Unrestricted Adversarial Attacks on Vision Transformers

Rajat Sahay
Vellore Institute of Technology
Vellore, Tamil Nadu, India
rajat.sahay2018@vitstudent.ac.in

Abstract

Recent advances in attention-based networks and following the success in advancing natural language processing and understanding have shown that Vision Transformers (ViTs) are expected to eventually replace traditional convolutional neural networks (CNNs). They have already shown to achieve state-of-the-art performance on most image classification tasks. While most CNNs have been shown to be vulnerable against adversarial examples, the same cannot be said for ViTs. This paper explores the area of unrestricted perturbations, which semantically manipulate image-based descriptors to generate naturalistic adversarial examples, and the robustness of Vision Transformers to these adversarial examples. We show that these types of adversarial attacks are effective against both ViTs alone and a combination of ViTs and CNNs used in conjunction with each other.

1. Introduction

For most deep learning tasks that include images or other visual media, convolutional neural networks are known have become a model choice [14, 17] while deep learning tasks for textual data usually use self-attention based architectures [22]. Based on the success of transformers and attention-based models in the field of natural language processing, there have been numerous attempts to use self-attention methods as a standalone [19] or by combining them with CNNs [24, 4] for a variety of image processing tasks. Dosovitskiy *et al.* [9] particularly, split images into patches and treated those patches the same way as tokens are treated in NLP applications by feeding them directly to the Transformer. The results showed that the Transformer itself is capable of competing with CNNs on image classification tasks. Since then, Vision Transformers (ViTs) have been adapted to be used in various visual tasks and are known to show equal or better performance that most convolutional neural networks or recurrent neural networks (RNNs) [29, 7]. The training in [9] is also unique to an extent since they first trained the Transformer on the ImageNet-21K dataset before training on a smaller dataset to achieve state-of-the-art accuracy on ImageNet, CIFAR-10 and CIFAR-100.

With the advancements and the increasing number of applications where ViTs are being used [6, 2], there are uncertainties on their robustness against adversarial attacks. It has been previously documented that CNNs are vulnerable to adversarial examples [5, 16, 25], harmless input images with small perturbations added which usually causes the network to incorrectly classify the image with high confidence.

There are a variety of approaches an attacker can use to generate adversarial examples [23, 25, 12]. Most of these are "restricted" in nature, that is, they search for adversarial perturbations within a bounded ε space in order to preserve their photorealism. However, previous works [13] have shown that the ε space cannot be considered a viable metric to judge visual similarity between adversarial and real images. Bhattad *et al.* [3] proposed unrestricted attack strategies which explicitly manipulated semantic visual representations to generate naturalistic adversarial examples. These examples are quite distant from the original image in the ε space, and are created by adaptively choosing locations in the images and producing substantial perturbations by changing colors or textures.

In this work, we examine the robustness of ViTs against unrestricted adversarial perturbations on image classification tasks and make comparisons with CNN benchmarks. In particular, we make use of the cADV attack proposed in [3] for our experiments and port a similar methodology towards the ColTran proposed in [15]. We also illustrate the robustness of ViTs and hybrid models of CNNs and ViTs in a white-box attack setting against these semantically manipulated adversarial examples.

2. Related Work

Transformers [22] and other attention-based architectures have been able to achieve remarkable performance on many important Natural Language Processing (NLP) tasks.

They have been widely studied from an adversarial perspective in the NLP domain [28], and have usually shown better robustness against conventional models. [21] analyzed the complex relationship between self-attention layers and proposed a robustness verification methodology for Transformers. However, due the discrete nature of NLP models, these works focus on discrete perturbations which differ significantly from the continuous imperceptible perturbations in computer vision tasks. [20] studied the effects of adversarial examples with restricted perturbations on Vision Transformers to some extent.

We also look at some of the previous works on the existing unrestricted and semantic adversarial examples. Xiao *et al.* [25] proposed the spatial distortion of pixels within an image to create adversarial examples. While these did produce natural-looking images, it did not take the semantics of the image into account. There have been previous attempts to generate adversarial perturbations by changing the hue and saturation randomly [11] but the images turned out to be unrealistic.

3. Methodology

We begin by briefly reviewing the architectures of the models that were included in our experiments. These include Vision Transformers (ViTs) and convolutional neural network (CNN) models. We then take a look at the robustness of the aforementioned models against unrestricted and semantically manipulated adversarial perturbations in a white-box setting. Our attack majorly uses the colorization attack [3] to produce adversarial examples which could be misclassified. Since the attack methodology was configured for confusing only CNN-based models, we also ported the methodology to work on a conditional variant of the Axial Transformer [10], known as ColTran. For brevity, we do not add a detailed algorithm in this paper but would be soon updating it here [1].

