Paper Review of Practical No-box Adversarial Attacks against DNNs

By: Zhiying Li, Jiajia Li

Li, Qizhang, Yiwen Guo, and Hao Chen. "Practical no-box adversarial attacks against dnns." *Advances in Neural Information Processing Systems* 33 (2020): 12849-12860.

Outline

- Background - Zhiying
- Motivation - Zhiying
- Problem Formulation - Zhiying
- Method - Zhiying, Jiajia
- Experiments -Jiajia
- Results - jiajia
- Insight and Discussion - Zhiying, Jiajia
- Summary - Zhiying, Jiajia

Outline

- Background
- Motivation
- Problem Formulation
- Method
- Experiments
- Summary
- Insight and Discussion

Outline

- **Background**
- Motivation
- Problem Formulation
- Method
- Experiments
- Summary
- Insight and Discussion

Background

- Evasion Attack: generate small perturbation to fool a trained ML system

$$\min_{\delta} \quad \ell_{\text{atk}}(\mathbf{x} + \delta;\theta)$$
$$\text{s.t.} \quad \|\delta\|_p \leq \epsilon, x$$

- We find the input perturbation by input gradient through backpropagation
- Based on knowing victim model parameter or not:
  - White-box attack: directly calculate the gradient
  - Black-box attack: estimate the gradient through queries

Outline

- Background
- **Motivation**
- Problem Formulation
- Method
- Experiments
- Summary
- Insight and Discussion

Motivation

- Impracticality for real-world cases:
  - Victim model parameters cannot always allow to be known (White-box)
  - We are not allowed to query frequently (Black-box)
- No-box Attack:
  - Attack by only levearging small amount of training data
  - without knowing victim model parameters
  - without querying the victim model

Outline

- Background
- Motivation
- **Problem Formulation**
- Method
- Experiments
- Summary
- Insight and Discussion

Problem Formulation

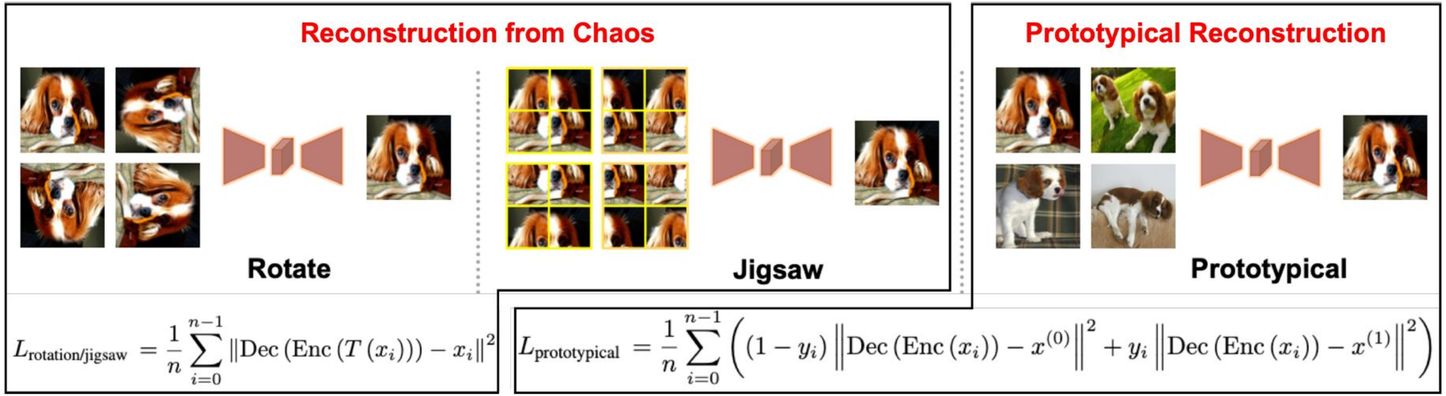
- Assume x_0 is to be perturbed
- We aim to train a “substitute” discriminative model
  - On a small and easily gathered auxiliary dataset, which includes x_0
- The adversarial perturbation is retrieved by attacking the “substitute” mode
- Attack the victim model with x_0 under such perturbation
  - In this way, we do not need to know about the victim model parameters

Outline

- Background
- Motivation
- Problem Formulation
- **Method**
- Experiments
- Summary
- Insight and Discussion

