Unrestricted Adversarial Attacks on Vision Transformers

Rajat Sahay
Vellore Institute of Technology, Vellore

CVPR VIRTUAL JUNE 19-25

Problem Definition and Contribution

Goal: Semantically manipulating image descriptors for adversarially attacking Vision Transformers.

Motivations:

- Existing adversarial attack methods induce high frequency noise which reduces photorealism of the image.

Original Restricted Unrestricted

- Current approaches are not effective against Vision Transformers and other self-attention based models.

Key Contributions:

- A simple approach to generate smooth adversarial perturbations while maintaining the photorealism of the image.
- Adapting existing approaches with colorization models and axial transformers to generate misclassifications from ViTs.
- Combining existing and proposed approaches to create a pipeline capable of generating adversarial examples that can fool ensemble (CNN+ViT) models.

Problem Formulation

Main idea: We assume a white-box attack setting and the goal is to adversarially color an image leveraging a pretrained colorization model.

- The Projected Gradient Descent iteratively computes:

$$\bar{x}_{adv} = x_{adv} + \epsilon * sign(T_G(x_{adv})) \quad (1)$$

where x_{adv} and ϵ denote the adversarial image and the step-size of the attack respectively.

- When given a context representation $c_g \in \mathbb{R}^{H \times W \times D}$, the Conditional Layer Norm would take a normalized input and globally scale it using learnable vectors. We aggregate c using spatial pooling into a one dimensional representation:

$$C_L = \prod_{i=1}^{H} \prod_{j=1}^{W} J_{loss}(x_{adv}, \bar{c_g}; t) \quad (2)$$

where J_{loss} is the loss of the ColTran transformer.

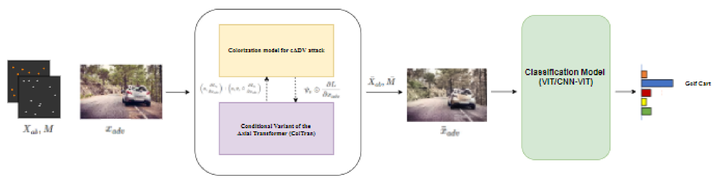
- Varying the input hints (X_{ab}) and mask (M) to semantically manipulate image:

$$\bar{X_{ab}}, \bar{M} = \underset{X_{ab}, M}{argmin}(C_L(R(C(x_{adv}, X_{ab}, M; \theta)), t)) \quad (3)$$

where C is the colorization network, R is the network in the ensemble and t is the target class.

Method

Our approach consists of the colorization network used by previous studies which experimented unrestricted adversarial methods. In addition, we also used a conditional variant of the Axial Transformer to port it to get misclassifications by self-attention based models.



T_G function for Transformer:

$$T_G(x_{adv}) = \psi_v \odot \frac{\partial L}{\partial x_{adv}} \quad (4)$$

L can be considered as the loss of the targeted Vision Transformer. $\odot$ is used to signify element-wise Hadamard product. ψ_v is the self-attention map associated with the Transformer. It can be as:

$$\psi_v = \left[\prod_{l=1}^{n_l} \left(\sum_{i=1}^{n_h} (0.5 W_{l,i}^{(att)} + 0.5I)\right)\right] \odot x \quad (5)$$

T_G for Ensemble model:

$$T_G(x_{adv}) = \left(\alpha_c \frac{\partial L_c}{\partial x_{adv}}\right) + \left(\alpha_v \psi_v \odot \frac{\partial L_v}{\partial x_{adv}}\right) \quad (6)$$

- L_c: Loss of the convolutional neural network in the ensemble
- L_v: Loss of the Transformer in the ensemble
- α_c, α_v: Weighting models to balance the emphasis on different types of models (CNNs, ViTs). These are considered as hyperparameters for the purposes of this paper.