3.1. Model Architectures

We consider the two major types of Visual Transformers, which include the original Vision Transformer as well as the hybrid model of CNN and ViT also proposed in the same paper [9].

Vision Transformer (ViT) Initially proposed by [9], a ViT follows the design methodology of a conventional Transformer [22] used in NLP tasks. A 2D image $x \in \mathbb{R}^{H \times W \times C}$, where C is the number of color channels, is divided into a sequence of N flattened patches where $N = \frac{H.W}{k^2}$. For further support, each of the patches are encoded into embeddings using a simple convolutional layer with stride $k \times k$. Drawing inspiration from BERT [8], a token is added along with positional embeddings for the classification process. For the purposes of our experiments we consider ViT-S/16, ViT-B/16 and ViT-L/16 which are pretrained on ImageNet-21k and fine-tuned on ImageNet-1k. The 'S', 'L' and 'B' stand for 'Small', 'Large' and 'Base' while 16 denotes the number of patches the image is divided into. These are able to achieve competitively comparable results with traditional CNNs.

Combined CNN and ViT (CNN-ViT) We would also be using another architecture that was proposed in [9] which creates a hybrid between a convolutional encoder and ViTs. Unlike the traditional ViT architecture where patches of the original image are provided directly to the ViT as input, the patches here are extracted from a feature map generated by a CNN. We investigate semantically manipulated adversarial perturbations on the ViT-B/16-Res, where a ResNet50 is used to get the feature maps.

3.2. Semantic Unrestricted Adversarial Attack

We assume a white-box attack setting for the purposes of our experiments. This means that the adversary has knowledge of the trained parameters as well as the models that make up the ensemble defense. The aim would be to craft an adversarial example x_{adv} from x with unrestricted adversarial perturbations which is misclassified by all members of the ensemble.

The goal of the attack is to adversarially color an image using a pretrained colorization model. Similar to the original experiment, we also use the model proposed by Zhang *et al.* [26, 27] for this paper. Unlike most conventional attacks, which tend to generate minimized high-frequency perturbations to make them invisible to the human eye, this network is used to introduce smooth and consistent adversarial noise into the image which have a large magnitude and keep retain the photorealism of the image. The inputs consist of the L channel of the image in the CIELAB color space $X_L = \mathbb{R}^{H \times W \times 1}$, the input hints, which provide the network with the ground truth color patches to guide colorization $X_{ab} = \mathbb{R}^{H \times W \times 2}$ and the binary mask which indicates the spatial location of those patches $M \in \mathbb{B}^{H \times W \times 1}$. As with [3], we also display control over the colorization by clustering the ground truth AB space of the image using k-Means and sampling hints from the clusters with the lowest entropy.

The Projected Gradient Descent (PGD) attack [18] iteratively computes Equation 1 by taking gradient ascent:

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

In this case, the initial value of $\bar{x}_{adv}$ is the same as the input given to the network. ϵ is the step-size for the attack. The function $T_G(x_{adv})$ can be defined as

$$T_G(x_{adv}) = \psi_v \odot \frac{\partial L}{\partial x_{adv}} \tag{2}$$

[1] https://github.com/rajatsahay/unrestricted-attacks-vit/

Here, L can be considered as the loss of the targeted Vision Transformer. Similar to the cADV method, we use a Colorization Transformer known as ColTran [15] to generate colours with adversarial perturbations. A ColTran consists of three major parts, the ColTran colorizer, an autoregressive, self-attention based conditional Axial Transformer [10] which downsamples images and produces a low resolution coarse, pixel-by-pixel colorization and an auxiliary parallel color model for conditioning and regularizing as well as learning richer representations. In addition, there are also color and spatial upsampling models. In this case, we would be minimizing the targeted adversarial loss directly. When given a context representation $c \in \mathbb{R}^{H \times W \times D}$, the Conditional Layer Norm would take a normalized input and globally scale it using learnable vectors. We aggregate c into a one-dimensional representation using spatial pooling and then apply a linear projection to $\bar{c}$ to predict the learnable vectors in turn. While this does make us lose some control over the extent of colorization, with a sufficiently small value of ϵ combined with the learnable vectors, the adversarial image still manages to retain photorealism. L can be formulated as

$$L = \prod_{i=1}^{H} \prod_{j=1}^{W} J_{loss}(x_{adv}, \bar{c_g}; t) \tag{3}$$

Here, J_{loss} is the loss of ColTran as defined in [15] and t is the target class. We use $\odot$ to signify the element-wise Hadamard product in all our equations. ψ_v is the self-attention map associated with the transformer. ψ_v can be computed using self-attention roll-out [1] as

$$\psi_v = \left[\prod_{l=1}^{n_l} \left(\sum_{i=1}^{n_h} (0.5 W_{l,i}^{(att)} + 0.5I) \right) \right] \odot x \tag{4}$$