“Substitute” Model

- Not DNN classifier
    a. Overfitting due to small dataset size
- Two Autoencoder Training Mechanisms
    a. **Reconstruction from Chaos**: train an autoencoder to recover the original image from “rotation” or “Jigsaw Puzzle”
    b. **Prototypical Reconstruction**: train an autoencoder to select a single sample image from the input of class-specific subset dataset



Attack the “Substitute” Model

- Attack loss (negative cross entropy loss)

$$L_{\text{adversarial}} = -\log p\left(y_i \mid x_i\right) \quad \text{where} \quad p\left(y_i \mid x_i\right) = \frac{\exp\left(-\lambda \left\|\operatorname{Dec}\left(\operatorname{Enc}\left(x_i\right)\right) - \tilde{x}_i\right\|^2\right)}{\sum_j \exp\left(-\lambda \left\|\operatorname{Dec}\left(\operatorname{Enc}\left(x_i\right)\right) - \tilde{x}_j\right\|^2\right)},$$

- Maximizing L_adversarial is to maximizing the difference between Dec(Enc(x_i)) and \tilte{x_i}(correct output)
- Thus, minimizing the the likelihood of correct output under input perturbation
- Note: This is done with model parameters of “substitute model”, and the Attack has no knowledge of the victim model.

Intermediate Level Attack (ILA)

- In general, ILA is a method to enhance the transferability of a black-box evasion attack by increasing the perturbation on a pre-specific layer of the model
- For No box Attack, this is applied at the output layer of the encoder.
- The purpose is to improve the transferability of “substitute” model’s adversarial examples to be also robust with the victim model

Outline

- Background
- Motivation
- Problem Formulation
- Method
- **Experiments**
- Summary
- Insight and Discussion

Experimental Setup

Implement on two computer vision tasks:

- image classification:
    - Generate adversarial examples based on benign ImageNet images, maximum perturbation being no greater than 0.1 or 0.08
- face verification
    - First attack open-source models on the LFW (Labeled Face from Wild) dataset, under perturbation being 0.1
    - Then test with a commercial system held by clarifai.com
    - Faces were aligned using MTCNN

Evaluation metric: prediction accuracy of victim models on the generated adversarial examples

Experimental Approach

Process steps:

1. Train a substitute model
2. Execute a baseline attack (e.g., I-FGSM) for 200 iterations
3. Run Intermediate Level Attack (ILA) for another 100 iterations

Training mechanisms: two unsupervised (i.e., reconstruction from rotation and jigsaw) and one supervised (i.e., prototypical reconstruction) training mechanisms

Experimental Approach

Baselines:

1. Transferring adversarial examples from ResNet with supervised training (e.g., naive†)
2. Auto-encoders conventionally trained on the same small-scale datasets with unsupervised training (e.g., naive‡)
3. transferring adversarial examples from models pre-trained on a large-scale dataset (e.g., Beyonder)

Victim models:

- Image classification task: 8 classical DNN models (e.g., VGG-19, ResNet152)
- Face verification task: 2 models, FaceNet and Cosface

Experimental Results

Image Classification:

- Our approach and the two baselines (i.e., naïve † and naïve ‡) involve <u>only 20 images</u> to train each substitute model

Table 1: Compare the transferability of adversarial examples crafted on different models on ImageNet. The prediction accuracy on adversarial examples under $\epsilon = 0.1$ are shown (lower is better).

Method	Sup.	VGG-19 [42]	Inception v3 [45]	ResNet [15]	DenseNet [17]	SENet [16]	WRN [56]	PNASNet [28]	MobileNet v2 [39]	Average
Naïve‡	✗	45.92%	63.94%	60.64%	56.48%	65.54%	58.80%	73.14%	37.76%	57.78%
Jigsaw	✗	31.54%	50.28%	46.24%	42.38%	59.06%	51.24%	62.32%	25.24%	46.04%
Rotation	✗	31.14%	48.14%	47.40 %	41.26%	58.20%	50.72%	59.94%	26.00%	45.35%
Naïve†	✓	76.20%	80.86%	83.76%	78.94%	87.00%	84.16%	86.96%	72.44%	81.29%
Prototypical	✓	19.78%	36.46%	37.92%	29.16%	44.56%	37.28%	48.58%	17.78%	33.94%
Prototypical*	✓	18.74%	33.68%	34.72%	26.06%	42.36%	33.14%	45.02%	16.34%	**31.26%**
Beyonder	✓	24.96%	51.12%	30.30%	27.12%	43.78%	33.94%	51.80%	27.02%	36.26%

* The prototypical models with multiple decoders. To be more specific, 20 decoders are introduced in each model.

