Evaluating Transfer-based Targeted Adversarial Perturbations against
Real-World Computer Vision Systems based on Human Judgments

Zhengyu Zhao, Nga Dang, Martha Larson
Radboud University, Nijmegen, Netherlands
z.zhao@cs.ru.nl, nga.dangthanhnga@student.ru.nl, m.larson@cs.ru.nl

Abstract

Computer vision systems are remarkably vulnerable to adversarial perturbations. Transfer-based adversarial images are generated on one (source) system and used to attack another (target) system. In this paper, we take the first step to investigate transfer-based targeted adversarial images in a realistic scenario where the target system is trained on some private data with its inventory of semantic labels not publicly available. Our main contributions include an extensive human-judgment-based evaluation of attack success on the Google Cloud Vision API, and additional analysis of the different behaviors of Google Cloud Vision in face of original images vs. adversarial images. Resources are publicly available at https://github.com/ZhengyuZhao/Targeted-Transfer/blob/main/google_results.zip.

1. Introduction

Adversarial image perturbations that are optimized with respect to one (known) model, called the source model, are known to have the ability to transfer their fooling effects to another (unknown) model, called the target model [7, 8]. Such transfer-based adversarial images are especially critical in the real-world adversarial scenarios since they do not require the specific knowledge of the target model. Most existing evaluation of transfer-based adversarial images has been limited to the closed-world assumption that the source and target models are trained on some disciplined benchmark dataset (e.g. ImageNet) [1, 4, 6, 10, 12].

However, this assumption might not hold in real-world scenarios where the target system is trained on some private data and does not usually make its inventory of semantic labels publicly available. For this reason, it remains unclear whether the targeted adversarial images that have been optimized towards a specific target class w.r.t the label set of the white-box source model can actually succeed on the target model, i.e. fooling the target model into predicting a semantically similar target class. To help answer this question, in this paper, we take the first step to study a realistic transfer-based attack scenario that abandons this unrealistic assumption. Specifically, we generate targeted adversarial images to fool both the object detection and label detection services of the Google Cloud Vision API\(^1\). Especially, we resort to human judgments in order to evaluate attack success since the semantic label set of the target model, Google Cloud Vision, is unknown.

Our results reveal the vulnerability of Google Cloud Vision API to transfer-based adversarial image perturbations, and additional analysis provides new insights into the se-
semantic pattern of API predictions on original vs. adversarial images. For reproducibility, all our human judgments will be made publicly available should the paper be accepted.

2. Transfer-Based Adversarial Images

Transfer-based adversarial images have been extensively studied in the non-targeted setting [1, 2, 10] as well as the more challenging, targeted setting [6, 12]. We focus our work on the iterative attacks that generate targeted adversarial images using pre-trained ImageNet classifiers as source models without additional data and model training.

We compare three attacks with different loss functions:

- **CE**, the widely used cross-entropy loss [5].
- **Po+Trip**, an effective loss for targeted transferability, based on Poincaré distance and Triplet loss [6].
- **Logit**, the state-of-the-art loss for targeted transferability, by solely maximizing the target logit value [12].

For each of these three attacks, a combination of the following three widely-used transfer techniques is applied:

- **MI-FGSM**, using a momentum term to accumulate previous gradients for more accurate updating [1].
- **TI-FGSM**, applying random translation to augment original images for preventing the attack optimization from overfitting to the white-box source model [2].
- **DI-FGSM**, applying random resizing and padding for input augmentation, but also varying the augmentation parameters over iterations [10].

3. Experiments

In this section, we report the results of the three transfer-based attacks as presented in Section 2 on the two widely used services of the Google Cloud Vision API: object detection and label detection. The evaluation is specifically made possible by our human judgments.

We used the NIPS 2017 ImageNet-Compatible Dataset\(^2\), which consists of 1000 \(299 \times 299\) images. Each image is associated with one of the 1000 ImageNet ground truth labels and one randomly assigned target label used for attacking. Due to the time overhead for human judgments, we selected the first 400 images from the dataset for our experiments. These 400 original images involve 266 unique ground truth labels and 310 unique target labels.

Following common practices, the perturbations were restricted by \(L_\infty\) norm with \(\epsilon = 16\), and the step size of the attack optimization was set as 2. To make sure all attacks can converge, we applied 300 iterations, following [12]. In order to further boost transferability, an ensemble of four pre-trained ImageNet classifiers (ResNet-50, DenseNet-121, VGGNet-16, and Inception-v3) is used as the source model.

3.1. Evaluation based on Human Judgments

We measure both the targeted and non-targeted success of our three considered attacks. We make use of human judgments to determine whether the prediction of the classifier on an image should be considered to be different from the image’s ground truth (for non-targeted success) or the same as the target class (for targeted success). Human judgments are necessary because the semantic label set of Google Cloud Vision does not match the 1000 ImageNet classes, which are used by the source models. The human judgments involve determining in which cases two classes should be considered semantically similar enough to be treated as the same class.

We collected prediction results by the simple process of taking screenshots of the Google Cloud Vision interface. In each session, an image in its 4 (1 original and 3 adversarial) versions were uploaded, and then 8 screenshots of prediction results (for both object detection and label detection) were captured. In total, for all the 400 different sessions of images, 3200 screenshots were gathered. After optimization, it took about 4 minutes for each session, which amounted to 27-hour manual work of one student.

For both services, the prediction list consists of semantic classes along with confidence scores. Note that the confidence score here is not a probability (which would sum to one). Specifically, at most 10 classes with high confidence score (\(\geq 50\%\)) are returned, and we did not limit our measure of success rates to only top-1 prediction. Instead, a non-targeted success is achieved when the ground-truth class does not appear in the top-10 predictions, while a targeted success is achieved when the target class appears.

3.2. Results on Attack Success

The detailed attack results are reported in Table 1. We can observe that as expected, it is much more difficult for an attack to achieve targeted success than non-targeted success. When comparing different attacks, we can observe

---

\(^2\)Publicly available at [this link](https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.1.0/examples/nips17_adversarial_competition/dataset).
Table 2. Five most frequently occurred objects that cause misclassification due to their multiple occurrences in an image.

| Object                | Frequency |
|-----------------------|-----------|
| Person Animal Luggage and bags Shoe wheel | 193 186 80 66 55 |

Table 3. Quantity of predictions per image on different image sets (Ori, CE, Po+Trip, Logit) for both object detection and label detection services.

| Services         | Ori | CE | Po+Trip | Logit |
|------------------|-----|----|---------|-------|
| Object detection | 2.58| 2.92 | 2.68   | 3.37  |
| Label detection  | 9.98| 9.85| 9.83    | 9.86  |

Table 4. Diversity of predictions on different image sets in terms of the total number of unique predictions.

| Services | All | Ori | CE | Po+Trip | Logit |
|----------|-----|-----|----|---------|-------|
| Object detection | 192 | 161 | 82 | 98      | 83    |
| Label detection  | 1333 | 1162 | 730 | 771     | 1008   |

that the Logit attack achieved the best performance in all cases. This is consistent with the finding in [12]. When comparing the two services, we can observe that it is much easier to mislead the object detection than label detection.

Note that for object detection, even the original image set has caused substantial misclassification (31.50%). It is mainly because in many cases, some objects (e.g., person) occur multiple times in a single image, making the given ground-truth object out of the returned top-10 predictions. Table 2 reports the frequencies of the five most frequently occurred objects. Such failure caused by multiple occurrences of objects is understandable because during the collection of ImageNet, the annotators were not asked to identify all the objects in an image, but rather to validate whether a specific candidate object appears without knowledge of other objects [9].

3.3. Predictions on Original vs. Adversarial Images

In this subsection, we provide additional analysis on the semantic patterns of predictions given by Google Cloud Vision in face of original vs. adversarial images.

