SUPPLEMENTARY FILE
Transferable Clean-Label Poisoning Attacks on Deep Nerual Nets

1. Proof of Proposition 1

$2 \implies 1$

For multi-class problems, the condition for $\phi(\boldsymbol{x})$ to be classified as ℓ_p is

$$\boldsymbol{w}_{\ell_p}^\top \phi(\boldsymbol{x}) + b_{\ell_p} > \boldsymbol{w}_i^\top \phi(\boldsymbol{x}) + b_i, \quad \text{for all } i \neq \ell_p.$$

Each of these constraints is linear, and is satisfied by a convex half-space. The region that satisfies all of these constraints in an intersection of convex half-spaces, and so is convex. Under condition (2), $\phi(\boldsymbol{x}_t)$ is a convex combination of points in this convex region, and so $\phi(\boldsymbol{x}_t)$ is itself in this convex region.

$1 \implies 2$

Suppose that (1) holds. Let

$$\mathcal{S} = \{\sum_i c_i \phi(\boldsymbol{x}_p^j) | \sum_i c_i = 1, 0 \leq c_i \leq 1\}$$

be the convex hull of the points $\{\phi(\boldsymbol{x}_p^j)\}_{j=1}^k$. Let $\mathbf{u}_t = \arg\min_{\mathbf{u} \in \mathcal{S}} \|\mathbf{u} - \phi(\boldsymbol{x}_t)\|$ be the closest point to $\phi(\boldsymbol{x}_t)$ in $\mathcal{S}$. If $\|\mathbf{u}_t - \phi(\boldsymbol{x}_t)\| = 0$, then (2) holds and the proof is complete. If $\|\mathbf{u}_t - \phi(\boldsymbol{x}_t)\| > 0$, then define the classifier function

$$f(\mathbf{z}) = (\mathbf{u}_t - \phi(\boldsymbol{x}_t))^\top (\mathbf{z} - \mathbf{u}_t).$$

Clearly $f(\phi(\boldsymbol{x}_t)) < 0$. By condition (1), there is some j with $f(\phi(\boldsymbol{x}_p^j)) < 0$ as well. Consider the function

$$g(\eta) = \frac{1}{2}\|\mathbf{u}_t + \eta(\phi(\boldsymbol{x}_p^j) - \mathbf{u}_t) - \phi(\boldsymbol{x}_t)\|^2.$$

Because $\mathbf{u}_t$ is the closest point to $\phi(\boldsymbol{x}_t)$ in $\mathcal{S}$, and g is smooth, the derivative of g with respect to η, evaluated at $\eta = 0$, is 0. We can write this derivative condition as

$$g'(0) = (\mathbf{u}_t - \phi(\boldsymbol{x}_t))^\top (\phi(\boldsymbol{x}_p^j) - \mathbf{u}_t) = f(\phi(\boldsymbol{x}_p^j)) \geq 0.$$

However this statement is a contradiction, since $f(\phi(\boldsymbol{x}_p^j)) < 0$.

2. Comparison of Validation Accuracies

To make data poisoning attacks undetectable, in addition to making the perturbations to nonobvious, the accuracy of the model fine-tuned on the poisoned dataset shall not drop too significantly, compared with fine-tuning on the same (except for the poisons) clean dataset. Figure 1 shows that the drop in accuracy is indeed not obvious.

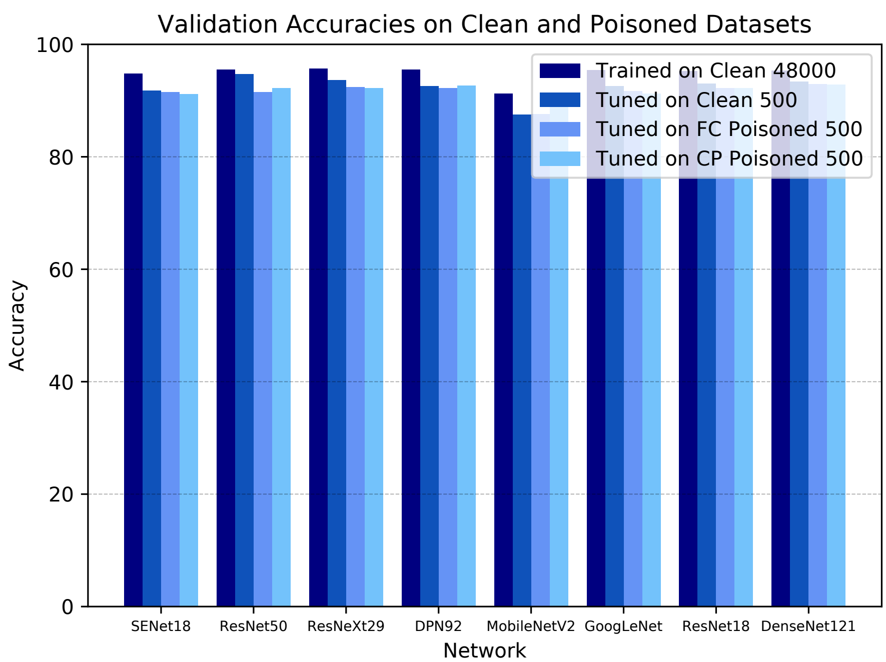

Figure 1: Accuracies on the whole CIFAR10 test for models trained or fine-tuned on different datasets. The fine-tuned models are initialized with the network trained on the first 4800 images of each class. There is little accuracy drop after fine-tuned on the poisoned datasets, compared with fine-tuning on the clean 500-image set directly.