Experimental Results

Image Classification:

- The rotation and jigsaw mechanisms both outperform the unsupervised baseline

Table 1: Compare the transferability of adversarial examples crafted on different models on ImageNet. The prediction accuracy on adversarial examples under $\epsilon = 0.1$ are shown (lower is better).

Method	Sup.	VGG-19 [42]	Inception v3 [45]	ResNet [15]	DenseNet [17]	SENet [16]	WRN [56]	PNASNet [28]	MobileNet v2 [39]	Average
Naïve‡	✗	45.92%	63.94%	60.64%	56.48%	65.54%	58.80%	73.14%	37.76%	57.78%
Jigsaw	✗	31.54%	50.28%	46.24%	42.38%	59.06%	51.24%	62.32%	25.24%	46.04%
Rotation	✗	31.14%	48.14%	47.40 %	41.26%	58.20%	50.72%	59.94%	26.00%	45.35%
Naïve†	✓	76.20%	80.86%	83.76%	78.94%	87.00%	84.16%	86.96%	72.44%	81.29%
Prototypical	✓	19.78%	36.46%	37.92%	29.16%	44.56%	37.28%	48.58%	17.78%	33.94%
Prototypical*	✓	18.74%	33.68%	34.72%	26.06%	42.36%	33.14%	45.02%	16.34%	**31.26%**
Beyonder	✓	24.96%	51.12%	30.30%	27.12%	43.78%	33.94%	51.80%	27.02%	36.26%

* The prototypical models with multiple decoders. To be more specific, 20 decoders are introduced in each model.

Experimental Results

Image Classification:

- Prototypical models with multiple decoders yield the most transferable adversarial examples overall

Table 1: Compare the transferability of adversarial examples crafted on different models on ImageNet. The prediction accuracy on adversarial examples under $\epsilon = 0.1$ are shown (lower is better).

Method	Sup.	VGG-19 [42]	Inception v3 [45]	ResNet [15]	DenseNet [17]	SENet [16]	WRN [56]	PNASNet [28]	MobileNet v2 [39]	Average
Naïve[‡]	✗	45.92%	63.94%	60.64%	56.48%	65.54%	58.80%	73.14%	37.76%	57.78%
Jigsaw	✗	31.54%	50.28%	46.24%	42.38%	59.06%	51.24%	62.32%	25.24%	46.04%
Rotation	✗	31.14%	48.14%	47.40 %	41.26%	58.20%	50.72%	59.94%	26.00%	45.35%
Naïve[†]	✓	76.20%	80.86%	83.76%	78.94%	87.00%	84.16%	86.96%	72.44%	81.29%
Prototypical	✓	19.78%	36.46%	37.92%	29.16%	44.56%	37.28%	48.58%	17.78%	33.94%
Prototypical*	✓	18.74%	33.68%	34.72%	26.06%	42.36%	33.14%	45.02%	16.34%	**31.26%**
Beyonder	✓	24.96%	51.12%	30.30%	27.12%	43.78%	33.94%	51.80%	27.02%	36.26%

* The prototypical models with multiple decoders. To be more specific, 20 decoders are introduced in each model.

Experimental Results

Image Classification:

- Training curves of authors' multiple-decoder prototypical models
- Less over-fitting and higher benign-set accuracy of the substitute models in comparison with the conventional supervised models

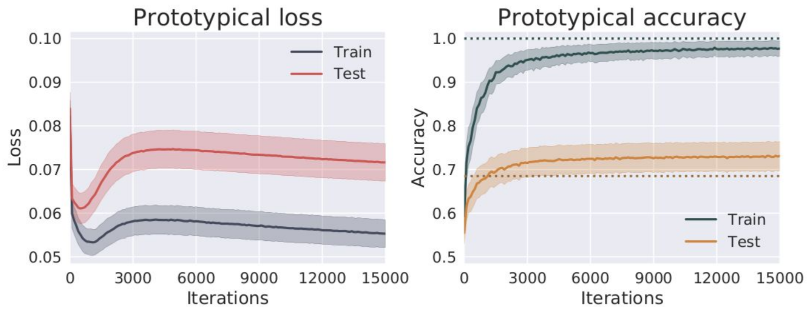


Figure 4: Our prototypical reconstruction mechanism leads to less over-fitting and higher *benign-set accuracy* of the substitute models in comparison with the conventional supervised models in Figure 1 and 2, using a small number of training images. The shaded areas indicate the amount of variance, and the dotted lines indicate final accuracies of the regularized VGG models in Figure 2.

