Adversarial Attacks Against Vision Transformers

Tamar Noselidze
Faculty of Mathematics and Physics, Charles University

Introduction

Deep neural networks, especially vision models, are highly vulnerable to adversarial attacks — small, often imperceptible changes that cause misclassification. Among them, we mainly focus on **adversarial patch attacks**, which place a visible patch on the input image to trigger incorrect predictions.

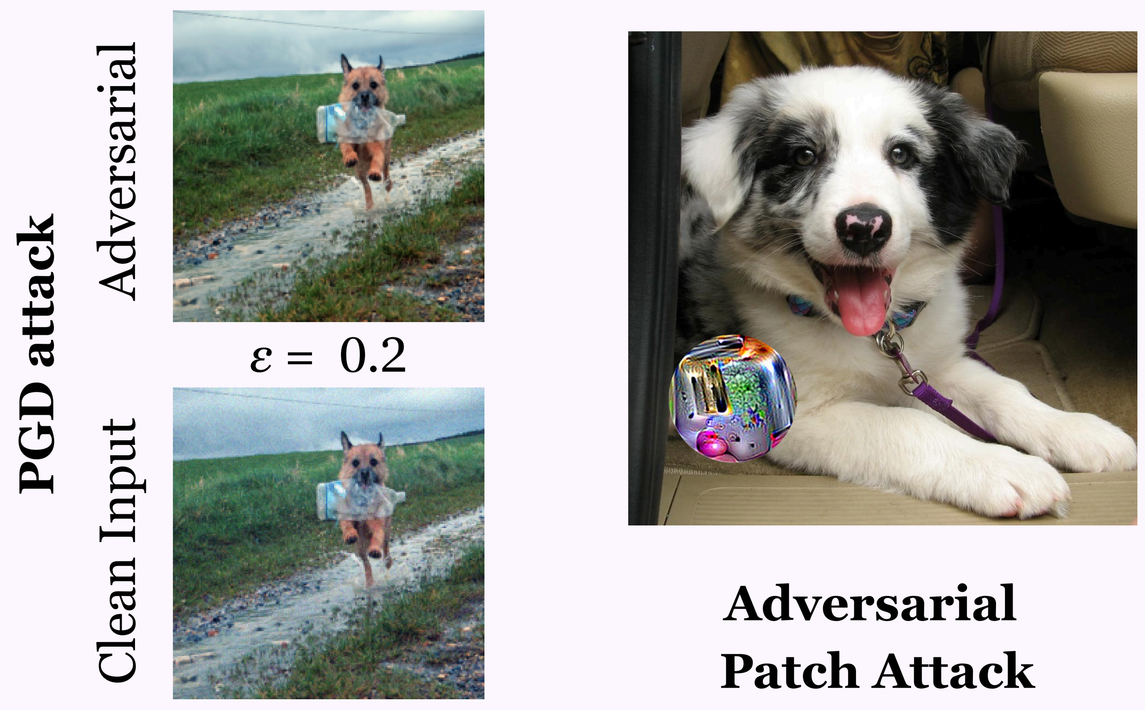

Fig 1. Adversarial examples: Gradient-based PGD method on the left, patch-based attack on the right.

Goals

- Reproduce and evaluate *standard* adversarial attacks.
- Implement a new **Random Position Patch Attack**.
- Design **Mini Patch Attacks** targeting critical regions.
- Analyze *transferability* across model architectures and families.

