CDTA: A Cross-Domain Transfer-Based Attack with Contrastive Learning

Zihan Li[1], Weibin Wu[1], Yuxin Su[1*], Zibin Zheng[1], Michael R. Lyu[2]

[1] School of Software Engineering, Sun Yat-sen University
[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong
lizh523@mail2.sysu.edu.cn, {wuwb36, suyx35, zhzibin}@mail.sysu.edu.cn, lyu@cse.cuhk.edu.hk

Abstract

Despite the excellent performance, deep neural networks (DNNs) have been shown to be vulnerable to adversarial examples. Besides, these examples are often transferable among different models. In other words, the same adversarial example can fool multiple models with different architectures at the same time. Based on this property, many black-box transfer-based attack techniques have been developed. However, current transfer-based attacks generally focus on the cross-architecture setting, where the attacker has access to the training data of the target model, which is not guaranteed in realistic situations. In this paper, we design a Cross-Domain Transfer-Based Attack (CDTA), which works in the cross-domain scenario. In this setting, attackers have no information about the target model, such as its architecture and training data. Specifically, we propose a contrastive spectral training method to train a feature extractor on a source domain (e.g., ImageNet) and use it to craft adversarial examples on target domains (e.g., Oxford 102 Flower). Our method corrupts the semantic information of the benign image by scrambling the outputs of both the intermediate feature layers and the final layer of the feature extractor. We evaluate CDTA with 16 target deep models on four datasets with widely varying styles. The results confirm that, in terms of the attack success rate, our approach can consistently outperform the state-of-the-art baselines by an average of 11.45% across all target models. Our code is available at https://github.com/LiulietLee/CDTA.

Introduction

Deep neural networks (DNNs) have been applied to many applications, even in some security-critical areas. Examples include autonomous driving (Kumar et al. 2021), face recognition (Singh et al. 2020), and assisted medical diagnosis (Li et al. 2020). However, large numbers of works have shown that DNNs can be easily fooled by adversarial examples, which are usually generated by attaching a tiny but carefully crafted perturbation to a benign image (Szegedy et al. 2014; Wu et al. 2020b). Therefore, developing adversarial attack methods to figure out the vulnerability of deep models is important for improving their security.

There are mainly two categories of adversarial attacks: white-box attacks and black-box attacks. In most cases, deep models are opaque to their users, so a white-box attack, which requires full knowledge of the target model, can hardly be used in real life. Black-box attacks, as a type of approach that restrict the information available to attackers, can better evaluate the robustness of a model in practice.

Black-box attacks can be roughly divided into transfer-based attacks and query-based attacks (Akhtar and Mian 2018). Query-based attacks generate adversarial perturbations by massively querying target models (Guo et al. 2019). However, massive queries can be easily detected and then blocked, which limits the practical applicability of query-based attacks. Transfer-based attacks work by employing the transferability of adversarial samples. The transferability of adversarial samples refers to the phenomenon that the adversarial examples generated via one deep model can also fool another deep model, even though they have different architectures and parameters. Therefore, transfer-based attacks can pose a severe threat to the security of DNNs. Nevertheless, current studies on transfer-based attacks are majorly concentrated on the cross-architecture setting (Wu et al. 2020a, 2021; Zhang et al. 2022a). This setting assumes that the surrogate models used by attackers share the same training data distribution with the target models. However, this assumption hardly holds in practice, because the owners of deployed deep models scarcely release their training data to the public.

Therefore, in this work, we focus on transfer-based attacks in a cross-domain setting, which can more faithfully reflect the realistic threat to DNNs. Specifically, a cross-domain transfer-based attack works by generating adversarial samples via a surrogate model on a source domain, and then directly using the generated adversarial samples to fool target black-box models with different architectures and parameters on *completely different target domains*. Existing cross-domain transfer-based attacks usually have limited attack success rates (Naseer et al. 2019; Zhang et al. 2022b), which makes cross-domain transfer-based attacks under-explored.

To this end, we propose a Cross-Domain Transfer-Based Attack (CDTA). Specifically, we train a feature extractor on a source domain and use it to produce adversarial examples to fool target black-box models on completely different

[*] Corresponding author.

target domains. To improve the attack success rates of our CDTA, we need to enhance the domain-agnostic property of the feature extractor. Inspired by self-supervised learning (He et al. 2020; Grill et al. 2020; Chen and He 2021), we propose a novel contrastive spectral training method. This no-label training method is more likely to fit the task of training a cross-domain feature extractor because no label information in the training phase means that the deep model will not overfit any specific classification task. Therefore, disrupting the feature representations output by the trained feature extractor can destroy more cross-domain semantic components in benign images, which leads to a better cross-domain attack performance.