Experimental Results

Image Classification:

- Training curves of authors' multiple-decoder prototypical models
- Less over-fitting and higher benign-set accuracy of the substitute models in comparison with the conventional supervised models

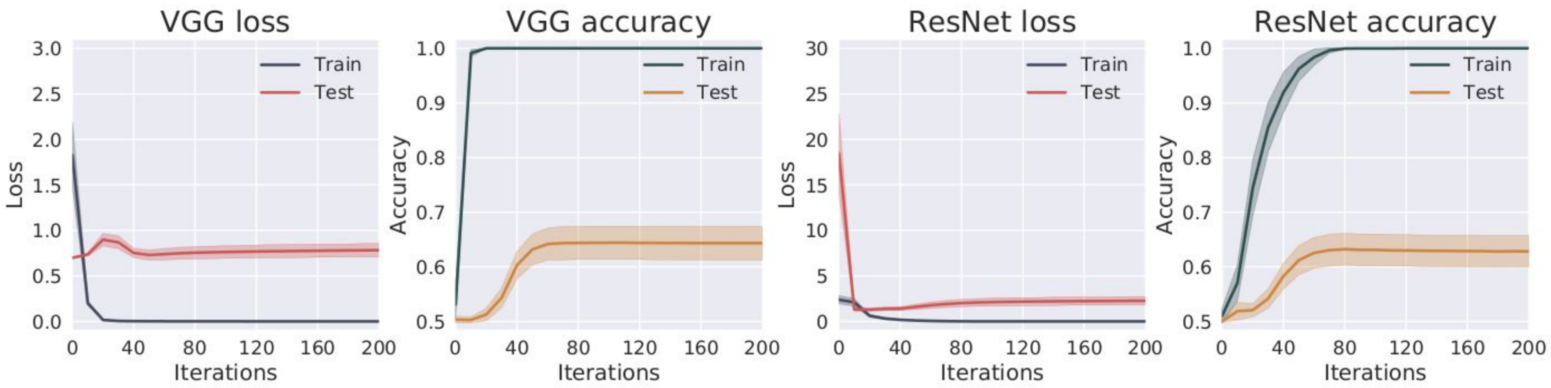


Figure 1: With limited training data, conventional supervised learning suffer from severe over-fitting.

Experimental Results

Image Classification:

- Training curves of authors' multiple-decoder prototypical models
- Less over-fitting and higher benign-set accuracy of the substitute models in comparison with the conventional supervised models

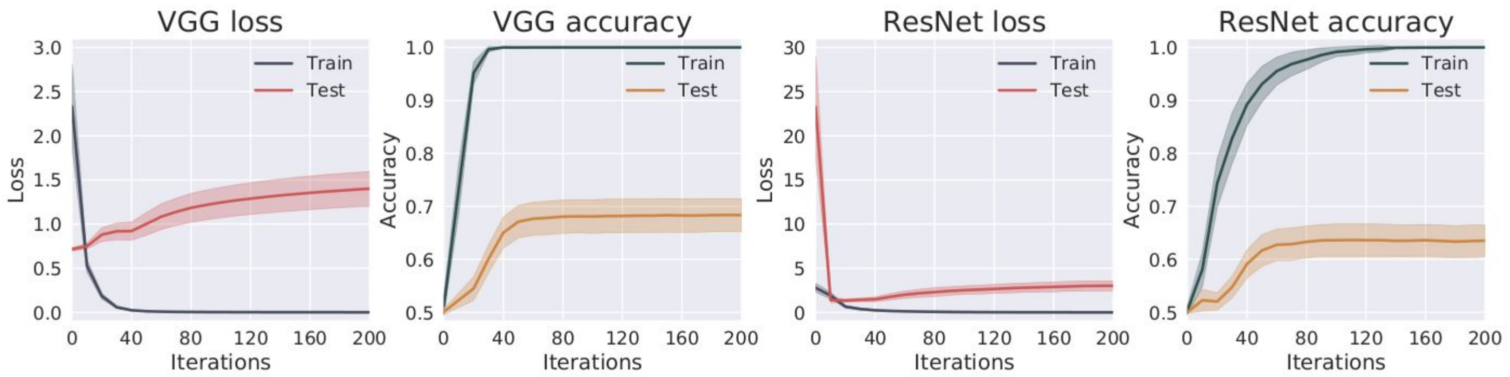


Figure 2: Data augmentations and regularizations help to a limited extent in the conventional supervised setting. Weight decay, dropout, and some popular data augmentations are adopted.