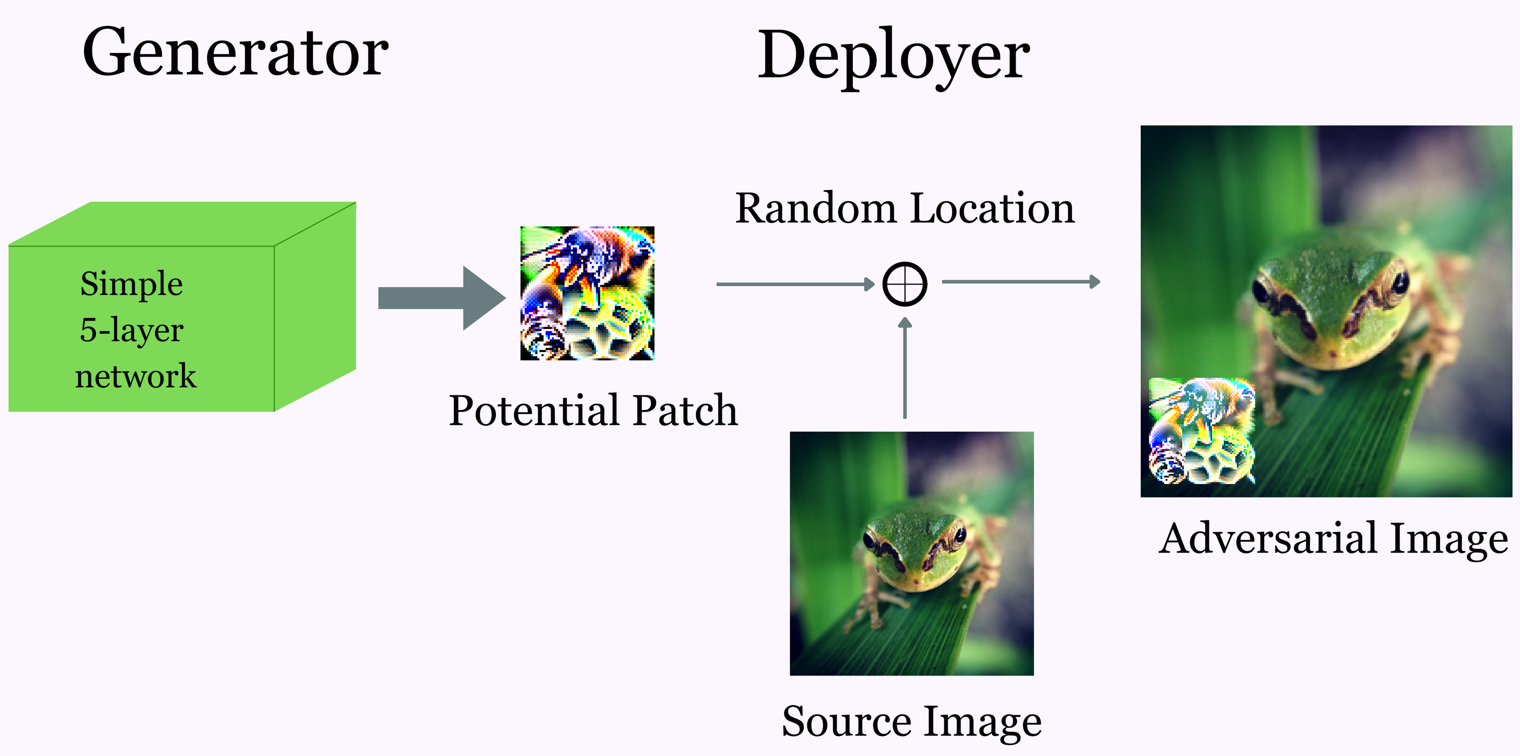

Fig. 2. Overview of the patch generating model.

Patch-based Attacks

Both variants of patch based attacks use a *Generator* to create a patch of the desired target class (of a given size).

- **G-Patches** (sizes 64x64 or 80x80) achieved consistently high attack success rates;
- **Mini-Patches** yield different success rates, based on the patch deployment approach. Experiments show that patches utilizing the internal architecture of ViT tokens perform better.

Fig. 4. Visual comparison of the initial random noise, and patches of target class **Maltese Dog** at epoch 1 and the best-performing epoch.