Besides, recent studies discovered that disrupting intermediate layer features is more effective than directly attacking the final predictions of surrogate models (Ganeshan, BS, and Babu 2019; Inkawhich et al. 2020b; Wang et al. 2021; Zhang et al. 2022a). Therefore, we propose to generate adversarial samples with a new loss function, which combines the intermediate layer features and the output "prediction" of the trained feature extractor. Our main contributions are as follows:

- We propose a novel Cross-Domain Transfer-Based Attack (CDTA) to craft adversarial examples in the cross-domain setting, which is the most strict black-box setting. In this setting, an attacker has zero knowledge about the target model, such as its parameters, architecture, training data, and label space.
- We are the first to show that a feature extractor trained by self-supervised contrastive learning can be an effective surrogate model for cross-domain transfer-based attacks. Besides, adversarial samples generated by attacking the feature extractor can achieve much better performance in the cross-domain scenario than those generated by attacking a classification network obtained by supervised training.
- Experiments confirm that, compared with the state-of-the-art baselines, CDTA can largely improve the attack success rate by 11.45% on average across multiple datasets.

Related Work

Self-Supervised Learning

Self-supervised learning methods have two major categories: generative and contrastive (Liu et al. 2021). Our work is highly motivated by contrastive learning. Therefore, we make a brief introduction here.

The basic idea of contrastive learning is to make the representations of positive samples stay close to each other while those of negative samples be far apart. In particular, contrastive learning regards each data point in the dataset as a class on its own. An image and the image enhanced by data augmentation methods form a positive sample pair. The closer their corresponding feature vectors are to each other, the better the contrastive learning results. On the contrary, two different pictures form a negative sample pair. The further apart their corresponding feature vectors are, the better the contrastive learning results. In the end, the feature vectors of images with similar content are close to each other, while those of images with greater content differences are further apart.

Hjelm et al. involve mutual information to learn meaningful representations (Hjelm et al. 2019). Wu et al. store every feature vector in a memory bank to increase negative samples as many as possible (Wu et al. 2018). He et al. enrich negative samples by a dictionary queue and proposes a momentum update strategy to keep the consistency of the dictionary (He et al. 2020). Grill et al. achieve state-of-the-art performance without negative sample pairs (Grill et al. 2020). Chen and He simplify the architecture and can converge faster than previous methods (Chen and He 2021).

Transfer-based Attack

Most transfer-based black-box attacks closely follow the previously proposed white-box methods, adding some regularization terms and optimization tricks to alleviate overfitting and boost the transferability of crafted adversarial examples. Some recent works have been proposed to expand the cross-domain attack ability of transfer-based attacks (Naseer et al. 2019; Zhang et al. 2022b).

There are two approaches to optimizing perturbations: decision-space attacks and feature-space attacks.

Decision-space attacks try to push predictions out of the correct decision boundaries by directly focusing on the output layers of the classifiers. Usually, it is achieved through optimizing classification loss. Dong et al. design a momentum-based iterative algorithm to boost adversarial attacks (Dong et al. 2018). Dong et al. propose a translation-invariant method to improve the transferability of adversarial examples (Dong et al. 2019). Xie et al. use random transformations to the input images at each iteration to increase the diversity (Xie et al. 2019). Li, Guo, and Chen use reconstructions of images from rotations, jigsaw puzzles, and prototypes to train a surrogate model on small sub-datasets of target domains, which is then used to perform transfer-based attacks on specific images (Li, Guo, and Chen 2020).

Feature-space attacks craft adversarial perturbations by disrupting intermediate feature layers instead of the classification layer. Ganeshan, BS, and Babu show the drawbacks of decision-space attacks and propose a new attack called FDA (Ganeshan, BS, and Babu 2019). Inkawhich et al. present a new adversarial attack based on the modeling of class-wise and layer-wise deep feature distributions (Inkawhich et al. 2020b). Inkawhich et al. design a flexible attack framework that allows multi-layer perturbations and achieves state-of-the-art targeted transfer-based attack performance (Inkawhich et al. 2020a). Yet, this approach requires that the label spaces of the white-box model and the black-box model should overlap. Lu et al. and Naseer et al. abandon the task-specific loss function and corrupt images by attacking intermediate features of deep models (Lu et al. 2020; Naseer et al. 2020). However, they focus on attacking different vision tasks but not cross-domain settings. Inkawhich et al. propose a correlation matrix-based attack method, but it still needs partial data from the target domain (Inkawhich et al. 2021).

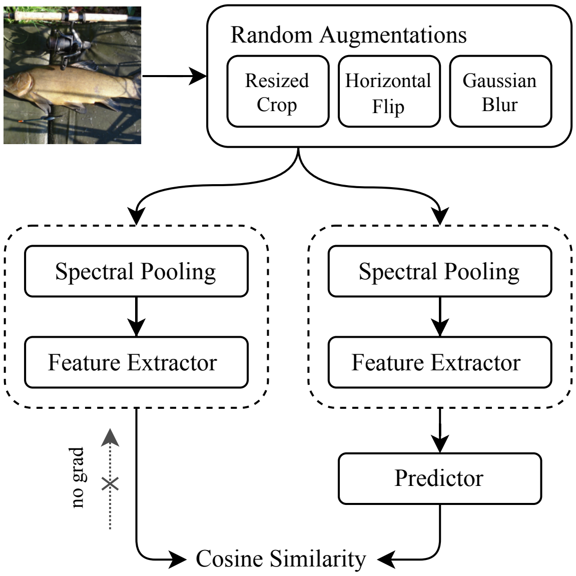

Figure 1: Contrastive spectral training. We first apply different random augmentations to the same training image to obtain a positive sample pair. We feed the sample pair to our feature extractor and that combined with an MLP predictor, respectively. Then we maximize the cosine similarity between the output vectors by only updating the feature extractor combined with an MLP predictor.