Here, x, n_h and n_l are the input image matrix, the number of attention heads and layers in the transformer respectively. $W_{l,i}^{(att)}$ is the weight attention matrix in each head and I is the identity matrix. The attention values obtained from different attention heads in the same layer are averaged while those in different layers are recursively multiplied.

To further demonstrate the generality of our attack approach, we also carry out a similar targeted attack in a white-box setting on a hybrid model of both ViTs and CNNs. Unlike conventional attack strategies, where the focus is on optimizing against a singular model, we propose to break all models in the ensemble. In this case, we generate perturbations that directly attack the CNN (ResNet50) as well as the Transformer (ViT-B/16). We use the cADV colorization scheme to attack the CNN while the ColTran-based colorization scheme to attack the transformer. Similar to Equation 1, our attack would be iteratively carried out. In this case, the value of $T_G(x_{adv})$ could be defined as

$$T_G(x_{adv}) = \left(\alpha_c \frac{\partial L_c}{\partial x_{adv}} \right) + \left(\alpha_v \psi_v \odot \frac{\partial L_v}{\partial x_{adv}} \right) \tag{5}$$

Equation 5 is majorly made up of 2 parts. The second part is the same as Equation 2 and follows the same process. The first part of the equation is where the colorization attack defined in previous works is used. L_c is the loss of the CNN used in the ensemble. Since the hints provide the patches which are responsible for the colorization as well as those from the feature maps to the ViT, and the mask provides the spatial location, we are able to exercise a suitable level of control over the colorization process of both models in the ensemble by varying the input hints (X_{ab}) and mask (M). The updated versions can be formulated as follows

$$\bar{X_{ab}}, \bar{M} = \underset{X_{ab}, M}{argmin} L_c(R(C(x, X_{ab}, M; \theta)), t) \tag{6}$$

Here, C is the colorization network by Zhang *et al.* [26], R is the network in the ensemble (a ResNet50 in our case) and t is the target class. α_c and α_v can be considered as weighting models to balance the emphasis on different models. They can be considered as hyperparameters and can be changed based on the types of ViTs or CNNs in the ensemble.

3.3. Implementation Details

We use both ViT-B/16 as well as ViT-L/16 as the Vision Transformers for image classification. For the ensemble of CNN+ViT, we use a R50+ViT-B/16 model. While the original ResNet50 has $[3, 4, 6, 3]$ blocks each of which reduce the resolution of the image by a factor if two; with the ResNet stem, the resultant would be a patch size of $(1, 1)$ and the ViT-B/16 model cannot be realized anymore. For that reason, $[3, 4, 9]$ blocks are used with the R50+ViT-B/16 hybrid. We were able to obtain pre-trained models of the Transformers as well as ColTran from their respective repositories [9, 15].

4. Experimental Results

In these experiments, we study the robustness of ViT and hybrid ViT+CNN models under a white-box, targeted attack settings. We present the results on the above models in both, qualitative as well as a quantitative fashion. The quantitative results are described in Section 4.3. We randomly sampled images from different classes in ImageNet which were predicted correctly before perturbations and misclassified after adversarial perturbations. We carry out experiments in a targeted as well as an untargeted fashion.

4.1. Individual Attack

This attack methodology defines the adversarial attack on a singular Vision Transformer on its own. The loss and

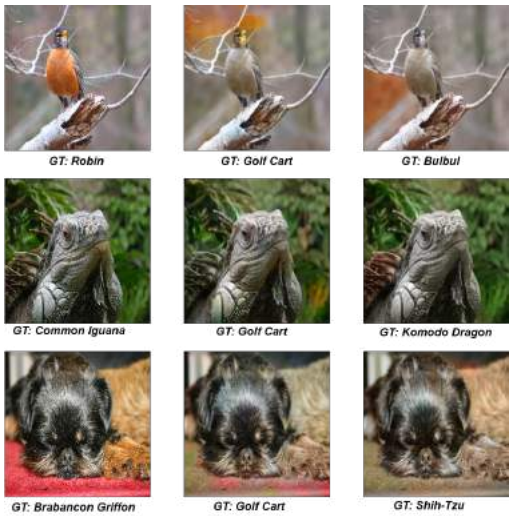

Figure 1. Unrestricted adversarial attack on Vision Transformers. The first row denotes the image, the second row denotes the image in a targeted attack (with the ground truth being *Golf Cart* and the third row denotes the prediction by the ViT with untargeted adversarial perturbations.