Experimental Results

Image Classification(Visual explanations):

- Visualize some adversarial examples and the model attention on the examples using Grad-CAM

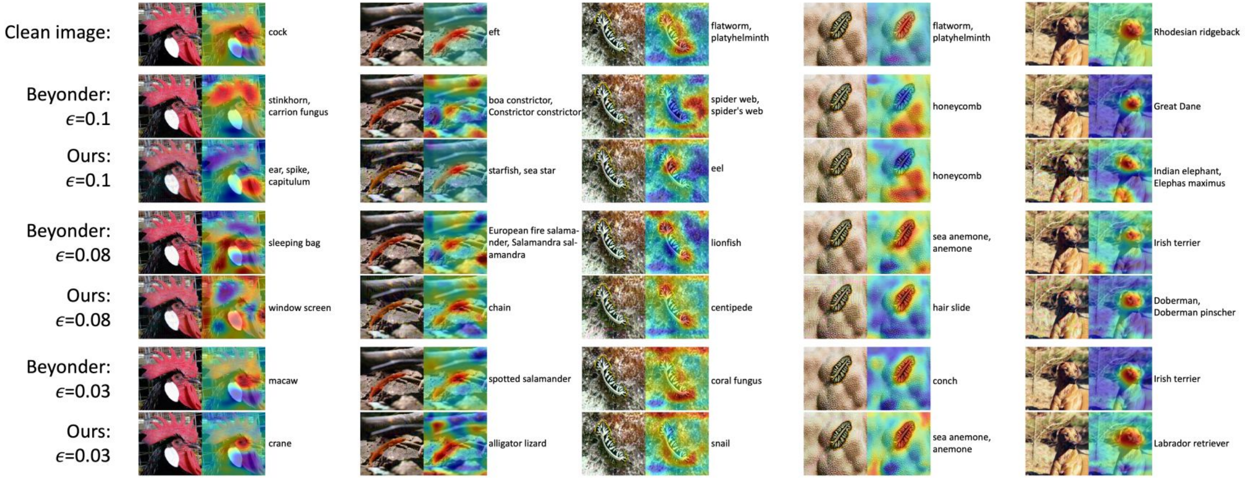


Figure 6: Visual explanation of how the Beyonder adversarial examples and our no-box adversarial examples fool the VGG-19 victim model. Grad-CAM is used.

Experimental Results

Image Classification(Visual explanations):

- The authors' adversarial examples <u>divert the model attention from important image regions</u>