Methodology

Overview

Before diving into the detail of our method, let's clarify the research question first.

Currently, deep models mainly face two types of attacks: white-box attacks and black-box attacks. The white-box scenario allows attackers to fully access all information about the target model. However, the black-box scenario is actually a complicated concept. Black-box scenarios in different papers are not exactly the same. So here we group them into the following categories:

- **Cross-architecture scenario.** This setting usually restricts attackers' access to the architecture, parameters, and gradients of the target model. But generally, attackers can use the same dataset to train surrogate models, so training data and label space of the target model are actually accessible (Wu et al. 2021).
- **Relaxed cross-domain scenario.** As we mentioned before, some works try to further limit the information available to transfer-based attacks. In this setting, attackers can only get a small portion of the training data or labels, and no other information is available (Li, Guo, and Chen 2020).
- **Strict cross-domain scenario.** This setting is the most strict black-box setting that we focus on in this paper. It is challenging because no information about the target model is available to attackers. In this case, attackers need to propose a highly generalizable method to attack the target model. (Zhang et al. 2022b).

Let $x_t \in R^{3\times H\times W}$ be a colored image in the target domain D_t, and y^* be the associated ground-truth label. Given a feature extractor h trained on the source domain D_s, a feature loss L, and a target deep classifier f trained on the target domain D_t, our goal is to derive an adversarial example $x_t' = x_t + \delta$ with the feature extractor h so that we can fool the target model f, i.e., $f(x_t') \neq y^*$. To make sure that the difference between x_t' and x_t is imperceptible, the L_p norm of the perturbation is constrained to be smaller than a threshold, i.e., $\|x_t' - x_t\|_p \leq \epsilon$. The final optimization problem can be formulated as:

$$\begin{aligned} & f(x_t') \neq y^*, \\ s.t.\quad & x_t' = \arg\max_{x_t'} L(h(x_t')), \\ & \|x_t' - x_t\|_p \leq \epsilon. \end{aligned} \tag{1}$$

In the following sections, we will first analyze the shortcomings of conventional self-supervised contrastive learning in training the feature extractor h. Later we give our improvement measure, i.e., the contrastive spectral training method. After training the feature extractor, we use it to craft adversarial samples that are directly applied to attack target models. To this end, we design a suitable attack objective function $L_{efd+cos}$ for the feature extractor h based on the characteristics of the contrastive spectral learning. Finally, we describe the optimization approach for generating the adversarial example x_t' that satisfies the L_p norm constraint.

Contrastive Spectral Training

In this paper, a new deep model training method called contrastive spectral training is proposed based on the existing self-supervised learning methods. Compared with other training methods, this approach is more suitable for training surrogate models for crafting transferable adversarial examples in the cross-domain setting. Besides, this approach can be easily integrated into most self-supervised training methods.

As we mentioned above, because of the no-label property, self-supervised training methods are the preferred choice for training surrogate models. However, we find that the performance obtained by directly using the existing self-supervised contrastive training method is rather mediocre (see Table 4 in our ablation study).

We investigate the various steps of self-supervised training and find that the key to improving cross-domain transferability lies in data augmentation methods. Current augmentation methods used by self-supervised training intensely transform the input images. This causes trained models to have unbalanced shape and texture biases. To balance them, we weaken some texture-related data augmentations. Specifically, we remove data augmentations that have a strong correlation with color information, such as grayscale.

Note that we are *not* suggesting that weakening texture-related augmentations is beneficial to all self-supervised tasks. The best augmentation combination depends on the downstream task (Tian et al. 2020). Therefore, other tasks may not be suitable for using this method.

Besides, we add a spectral pooling layer (Rippel, Snoek, and Adams 2015) before the input layer of the feature extractor. In this layer, we transform the image from the spatial space to the frequency space with Fast Fourier Transform (FFT). After cutting off all high-frequency components, we convert the remaining low-frequency signals back to the spatial space by inverse FFT. In this way, we remove all details but retain image structures and high-level semantic information. Once the training is finished, the spectral pooling layer will be removed from the network architecture.