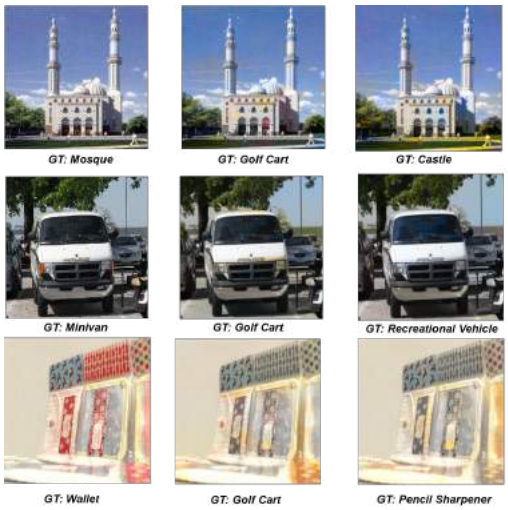

Figure 2. Unrestricted adversarial attack on a hybrid CNN+ViT model. The first row denotes the image, the second row denotes the image in a targeted attack (with the ground truth being *Golf Cart* and the third row denotes the prediction by the model with untargeted adversarial perturbations.

adversarial perturbations are computed using Equation 2 for this attack. Figure 1 shows interesting properties of the adversarial perturbations that were generated by the ColTran. We observe that the perturbations are of relatively lower frequencies, uniform and smooth in nature, which is in contrast to most conventional adversarial attacks generating high frequency perturbations. This phenomenon can be explained by the fact that even though the colorizer in the ColTran while produces coarse perturbations when adversarially trained, the colour and spatial upsampling networks remove the coarseness while keeping the image far from the original image in the ε space. The ψ_v that was computed in Equation 4 also takes into account the attention flow of each layer of the Transformer to the next layer, including the effect of skip connections.

4.2. Hybrid Attack

This attack methodology defines the adversarial attack on a hybrid of a CNN and ViT. Figure 2 shows that the generated image has similar properties as the ones generated in Section 4.1. The generated adversarial perturbations are smooth and bring the image far from the original image in the ε space, preserving the naturalistic look at the same time.

4.3. Quantitative Results

We present the results of the attack with different step sizes in Table 4.3. For the quantitative results, we use 4 different ensembles of models to ensure the generality of our approach. We use the ViT-S/16, ViT-B/16, ViT-L/16 and ViT-B/16-Res which is a R50+ViT-B/16 model. Initial qualitative comparisons over a range of color changes reveal that adversarial examples with a larger color change show more robustness against adversarial defenses. However, these large color changes cause the image to lose its sense of photorealism. We intend to conduct more extensive studies which would be able to quantify realism in images based on human perception.

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

Table 1. Quantitative Results. Attack success is defined by the $\%$ number of misclassifications in ImageNet. we use the Individual Attack for the first three and the Hybrid Attack for ViT-B/16-Res.

5. Conclusion

Our proposed approach extends previous works on colorization and generating low-frequency adversarial perturbations to Vision Transformers (ViTs). These semantic attacks shed light on the role of colors in influencing the classification prediction by ViTs. Furthermore, we also demonstrate that these attacks are successful on hybrid CNN+ViT models. We hope that these methods encourage future studies on more sophisticated defenses and improving the robustness of ViTs against unbounded adversarial attacks.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020. 3

[2] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3980–3989, 2017. 1

[3] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. *arXiv preprint arXiv:1904.06347*, 2019. 1, 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 1

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017. 1

[6] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021. 1

[7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 1

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 1, 2, 3

[10] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers, 2019. 2, 3

[11] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples, 2018. 2

[12] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*, 2018. 1

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 1

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[15] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *International Conference on Learning Representations*, 2021. 1, 3

[16] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 1

[17] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies, 2016. 1

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. 2

[19] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. 1

[20] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers, 2021. 2

[21] Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers, 2020. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 1, 2

[23] B Wang, FC Zou, and XW Liu. New algorithm to generate the adversarial example of image. *Optik*, 207:164477, 2020. 1

[24] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer, 2021. 1

[25] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples, 2018. 1, 2

[26] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2, 3

[27] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017. 2

[28] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020. 2

[29] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021. 1