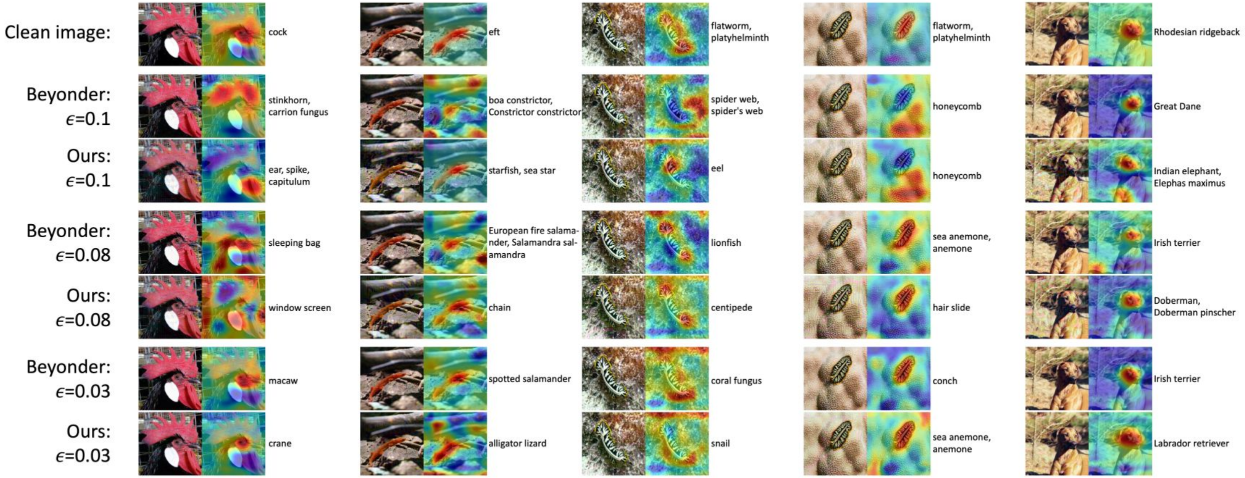


Figure 6: Visual explanation of how the Beyonder adversarial examples and our no-box adversarial examples fool the VGG-19 victim model. Grad-CAM is used.

Experimental Results

Image Classification(Visual explanations):

- The authors' no-box adversarial examples are intrinsically and perceptually very different from the Beyonder adversarial examples.

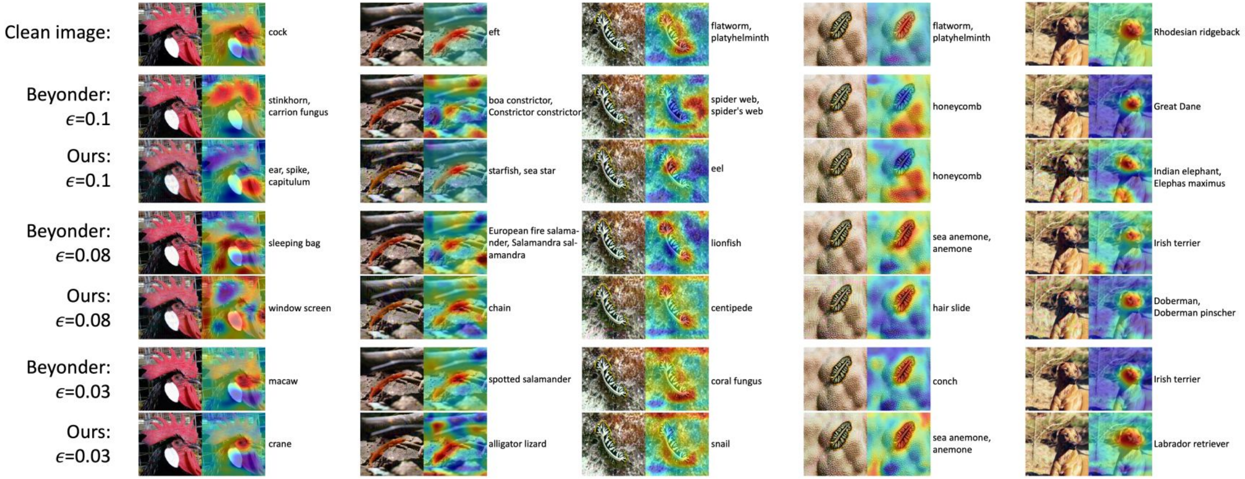


Figure 6: Visual explanation of how the Beyonder adversarial examples and our no-box adversarial examples fool the VGG-19 victim model. Grad-CAM is used.

Experimental Results

Image Classification(Number of training images):

- All the proposed mechanisms perform reasonably well with no more than 20 images (i.e., n ≤ 20) on ImageNet

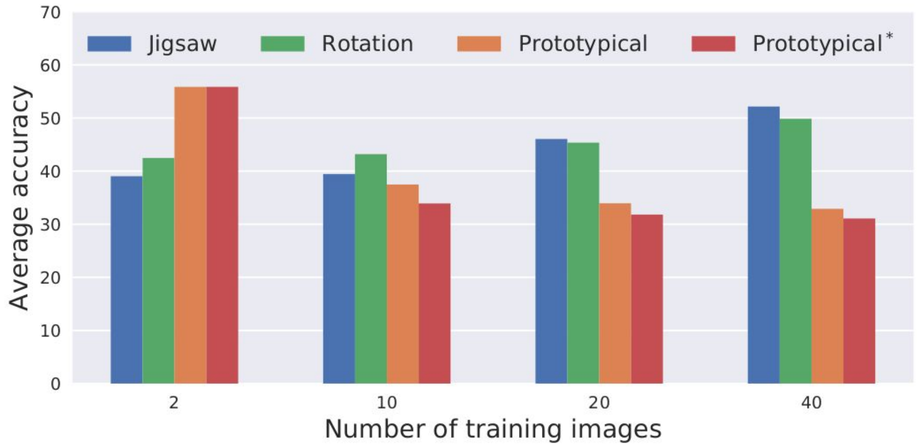


Figure 7: How the attack performance of our approach varies with the number of training images on ImageNet. Lower average accuracy indicate better performance in attacking the victim models.