Figure 1 shows the training procedure of our feature extractor. We first apply different random augmentations to the same training image x to obtain a positive sample pair (x_1, x_2). We feed the sample pair to our feature extractor with spectral pooling layer h^* and that combined with an MLP predictor p, respectively. Then we get four output vectors $y_1 = h^*(x_1)$, $y_2 = h^*(x_2)$, $z_1 = p(h^*(x_1))$, and $z_2 = p(h^*(x_2))$. We maximize the cosine similarity d between the output vectors by only updating the feature extractor combined with an MLP predictor. Specifically, we set the vectors y_1 and y_2 as constants and maximize the following objective function:

$$\frac{1}{2}\left(d\left(z_1, constant(y_2)\right) + d(z_2, constant(y_1))\right) \tag{2}$$

Attack Objective Function

Our work is highly motivated by recent studies of the feature-space attack. In particular, our attack formula is based on a standard transfer-based attack FDA+*fd* (Inkawhich et al. 2020b), which can be written as

$$L_{FDA+fd} = p\left(f\left(x+\delta\right) = y | f_l(x+\delta)\right) + \eta\frac{\|f_l(x+\delta) - f_l(x)\|_2}{\|f_l(x)\|_2}. \tag{3}$$

f_l is the l-th layer representation, and the first term $p(f(x+\delta) = y|f_l(x+\delta))$ measures the f_l's contribution to the final prediction y. To compute this term we should train a small binary auxiliary network on the feature f_l, and one auxiliary network can only compute the probability of one label, which means that if we want to attack K labels we need to train K auxiliary networks.

The second term $L_{fd} = \|f_l(x+\delta) - f_l(x)\|_2 / \|f_l(x)\|_2$ is called the feature disruption term. This term aims to encourage the adversarial example to move far away from the original benign image in the feature space. η is the trade-off hyper-parameter. Because of the heavy computation overhead and the reliance on label information of the FDA term, we discard it and only keep the second term. This term has no relationship with any specific label space or domain information, so it is suitable for the cross-domain scenario.

Besides, in order to make full use of the parameters in the deep neural network, we gather multiple intermediate representations and compute the sum of all L_{fd}s. Considering the general idea of contrastive learning, i.e., positive sample pairs are close to each other while negative sample pairs are far from each other, we add the cosine similarity loss L_{cos} to our attack loss formulation to help the images deviate from their original semantic features. Therefore, our final attack objective function is:

$$L_{efd+cos} = \sum_{l \in \mathscr{L}} \left[\frac{\|h_l(x+\delta) - h_l(x)\|_2}{\|h_l(x)\|_2}\right] + \frac{-h(x)^T h(x+\delta)}{\|h(x)\|_2 \|h(x+\delta)\|_2}, \tag{4}$$

where $\mathscr{L} = \{l_1, ..., l_k\}$ is the set of selected feature layers.

Cross-Domain Transfer-Based Attack

Here we introduce the optimization algorithm for generating adversarial perturbations:

$$\begin{cases} x_t^{(0)} = x_t, \\ x_t^{(k+1)} = \texttt{Clip}\left\{x_t^{(k)} + \alpha \cdot \delta^{(k)}\right\}, \\ \delta^{(k)} = \mathscr{G} * \texttt{Sign}\left(\bigtriangledown L_{efd+cos}\left(T\left(x_t^{(k)}\right)\right)\right), \\ x'_t = x_t^{(N)}, \end{cases} \tag{5}$$

where Clip is a projection operation to ensure that the L_p norm constraint holds. N is the iteration number, and α is the step size.

Outside the Sign function we have a convolution with a 15×15 Gaussian kernel $\mathscr{G}$. This operator comes from Dong et al., who show that the Gaussian kernel can be used to alleviate the issue of interest region shifting across different deep models (Dong et al. 2019). Dong et al. plug the kernel $\mathscr{G}$ inside the Sign function. Here we use it to suppress the high-frequency noise of the generated perturbations and make them smoother, so we move it to the outside of the Sign function.

T is a random linear transformation operation, including random cropping, random padding, and random resizing. This operation introduces extra data to the optimization to alleviate overfitting. Unlike previous methods (Xie et al. 2019), in this formula, there is no transformation probability p. This indicates that the probability of the image being transformed is 1.0.

Experimental Setup

This section introduces the basic settings of experiments, including the selection of datasets, competitors, and the implementation details of CDTA.

Datasets

There are two types of datasets in our experiments: the source domain dataset and the target domain dataset. The source domain dataset refers to the training data of the surrogate model. Generally, for models trained through self-supervised methods, the larger the training set, the better the

Attack	Comic Books				Oxford 102 Flower			
	Res34	DN161	Inc-v3	VGG16bn	Res34	DN161	Inc-v3	VGG16bn
MI-FGSM	32.1±2.2	31.5±1.1	35.1±3.0	42.3±2.6	14.5±1.2	15.7±1.0	15.2±0.8	21.9±0.7
DIM	35.1±2.7	36.4±1.4	36.3±1.5	46.4±1.7	15.2±0.8	17.9±0.7	16.1±0.6	20.9±0.7
TI-DIM	70.3±1.2	72.1±2.1	66.6±2.5	72.2±3.1	12.6±1.2	13.6±1.2	17.1±0.7	19.2±0.7
DR	46.2±2.2	39.8±2.2	43.7±1.9	53.1±2.3	13.3±1.1	15.8±0.8	16.3±0.9	20.6±1.1
SSP	48.9±2.0	53.3±1.4	45.9±1.8	62.5±2.5	12.9±0.7	16.7±0.6	18.5±1.4	25.4±0.7
BIA+RN	58.3±2.0	59.8±2.2	56.3±3.1	72.8±2.8	12.1±0.8	15.7±0.7	16.2±1.3	27.4±1.0
BIA+DA	66.7±2.8	67.8±2.5	59.6±3.2	76.4±0.9	9.8±0.4	18.6±1.4	23.7±1.8	24.1±1.9
CDA	78.1±2.3	75.7±1.5	69.0±2.3	75.0±1.7	**17.3±1.3**	18.0±0.9	21.0±0.9	29.0±0.9
CDTA	**86.0±0.7**	**87.0±2.0**	**75.4±1.7**	**86.8±1.8**	15.8±1.3	**24.8±1.8**	**30.6±1.8**	**40.0±1.7**

