al., "Synthesizing Robust Adversarial Examples," *ICML*, 2018.

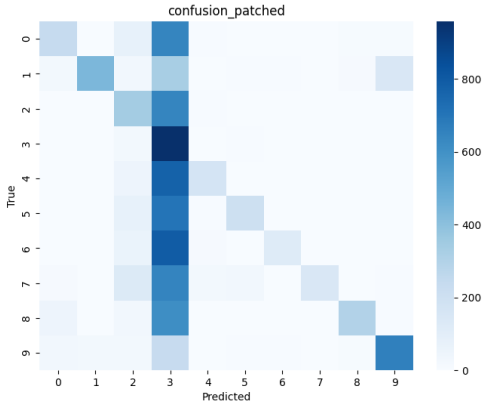

Fig. 4: Confusion matrix after applying adversarial patch.

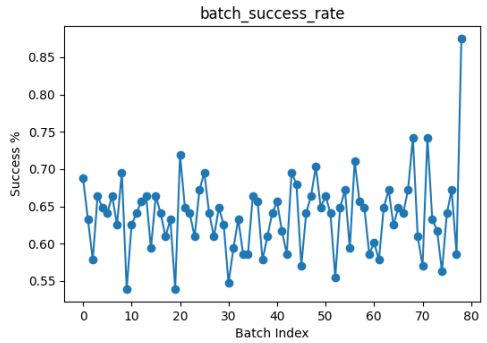

Fig. 5: Targeted attack success rate across test batches.

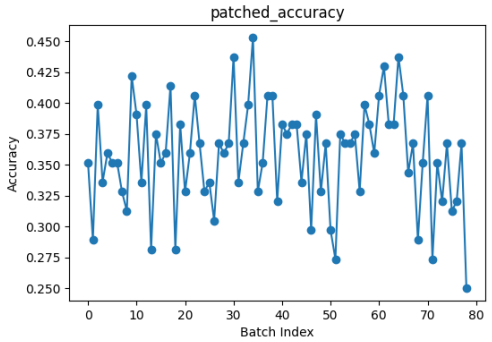

Fig. 6: Classifier accuracy degradation under attack.