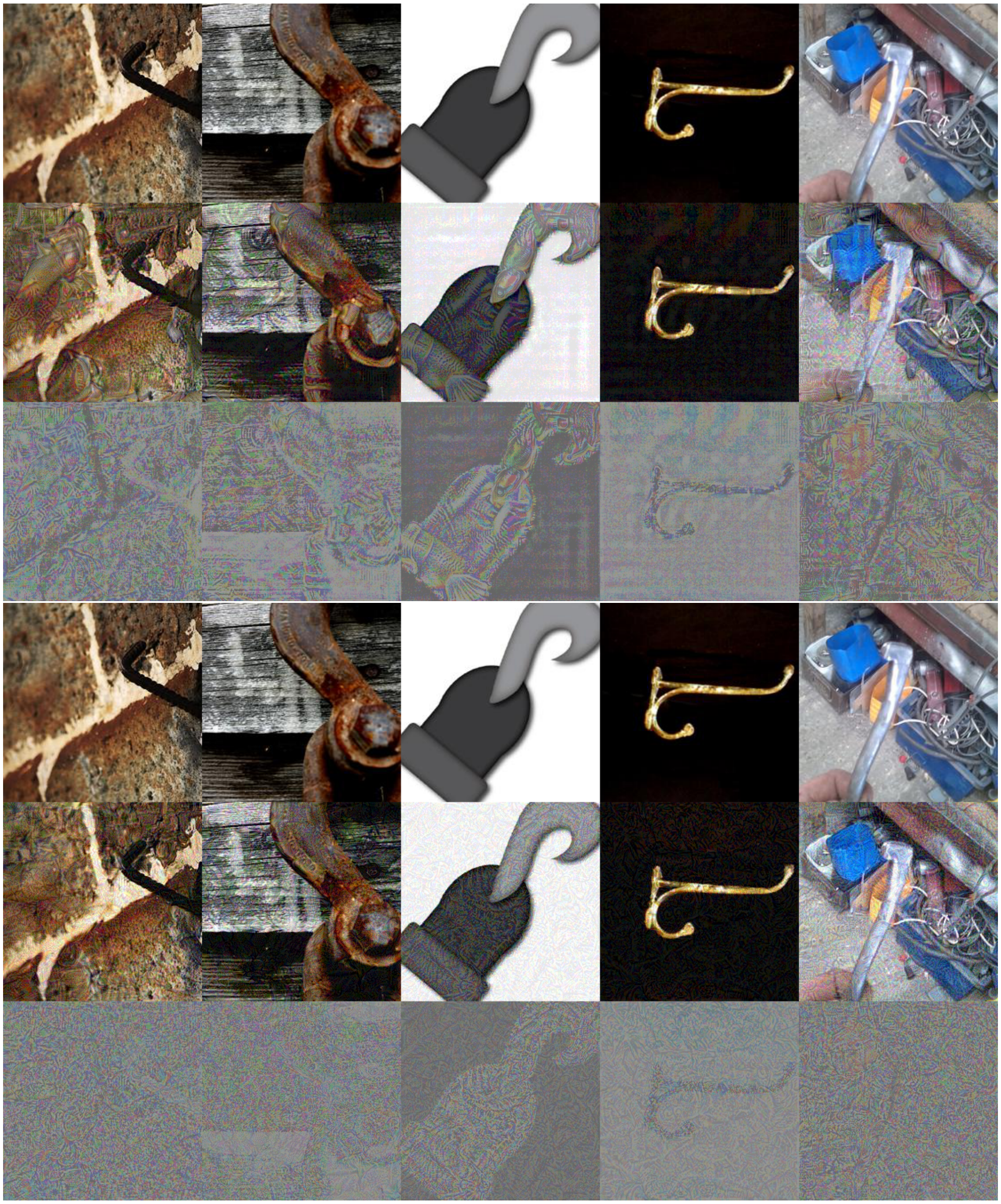

Figure 2: All the 5 poison images.

3. Details of the qualitative example

Both the target *fish* image and the five *hook* images used for crafting poisons come from the WebVision [1] dataset, which has the same taxonomy as the ImageNet dataset. Figure 2 gives all the five poison examples.

References

[1] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.