Table 1: Transferability comparisons on Comic Books and Oxford 102 Flower classification tasks. We report the performance of each attack method in the form of the *mean ASR (%)±standard deviation* for six repeats. The best results are marked in bold (the higher, the better).

Attack	BIRDS-400				Food-101			
	Res34	DN161	Inc-v3	VGG16bn	Res34	DN161	Inc-v3	VGG16bn
MI-FGSM	22.6±1.3	22.7±1.3	10.6±0.6	17.4±1.3	36.4±2.3	41.5±2.1	47.0±2.4	53.9±1.6
DIM	25.6±1.2	26.5±0.9	13.4±1.7	23.7±0.5	42.5±2.8	47.9±1.7	54.2±2.3	58.5±2.5
TI-DIM	54.7±1.8	46.7±1.6	41.8±2.3	54.9±1.6	74.2±1.5	73.9±2.0	73.9±1.1	71.5±1.4
DR	20.7±0.5	21.6±1.7	6.4±1.6	17.8±2.0	39.3±2.1	47.2±0.5	44.3±1.8	56.8±2.5
SSP	42.4±2.0	41.3±1.3	18.6±1.1	33.7±1.3	53.9±1.7	62.1±1.9	63.4±2.3	77.2±1.2
BIA+RN	49.2±1.3	48.7±1.7	18.6±0.6	53.7±0.5	64.4±2.1	87.9±1.9	82.4±2.4	**96.2±1.3**
BIA+DA	48.3±2.1	46.9±3.1	20.3±1.1	49.9±1.8	74.9±1.6	86.2±1.6	77.6±1.7	95.5±1.2
CDA	58.5±1.8	60.5±1.7	49.5±1.7	61.5±1.5	86.4±1.5	87.6±0.9	88.7±0.5	91.4±1.3
CDTA	**91.2±1.6**	**92.3±0.8**	**54.9±3.1**	**88.2±1.0**	**94.5±1.1**	**94.1±1.1**	**93.8±1.3**	94.6±0.6

Table 2: Transferability comparisons on BIRDS-400 and Food-101 classification tasks. We report the performance of each attack method in the form of the *mean ASR (%)±standard deviation* for six repeats. The best results are marked in bold (the higher, the better).

Attack	Dataset			
	C. B.	102 F.	B. 400	F. 101
MI-FGSM	35.2	16.8	18.3	44.7
DIM	38.5	17.5	22.3	50.7
TI-DIM	70.3	15.6	49.5	73.3
DR	45.7	16.5	16.6	46.9
SSP	52.6	18.3	34.0	64.1
BIA+RN	61.8	17.8	42.5	82.7
BIA+DA	67.6	19.0	41.3	83.5
CDA	74.4	21.3	57.4	88.5
CDTA	**83.8**	**27.8**	**81.6**	**94.2**

Table 3: Transferability comparisons across all classification tasks. Each result in this table means the average ASR (%) over four target models on each dataset. From left to right the datasets are Comic Books, Oxford 102 Flower, BIRDS-400, and Food-101, respectively.

model's performance. ImageNet (Deng et al. 2009) is large and diverse enough, so we choose ImageNet as our source domain dataset.

Our approach should have good results across multiple datasets. To compare with previous methods, we choose four different datasets as our target domains.

- **Oxford 102 Flower.** It consists of 102 flower categories (Nilsback and Zisserman 2008). The flowers are chosen to be flowers commonly occurring in the United Kingdom. Each class consists of between 40 and 258 images. The images have large scale, pose and light variations.
- **Food-101.** It is a challenging dataset of 101 food categories, with a total of 101,000 images (Bossard, Guillaumin, and Van Gool 2014). All images are rescaled to have a maximal side length of 512 pixels.
- **BIRDS-400.** It is a dataset of 400 bird species, with a total of 62,388 images (Gerry 2022). All images are 224 × 224 × 3 color images in the JPG format.
- **Comic Books.** This dataset has 52,156 colored images from 86 classes (Bircanoglu 2017). All images are resized to 288 × 432.