Experimental Results

Image Classification(Number of training images):

- By further increasing n to 40, the prototypical mechanism achieves even better performance in the sense of no-box transfer

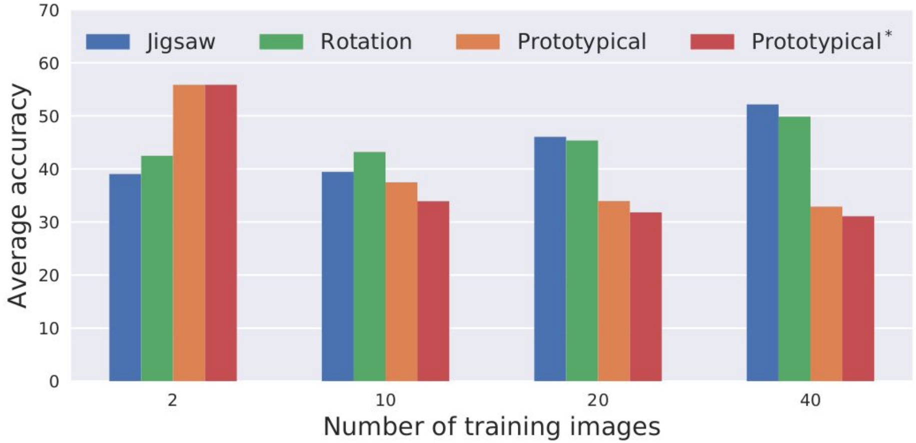


Figure 7: How the attack performance of our approach varies with the number of training images on ImageNet. Lower average accuracy indicate better performance in attacking the victim models.

Experimental Results

Image Classification(Number of training images):

- <u>Rotation and Jigsaw models</u> works better with less training images, due to faster training convergence within the limited number of training iterations.

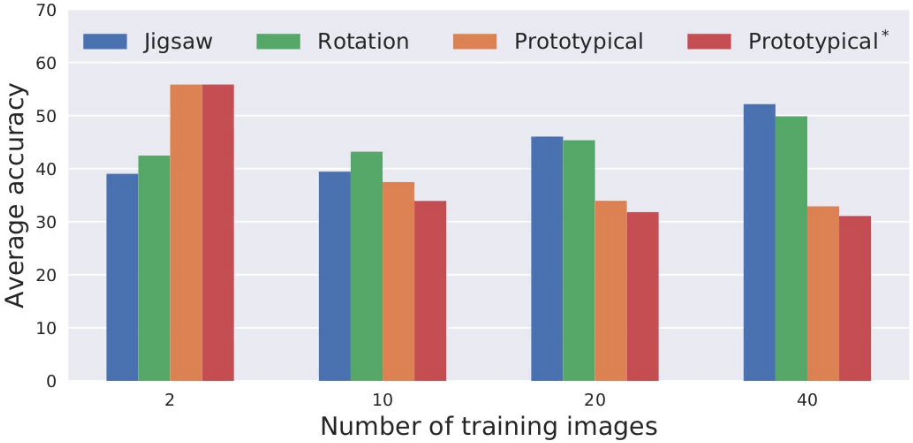


Figure 7: How the attack performance of our approach varies with the number of training images on ImageNet. Lower average accuracy indicate better performance in attacking the victim models.

Experimental Results

Image Classification(Number of prototypical decoders):

- The more decoders get involved, the higher attack success rates can be achieved
- Take longer to converge with more decoders, suggesting a trade-off between the attack success rate and training scale.
- Explain: richer supervision can be obtained from more decoders and more image anchors

Table 4: How the number of prototypical decoders impact attack performance on ImageNet victim models. Results are obtained under ℓ_∞ attacks with $\epsilon = 0.1$. Lower is better.