Experiments & Results

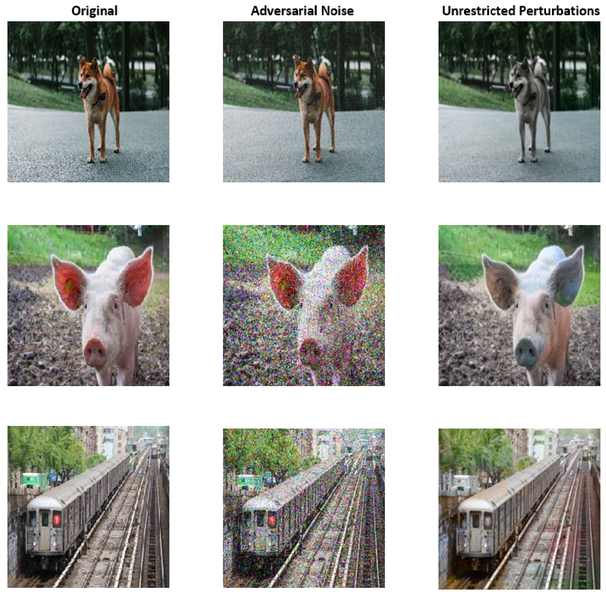
Models:

- We tested our approach by classifying images through a ViT and an ensemble (CNN-ViT).
- We considered ViT-S/16, ViT-B/16, ViT-L/16 for singular transformer models and used the R50+ViT-B/16 with some adaptations to the ResNet50 for the ensemble model.
- We used [3,4,9] blocks in the ResNet50 for the R50+ViT-B/16 ensemble instead of the conventional [3,4,6,3] blocks. Using the latter would result in a patch size of (1,1) and the ViT-B/16 model cannot be realized anymore.

Attack Success Rate on different models

Model	Clean Acc.	Attack Success
ViT-S/16	77.6	92.60
ViT-B/16	75.7	93.14
ViT-L/16	79.2	91.97
ViT-B/16-Res	84.0	90.02

Qualitative comparison of adversarial examples

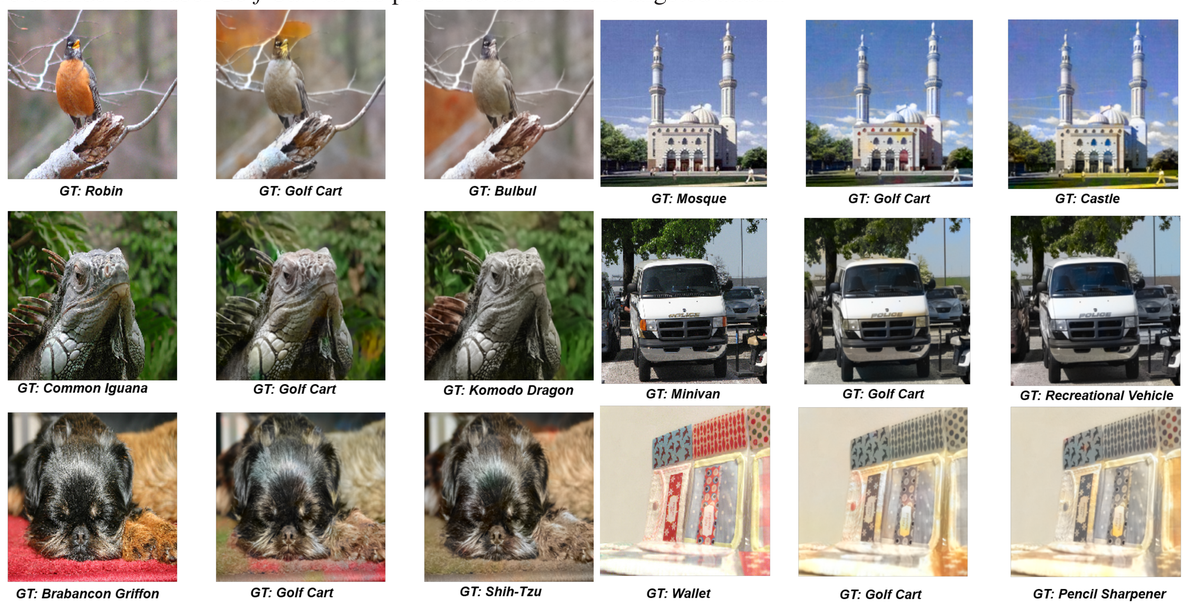


Observations after adversarial attacks:

- While the colorizer in the Axial Transformer we used for colorization generates coarse perturbations when adversarially trained, the color and spatial upsampling networks keep the image far from the original in ϵ space.
- ψ_v takes into account the attention flow of each layer of the Transformer to the next layer, including the effect of skip connections.

Qualitative results for targeted and untargeted attacks:

We consider the label *Golf Cart* as the preferred label for the targeted attack.



Results on ViT Results on CNN-ViT