Competitors

We adopt the most recent cross-domain transfer-based attacks as our baselines. Naseer et al. propose an attack method called CDA (Naseer et al. 2019). Specifically, they train a generative model with relativistic cross-entropy and use it to attack other models trained with different domains. Zhang et al. propose BIA (Zhang et al. 2022b). They propose two augmentation methods: random normalization (RN) and domain-agnostic attention (DA) to enhance the generalization capability of the generator used to craft adversarial sam-

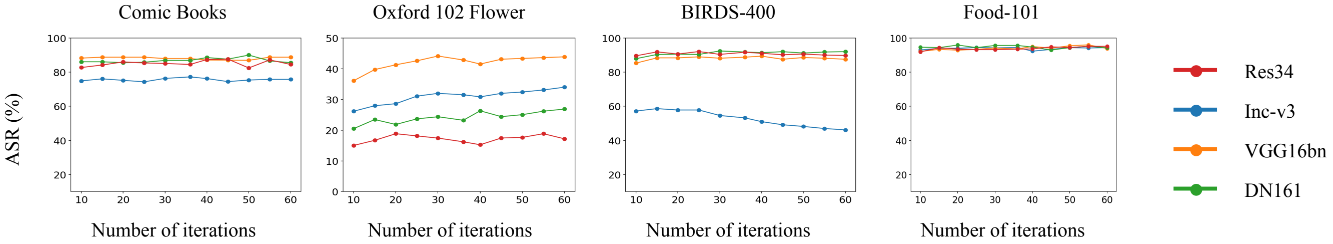

Figure 2: The ASRs (%) of CDTA against target models. The adversarial examples are generated with the number of iterations from 10 to 60.

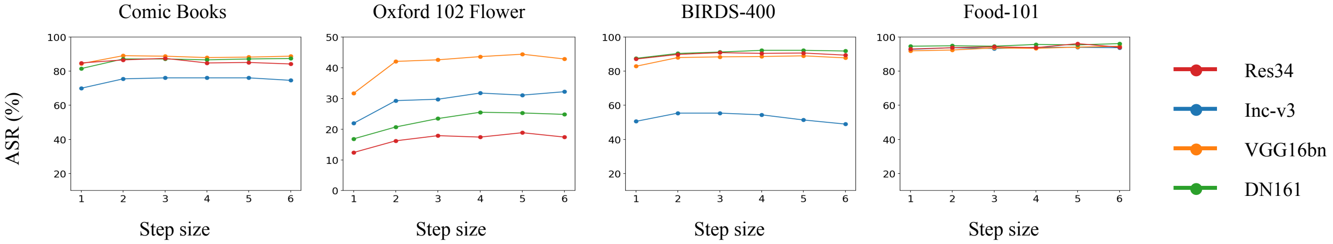

Figure 3: The ASRs (%) of CDTA against target models. The adversarial examples are generated with the step size from 1 to 6.

ples. We use the ImageNet pre-trained generative models released by these works to conduct the experiments. We follow the implementation of CDA to do Gaussian blurring on the generated adversarial examples to further improve their transferability.

We also use the classification model obtained from supervised learning as a surrogate model to compare our approach with the cross-architecture transfer-based attacks. We select state-of-the-art attacks of this kind as our baselines, including MI-FGSM (Dong et al. 2017), DIM (Xie et al. 2019), and TI-DIM (Dong et al. 2019). MI-FGSM and DIM set the step size to 1 and the iteration number to 20, while TI-DIM sets the step size to 1.6 and the iteration number to 10. We use the original output of the surrogate model as the pseudo-label to compute cross-entropy loss.

Two feature-based attack methods, DR (Lu et al. 2020) and SSP (Naseer et al. 2020), are also included as baselines. We follow the default setting of the original paper with the step size equaling 4 and the iteration number equaling 100 for DR. For SSP, the step size equals 2.55, and the iteration number equals 100. All of the cross-architecture transfer-based attacks use the ImageNet pre-trained ResNet-50 (He et al. 2016) classification model as the surrogate model. The pre-trained ResNet-50 is provided by the PyTorch library[1].

Note that we do not compare CDTA with relaxed cross-domain attacks, such as FDA+$xent$ (Inkawhich et al. 2020a) and No-box Attack (Li, Guo, and Chen 2020). This is because for these methods to work properly, they need to have access to information about the target domain, which is strictly prohibited in this work.

Implementation

The backbone architecture of the feature extractor is ResNet-50. The training configurations of the feature extractor h follow SimSiam (Chen and He 2021). We take the output of the first, the second, and the third bottleneck groups, and the output of the whole feature extractor as the target layers of CDTA.

For each target domain, we train a Res34 (ResNet-34), a DN161 (DenseNet-161) (Huang et al. 2017), an Inc-v3 (Inception-v3) (Szegedy et al. 2016), and a VGG16bn (VGG16 with batch normalization) (Simonyan and Zisserman 2015). All models are trained from scratch instead of being fine-tuned based on the ImageNet pre-trained weights. So the parameters of these models are completely independent of ImageNet. The input size of Inc-v3 is $3 \times 299 \times 299$, while the others are $3 \times 224 \times 224$.

Experimental Results

Cross-Domain Attack Results

We first investigate the effectiveness of CDTA compared with competitors. We conduct experiments with 16 different models on four datasets. We set the maximum perturbation to 16 for all experiments with pixel values in [0, 255]. We set the iteration number of CDTA as 30 and the step size as 4. For each target model, we randomly choose 500 images from the *test* set of target domains and calculate the ASR corresponding to each attack method. To test the stability of each attack method, we repeatedly attack each target model 6 times with random initialization and different images. We report both the mean and standard deviation of ASRs for each target model.

As shown in Table 1 and 2, our method can consistently and substantially improve the ASR (attack success rate)

[1] https://pytorch.org/hub/pytorch_vision_resnet/

Target model		ST	SST	CDTA
C. B.	Res34	70.2±2.3	25.3±2.4	**86.0±0.7**
	DN161	72.2±2.2	28.0±1.9	**87.0±2.0**
	Inc-v3	58.5±2.8	25.7±2.6	**75.4±1.7**
	VGG16bn	65.6±1.9	25.3±2.1	**86.8±1.8**
102 F.	Res34	9.8±0.6	6.8±1.2	**15.8±1.3**
	DN161	8.5±1.2	6.1±0.7	**24.8±1.8**
	Inc-v3	13.6±1.4	7.7±0.4	**30.6±1.8**
	VGG16bn	13.2±1.0	6.2±1.0	**40.0±1.7**
B. 400	Res34	71.1±1.8	19.3±2.4	**91.2±1.6**
	DN161	69.3±2.1	21.0±1.5	**92.3±0.8**
	Inc-v3	47.4±2.0	7.6±1.0	**54.9±3.1**
	VGG16bn	71.2±1.3	23.1±1.9	**88.2±1.0**
F. 101	Res34	80.8±1.3	39.4±3.2	**94.5±1.1**
	DN161	77.9±0.9	41.2±1.6	**94.1±1.1**
	Inc-v3	79.5±2.5	46.4±2.9	**93.8±1.3**
	VGG16bn	79.6±1.8	36.5±1.3	**94.6±0.6**

Table 4: Cross-domain ASRs (%) when using differently trained surrogate models. "ST" uses a classification model obtained from supervised training. "SST" uses the model trained by SimSiam, a self-supervised training method. "CDTA" uses the model trained by contrastive spectral training, the method proposed in this paper. From up to bottom the datasets are Comic Books, Oxford 102 Flower, BIRDS-400, and Food-101, respectively.

compared with the state-of-the-art baselines. For example, if the target model is Res34, CDA achieves the second-best ASR of 78.1% on the Comic Books dataset, while our CDTA can effectively improve it to **86.0%**. On 14 out of 16 target models, our method outperforms the other state-of-the-art methods. We lag slightly behind the best results in the other two cases. For example, for Res34 on the Oxford 102 Flower dataset, CDA has the best ASR of 17.3%, while we achieve 15.8%.

Table 3 is a summary of Table 1 and Table 2. In terms of the mean ASR over four target models on each dataset, our proposed approach improves the state-of-the-art baselines by **9.35%, 6.50%, 24.17%, and 5.79%** on the Comic Books, Oxford 102 Flower, BIRDS-400, and Food-101 datasets, respectively.

Ablation Study

In this section, we show the experimental results of the proposed approach with different choices of hyper-parameters, i.e., the number of iterations N and the step size α. We also investigate the impact of other surrogate models on the attack success rate of cross-domain attacks.

In Figure 2, we attack the 16 target models with different iteration numbers that vary from 10 to 60. From the result, we can see that the ASR does not always increase with the number of iterations. When attacking Inc-v3 on the Comic Books and Food-101 datasets, large iteration numbers can even hurt cross-domain transferability. Therefore, we set the iteration number of CDTA to 30 to obtain balanced attack performances across different datasets.

Figure 3 shows the relationship between ASR and the step size. Similar to Figure 2, the ASR increases with the step size when the step size is small. However, it decreases when the step size is too large. This can be explained by the fact that if the step size is too small, the perturbation may fall into a local optimum. However, when the step size is too large, it is not conducive to the convergence of the loss function, and thus not good for the results. Therefore, a value that is not too big or too small (e.g., in this work) should be chosen as the step size.

We also investigate the effect of differently trained surrogate models on ASRs. The results are shown in Table 4. In the table, we compare the cross-domain attack performances when using different surrogate models, including the classification model pre-trained using the ImageNet dataset, the encoder trained by SimSiam (Chen and He 2021) (a self-supervised training method), and the feature extractor trained by the proposed contrastive spectral training method. We find that the encoder trained with SimSiam generally obtains *significantly lower performance* than the other surrogate models. Using the encoder trained with SimSiam cannot even beat using surrogate models trained by supervised training methods. In contrast, ASRs can be improved substantially using the method proposed in this paper.

The results show that attacking the feature extractor obtained by conventional self-supervised contrastive training cannot generate highly transferable adversarial samples in cross-domain settings. We think that the reason is that the deep encoder trained by SimSiam has unbalanced shape and texture biases. In contrast, our proposed method can balance the shape and texture biases of the trained feature extractor by weakening some texture-related augmentations.

