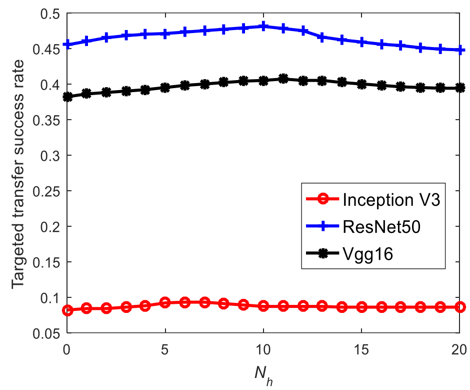
Enhancing targeted transferability via suppressing high-confidence labels: supplementary material

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

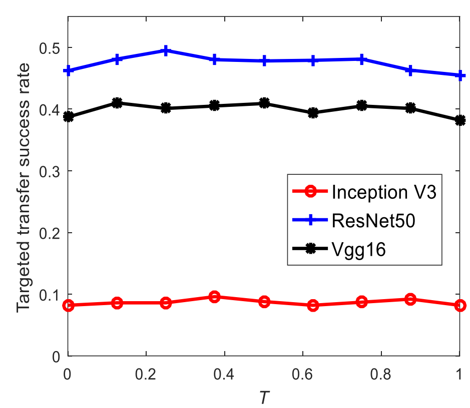
In this supplementary document, we provide the ablation study on N_h and T.

1) **Influence of the number of high-confidence labels**. We study the influence of N_h on the transfer success rate in the single-model, random-target transfer scenario, with the source model fixed as DenseNet121. As shown in Fig. 1(a), the optimal N_h values for different source-target model pairs are between 5 and 15. This can be explained as follows. A very small N_h cannot suppress enough high-confidence labels. On the other hand, the gradients associated with different labels may contradict each other when N_h is large. Hence, there is a trade-off in setting N_h. In our study, we set $N_h = 10$. It is also observed from Fig. 1(a) that the choice of $N_h = 10$ is strictly better than that of $N_h = 0$, which verifies the necessity of suppressing high-confidence labels. Note $N_h = 0$ means suppressing the original label y_o only (Eq. (7) of the paper).

2) **Influence of the timing T of introducing the orthogonal gradient**. Next, we study the influence of T on the transfer success rate in the attack scenario as above. In this experiment, we fix the other parameters of the proposed method and let T vary from 0 to 1. As shown in Fig. 1(b), the optimal T varies between different target models. The main takeaway is that the orthogonal gradient should not be introduced at the attack's very beginning or end. This phenomenon can be explained as follows. On the one hand, the high-confidence labels calculated at the beginning of the attack differ significantly from the high-confidence labels of the final AEs. On the other hand, $T = 1$ means the orthogonal gradient is not introduced. Based on the above considerations, we set $T = 0.75$ in this study.



(a)



(b)

Fig. 1. Ablation study on N_h and T. The source model is a pretrained DenseNet121 model. (a) Targeted transfer success rate as a function of N_h. (b) Targeted transfer success rate as a function of T.