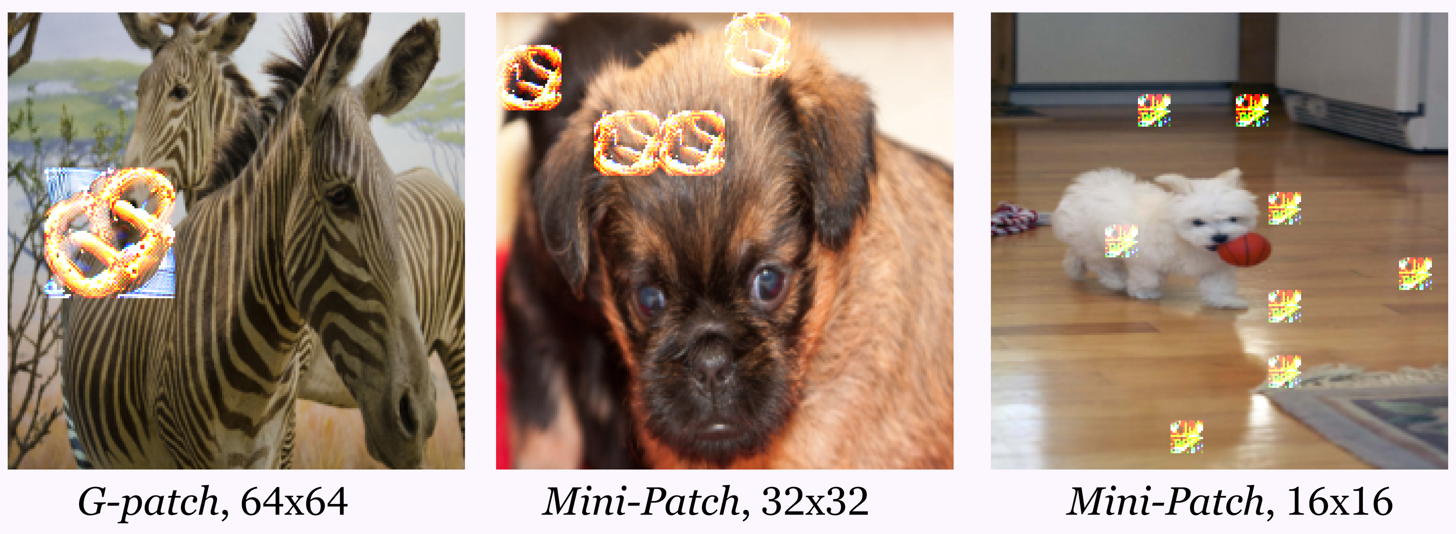

Fig. 3. Adversarial examples misclassified as **Pretzel**.

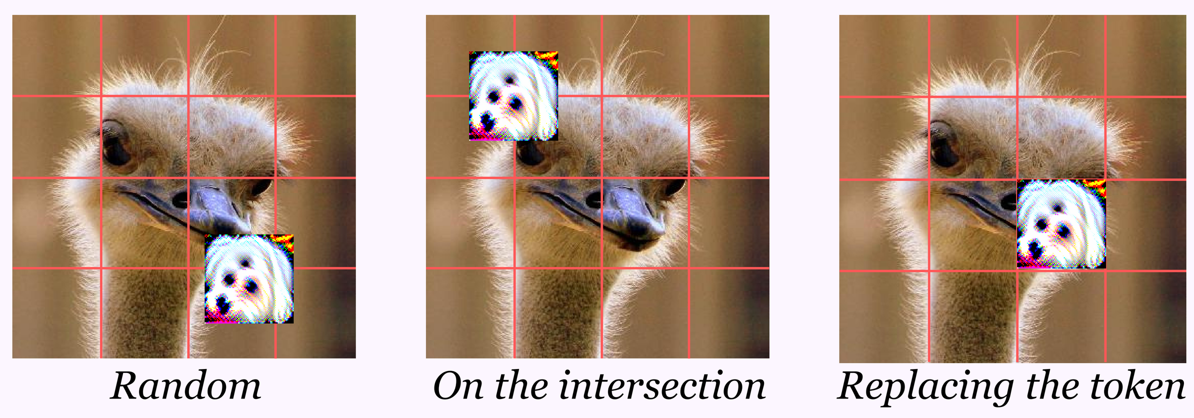

Fig. 5. Simplified visualization of 3 different approaches targeting ViT tokenization.

Transferability

For the transferability analysis, we conclude that:

- in the **G-Patch** setting, **intra-family** transferability is more effective, while inter-family variant favors patches trained on *mixed ensembles* of ViTs and CNNs.
- **Mini-Patches** targeted at *corner points* showed high sensitivity to model architectures and ensemble compositions, therefore yielding more unstable transferability results.

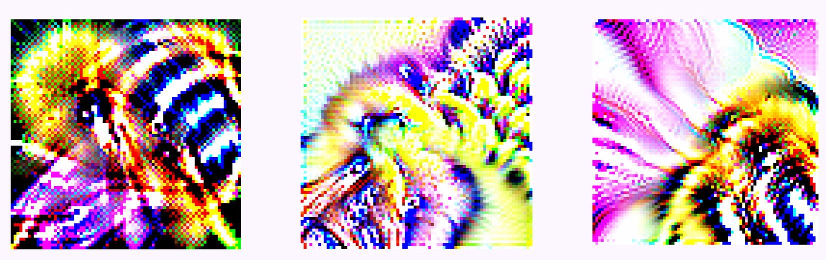

Fig. 6. Patches for target class **Bee** generated by ensembles of: ViTs, CNNs, mixed (from left to right).