#decoders	VGG-19 [7]	Inception v3 [8]	ResNet [1]	DenseNet [3]	SENet [2]	WRN [9]	PNASNet [4]	MobileNet v2 [5]	Average
1	19.78%	36.46%	37.92%	29.16%	44.56%	37.28%	48.58%	17.78%	33.94%
5	19.48%	34.32%	35.90%	26.44%	42.70%	34.72%	46.12%	17.37%	32.13%
10	19.16%	34.18%	35.00%	25.94%	42.14%	33.16%	45.22%	17.18%	31.50%
20	18.74%	33.68%	34.72%	26.06%	42.36%	33.14%	45.02%	16.34%	**31.26%**

Experimental Results

Face Verification:

- Test on the basis of LFW (Labeled Face from Wild) images
- <u>Multiple-decoder prototypical models</u> still achieve the best performance in attacking FaceNet, which is even better than that of Beyonder

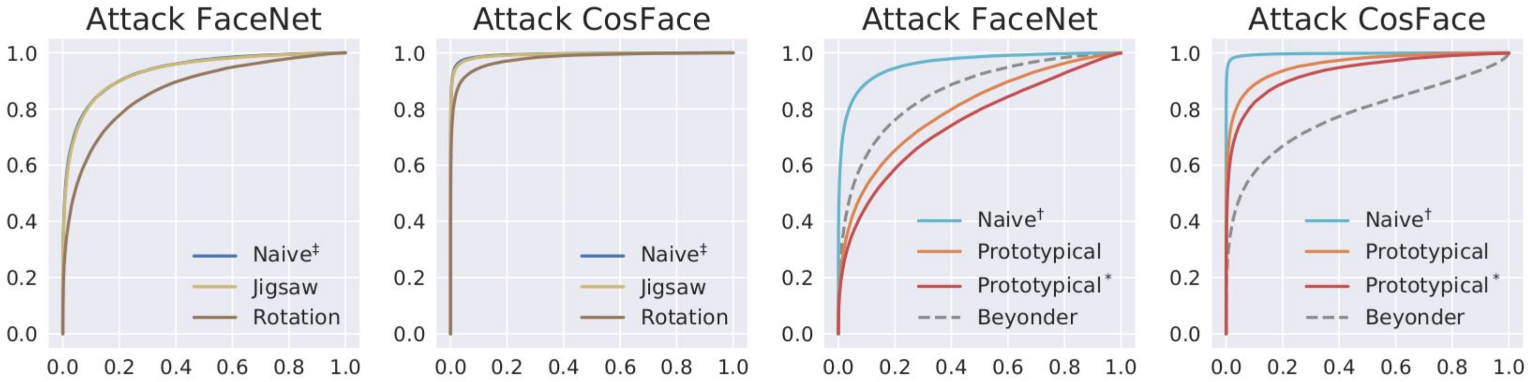


Figure 5: ROC curves of face verification on adversarial examples crafted on different substitute models. The left two sub-figures show *unsupervised* results and the right two show *supervised* results.

Outline

- Background
- Motivation
- Problem Formulation
- Method
- Experiments
- **Summary**
- Insight and Discussion

Summary

- The paper provides a novel way to achieve transfer evasion attack with small amount of training data
- It achieves "practicability of attack"
  - when model parameter is infeasible
  - when querying and large-scale training are infeasible
- Core of the method: 3 autoencoder substitute models
- Uses Intermediate Level Attack (ILA) to improve the transferability of perturbation
- Successfully diminish the prediction results of Image Recognition (31%) and Face Verification (14%)

Outline

- Background
- Motivation
- Problem Formulation
- Method
- Experiments
- Summary
- **Insight and Discussion**

Insights and Discussion

- Data-free attack approaches [1, 2]: build an attack without collecting private data
    - The proposed work requires a small number of auxiliary samples, such as 20 images, while sometimes, collecting the images for security-sensitive applications is difficult and infeasible.
- Training is time-consuming and inefficient when attacking a new sample out of the distribution
- No-box attack for more complicated applications, such as object detection [3] and segmentation

[1] Q. Zhang, C. Zhang, C. Li, J. Song, L. Gao, and H. T. Shen, “Practical no-box adversarial attacks with training-free hybrid image transformation,” arXiv preprint arXiv:2203.04607, 2022.
[2] C. Zhang, P. Benz, A. Karjauv, and I. S. Kweon, “Data-free universal adversarial perturbation and black-box attack,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp.7868–7877.
[3] Z. Cai, S. Rane, A. E. Brito, C. Song, S. V. Krishnamurthy, A. K. Roy-Chowdhury, and M. S. Asif, “Zero-query transfer attacks on context-aware object detectors,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15 024–15 034.