Conclusion

This paper proposes a Cross-Domain Transfer-Based Attack (CDTA) method to generate highly transferable adversarial examples in the most strict black-box setting, i.e., the strict cross-domain setting. Specifically, we propose a novel contrastive spectral training method to train a surrogate model. After that, CDTA crafts adversarial examples by destroying the intermediate features and final output representations of the trained feature extractor. We conduct extensive experiments to compare our CDTA with both state-of-the-art cross-domain attacks and cross-architecture attacks. Experimental results of 16 target deep learning models on four datasets show that our approach can consistently outperform the state-of-the-art competitors by a large margin (11.45% on average). We believe that our CDTA can serve as a strong benchmark to evaluate the robustness of deep learning models in the strict black-box setting.

Acknowledgments

We would like to thank the anonymous reviewers for their detailed and valuable comments. This work was supported by the National Key R&D Program of China (Grant No.2020YFB1707603), the National Natural Science Foundation of China (Grant No. 62206318) and the Fundamental Research Funds for the Central Univerisities, Sun Yat-sen University (Grant No. 22QNTD1201).

References

Akhtar, N.; and Mian, A. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6: 14410–14430.

Bircanoglu, C. 2017. Comic Books Images. https://www.kaggle.com/datasets/cenkbircanoglu/comic-books-classification. Accessed: 2022-08-12.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101 – Mining Discriminative Components with Random Forests. In *European Conference on Computer Vision*.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, Y.; Liao, F.; Pang, T.; Hu, X.; and Zhu, J. 2017. Discovering adversarial examples with momentum. arXiv:1710.06081.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4312–4321.

Ganeshan, A.; BS, V.; and Babu, R. V. 2019. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8069–8079.

Gerry. 2022. BIRDS 400 - SPECIES IMAGE CLASSIFICATION. https://www.kaggle.com/datasets/gpiosenka/100-bird-species. Accessed: 2022-08-12.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Guo, C.; Gardner, J.; You, Y.; Wilson, A. G.; and Weinberger, K. 2019. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, 2484–2493. PMLR.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Inkawhich, N.; Liang, K.; Wang, B.; Inkawhich, M.; Carin, L.; and Chen, Y. 2020a. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *Advances in Neural Information Processing Systems*, 33: 20791–20801.

Inkawhich, N.; Liang, K. J.; Carin, L.; and Chen, Y. 2020b. Transferable Perturbations of Deep Feature Distributions. In *International Conference on Learning Representations*.

Inkawhich, N.; Liang, K. J.; Zhang, J.; Yang, H.; Li, H.; and Chen, Y. 2021. Can Targeted Adversarial Examples Transfer When the Source and Target Models Have No Label Space Overlap? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 41–50.

Kumar, C.; Ramesh, J.; Chakraborty, B.; Raman, R.; Weinrich, C.; Mundhada, A.; Jain, A.; and Flohr, F. B. 2021. Vru pose-ssd: Multiperson pose estimation for automated driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15331–15338.

Li, K.; Yu, L.; Wang, S.; and Heng, P.-A. 2020. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, 775–783.

Li, Q.; Guo, Y.; and Chen, H. 2020. Practical no-box adversarial attacks against dnns. *Advances in Neural Information Processing Systems*, 33: 12849–12860.

Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; and Tang, J. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Lu, Y.; Jia, Y.; Wang, J.; Li, B.; Chai, W.; Carin, L.; and Velipasalar, S. 2020. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 940–949.

Naseer, M.; Khan, S.; Hayat, M.; Khan, F. S.; and Porikli, F. 2020. A Self-supervised Approach for Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Naseer, M. M.; Khan, S. H.; Khan, M. H.; Shahbaz Khan, F.; and Porikli, F. 2019. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32: 12905–12915.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

Rippel, O.; Snoek, J.; and Adams, R. P. 2015. Spectral representations for convolutional neural networks. *Advances in neural information processing systems*, 28.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Bengio, Y.; and LeCun, Y., eds., *International Conference on Learning Representations*.

Singh, R.; Agarwal, A.; Singh, M.; Nagpal, S.; and Vatsa, M. 2020. On the robustness of face recognition algorithms against attacks and bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13583–13589.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In Bengio, Y.; and LeCun, Y., eds., *International Conference on Learning Representations*.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33: 6827–6839.

Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; and Ren, K. 2021. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7639–7648.

Wu, W.; Su, Y.; Chen, X.; Zhao, S.; King, I.; Lyu, M. R.; and Tai, Y.-W. 2020a. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1161–1170.

Wu, W.; Su, Y.; Chen, X.; Zhao, S.; King, I.; Lyu, M. R.; and Tai, Y.-W. 2020b. Towards global explanations of convolutional neural networks with concept attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8652–8661.

Wu, W.; Su, Y.; Lyu, M. R.; and King, I. 2021. Improving the transferability of adversarial samples with adversarial transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9024–9033.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2730–2739.

Zhang, J.; Wu, W.; Huang, J.-t.; Huang, Y.; Wang, W.; Su, Y.; and Lyu, M. R. 2022a. Improving Adversarial Transferability via Neuron Attribution-Based Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14993–15002.

Zhang, Q.; Li, X.; Chen, Y.; Song, J.; Gao, L.; He, Y.; and Xue, H. 2022b. Beyond ImageNet Attack: Towards Crafting Adversarial Examples for Black-box Domains. In *International Conference on Learning Representations*.