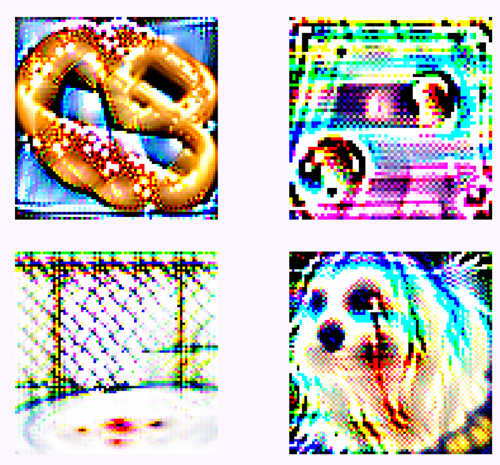

Fig. 7. Some of the best-performing patches for classes: **Pretzel**, **Cassette**, **Hockey Puck**, **Maltese Dog**.

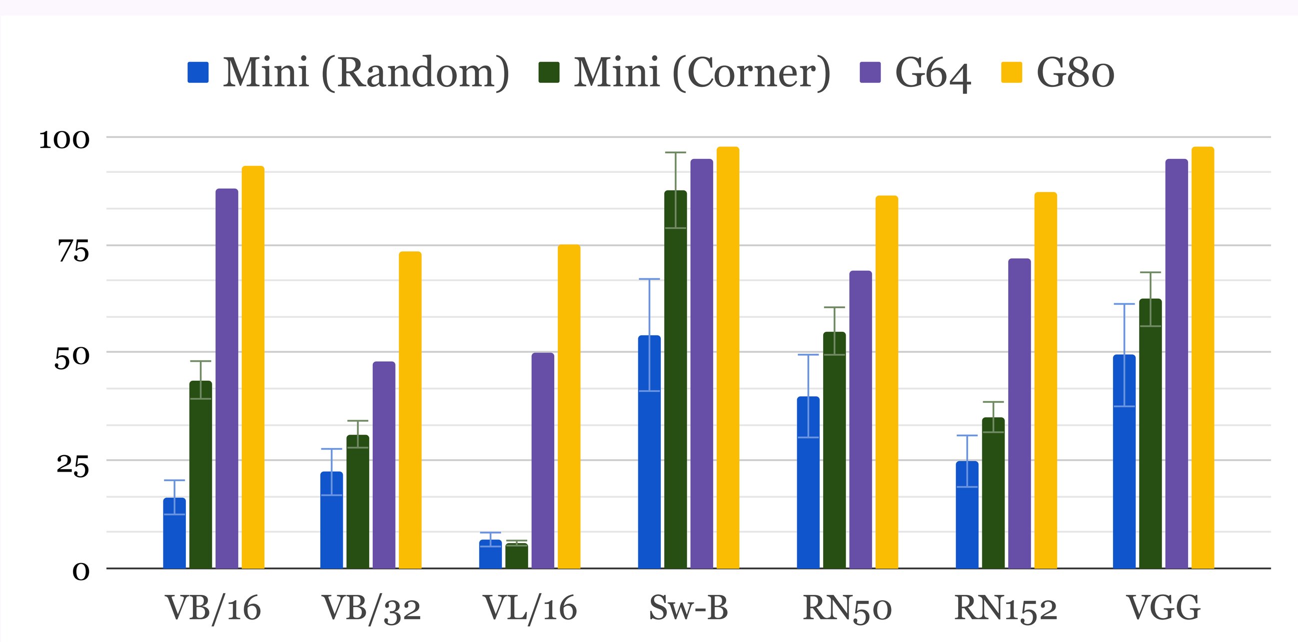

Fig. 8. Comparison of average ASRs across victim models for different types of patch attacks.

Conclusion

- G-Patches were consistently effective and transferable across models.
- Mini-Patches revealed effectiveness in the single model setting and architectural sensitivities.
- These results highlight the importance of both patch design and deployment strategy in understanding and improving the robustness of vision transformers.

Supervisor: Doc. Mgr. Martin Pilát, Ph.D.
Department of Theoretical Computer Science and Mathematical Logic

Project GitHub Page: