Enhancing targeted transferability via suppressing high-confidence labels

Hui Zeng, Tong Zhang, Biwei Chen, and Anjie Peng

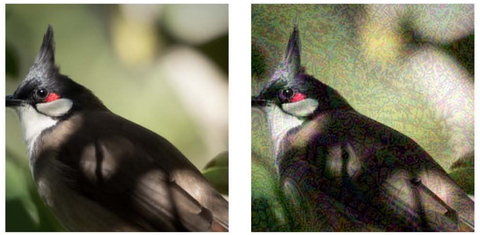
BACKGROUND

A targeted attack misguides a CNN model to whatever the attacker intends, i.e., $F(\boldsymbol{I}') = y_t$, where y_t is the target label. Targeted transferability is much more challenging than its untargeted counterpart:

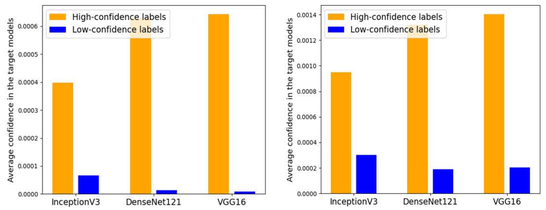
- It requires the output of an unknown model to be a specific label;
- Gradient vanishing;
- Original label restoration.

Hence, besides the widely-studied enhancements for untargeted transferability, tailored schemes are required for targeted transferability.

MOTIVATION

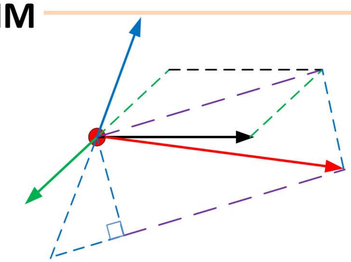


The original label of an AE is likely to be recovered when it is transferred to a target model. (L) the original image, whose label is '*bulbul*', (R) adversarial image, whose confidence with respect to '*bulbul*' in the source model is 0.0, and 0.87 in the target model.



High-confidence labels in the source model are likely to retain high confidence in the target model. (L) the original image, (R) adversarial image. The high/low-confidence labels are calculated on a pretrained ResNet50 model (source model).

ALGORITHM



The gradient with respect to the target label
The gradient with respect to the original label
The gradient with respect to other high-confidence labels
Our used gradient

The proposed method can be illustrated with the figure on the top.

1) Targeted attack (to y_t) while suppressing y_o.

-Mounting attack with the loss function $L_{combine} = -l_t(\boldsymbol{I}') + \beta_1 l_o(\boldsymbol{I}')$, and obtain an intermediate image $\boldsymbol{I}^{inter}$;

2) Suppressing high-confidence labels.

-Calculating high-confidence labels from $\boldsymbol{I}^{inter}$; Mounting attack with the gradient $\nabla L_{combine} + \beta_2 \nabla L_{combine}\perp$, where $\nabla L_{combine}\perp$ denotes the orthogonal component of $\nabla(l_{high-conf}(\boldsymbol{I}'))$ to $\nabla L_{combine}$.

In this way, we push the AEs away from high-confidence labels while approaching the target one.

CONCLUSION

1) High-confidence labels in the source model are likely to retain high confidence in the target model. Explicitly suppressing them in attack helps with targeted transferability.

2) Targeted transferability has its uniqueness and is not a simple extension of untargeted transferability.

Contacts

zengh5@mail2.sysu.edu.cn

github.com/zengh5/Transferable_targeted_attack

RESULTS

Dataset: ImageNet-compatible dataset.

Pretrained models: Inceptionv3, ResNet50, DenseNet121, and VGG16.

Competitors: CE, Po+Trip [1], Logit [2], and TTP [3].

[1] M. Li, et al., "Towards transferable targeted attack," *CVPR*, 2020.

[2] Z. Zhao, et al., "On success and simplicity: a second look at transferable targeted attacks," *NeurIPS*, 2021.

[3] M. Naseer, et al. "On generating transferable targeted perturbations," *ICCV*, 2019.

Table 1. Targeted transfer success rate (%) in the single-model, random-target scenario.

Attack	Source Model: Res50			Source Model: Dense121			Source Model: Inc-v3		
	→Inc-v3	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Res50	→Dense121	→VGG16
CE	3.9	44.9	30.5	2.3	19.0	11.3	1.8	2.1	1.5
Po+Trip	7.1	57.5	36.3	2.5	15.2	9.2	1.7	3.3	1.6
Logit	9.1	70.0	61.9	7.8	42.6	37.1	**2.4**	3.6	**2.2**
Proposed	**9.6**	**74.9**	**63.5**	**8.7**	**48.1**	**40.5**	2.3	**4.5**	**2.2**

Table 2. Targeted transfer success rate (%) in the single-model, most difficult-target scenario.

Attack	Source Model: Res50			Source Model: Dense121			Source Model: Inc-v3		
	→Inc-v3	→Dense121	→VGG16	→Inc-v3	→Res50	→VGG16	→Res50	→Dense121	→VGG16
CE	1.3	25.8	15.0	1.2	6.5	3.6	1.8	4.2	2.3
Po+Trip	2.8	40.5	20.5	0.9	6.1	2.5	2.4	4.1	2.7
Logit	3.6	51.6	38.6	3.5	22.7	18.3	3.8	**5.5**	3.2
Proposed	**4.0**	**54.5**	**41.6**	**4.0**	**24.5**	**21.2**	**4.0**	4.9	**3.4**

Targeted transfer success rate of TTP vs. iterative attacks, averaged over 3 models and 10 target classes. Source model: Res50.