**Quantity of predictions.** We first looked at the coarse-grained impact of different attacks on the predictions. To this end, we calculated the number of the returned predictions per image for original vs. adversarial images. As can be seen from Table 3, for object detection, adversarial images yielded generally more predictions than original images, with the Logit attack the most. In contrast, all results for label detection are very similar, and close to the maximum returned number, 10. This finding implies that label detection is more robust to adversarial attacks than object detection from a coarse-grained perspective.

**Diversity of predictions.** Table 4 reports the total number of unique predictions for original vs. adversarial images. As can be seen, for both two services, the predictions became less diverse after attacking. This result is somewhat unexpected since there were 310 unique target labels (for the adversarial images), which are more than the 266 unique ground truth labels (for the original images). This new finding implies that adversarial attacks were actually trying to force the classifier towards predicting certain classes, and...
this has led us to further looking into the actual distribution of all unique predictions. **Distribution of predictions.** Fig. 2 visualizes the distributions of all unique predictions for original vs. adversarial images. We can observe that the relative frequency of different classes for these two image sets are substantially different. We further visualize the most frequent predictions in these two image sets in Fig. 3. As can be seen, for object detection, the two image sets have different dominant classes (“person” vs. “animal”). Differently, for label detection, “animal” is not among the top predictions.

Interestingly, the added perturbations themselves have lead to the occurrence of some specific classes that typically have tiled textures and colors, such as “art” and “painting”. This observation is also consistent with recent findings that adversarial images tend to be misclassified into some dominant classes (e.g. “brain coral”) that share close semantics with random perturbation patterns [11].

## 4. Conclusion and Discussion

In this paper, we have investigated the impact of transfer-based adversarial attacks against a real-world system, the Google Cloud Vision API. Our investigation improves on existing studies by considering a more realistic scenario in which the semantic label set of the target model is unknown. We have proposed to rely on human judgments to evaluate the attack success in such a scenario, and our results demonstrate that Google Cloud Vision, especially its object detection service, is substantially vulnerable to existing transfer-based attacks that are simple to implement.

Additional analysis of the semantic patterns of predictions provides interesting insights into understanding the different behaviors of Google Cloud Vision in face of original vs. adversarial images. Specifically, we have identified differences in terms of the quantity, diversity, and distribution of predictions. We conjecture that these differences might be caused by specific properties of the training data, or by design for specific purposes, for example, outputting certain labels under conditions with low certainty. Investigations on how and why these differences exist would be promising for future work. In general, our work supports the recent trend towards documentation of datasets in order to address mismatches between data characteristics in the data creation and those in real-world use scenarios [3].

As we are the first to provide human-related insights into the impact of transfer-based attacks on real-world computer vision systems, our human evaluation might not be optimal. For example, in order to achieve more reliable analysis, we expect future research to involve more participants and conduct inter-participant variability testing.

In sum, this paper lays the groundwork for future research that gains deeper insight into transfer-based adversarial attacks. Moving forward, researchers should consider human perspective on the change in the semantics of the prediction caused by adversarial images and should carry out analysis of the semantic patterns of predictions, in order to understand the impact of an adversarial attack at the system level.

## References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In **CVPR**, 2018. 1, 2

[2] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In **CVPR**, 2019. 2

[3] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64:86–92, 2021. 4

[4] Nathan Inkawhich, Kevin J Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. In **NeurIPS**, 2020. 1

[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In **ICLR**, 2017. 2

[6] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In **CVPR**, 2020. 1, 2

[7] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In **ICLR**, 2017. 1

[8] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, 2016. 1

[9] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In **ICML**, 2020. 3

[10] Chihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In **CVPR**, 2019. 1, 2

[11] Chaoning Zhang, Philipp Benz, Adil Karjaou, and In So Kweon. Data-free universal adversarial perturbation and black-box attack. In **ICCV**, 2021. 4

[12] Zhengyu Zhao, Zhioran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. In **NeurIPS**, 2021. 1, 2, 3