Supplementary Materials: Zero-Shot Attribute Attacks on Fine-Grained Recognition Models

Nasim Shafiee and Ehsan Elhamifar

Northeastern University, Boston, USA
{shafiee.n,e.elhamifar}@northeastern.edu

1 More Details on CAUP Performance

In the main paper, we investigated the effect of CAUP ℓ_2-norm attack on seen and unseen accuracies and compared it against UAP [2] and GAP [1]. Cd-UAP [3] is another important baseline which generates class-wise UAPs and can not handle zero-shot scenarios(unseen classes). Hence, it is not applicable as a baseline for a fair comparison.

Figure 1 shows the fooling percentage as a function of ℓ_2-norm perturbation for different attacks (this figure is a detailed version of Figure 2 at the main paper). From the granularity perspective, we divided the attacks to two categories. The first category includes UAP, GAP and random attacks, which generate image-level perturbations. The second category consists of fine-grained attacks with four different settings of the CAUP attack: i) Uniform is a CAUP attack with uniform $\{w_a^c\}_{a\in A}$ instead of trainable class-attribute function, ii) CAUP-CE represents the CAUP attack with the Cross-Entropy loss function iii) CAUP-conf2 is a CAUP attack with the Ranking loss function and $\delta = 2$ iv) CAUP is the default version of attack with the Ranking loss function and $\delta = 8$.

On all datasets CAUP attack (type iv, which is our proposed attack in the paper) outperforms all attacks of the first category of methods (UAP, GAP, Random). These results show that the first category of attacks fail to fool the zero-shot fine-grained model properly, as they do not leverage fine attribute models. So, they are not a good suitable as (zero-shot) fine-grained perturbations.

On all datasets, CAUP-CE underperforms other CAUP versions. This confirms the fact that Ranking loss is a stronger tool to attack zero-shot fine-grained models. While we trained both CAUP and CAUP-conf2 on Ranking loss, CAUP shows a better performance than CAUP-conf2 on CUB and SUN and similar performance on AWA2. We can conclude that a higher confidence value makes the ℓ_2-norm CAUP attack to perform better. We additionally provide the performances on the harmonic mean H for both seen and unseen scenarios as a function of the magnitude (ℓ_2-norm and ℓ_∞-norm) of the perturbation:

$$H = 2 \times \frac{fool_s \times fool_u}{fool_s + fool_u} \tag{1}$$

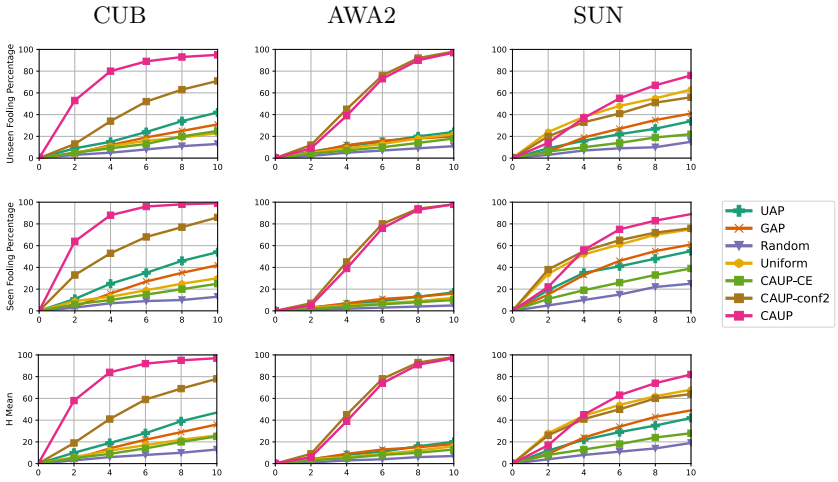


Fig. 1: Unseen (top) and seen (bottom row) fooling percentage of different universal attacks on DAZLE as a function of the magnitude (ℓ_2-norm) of the perturbation.

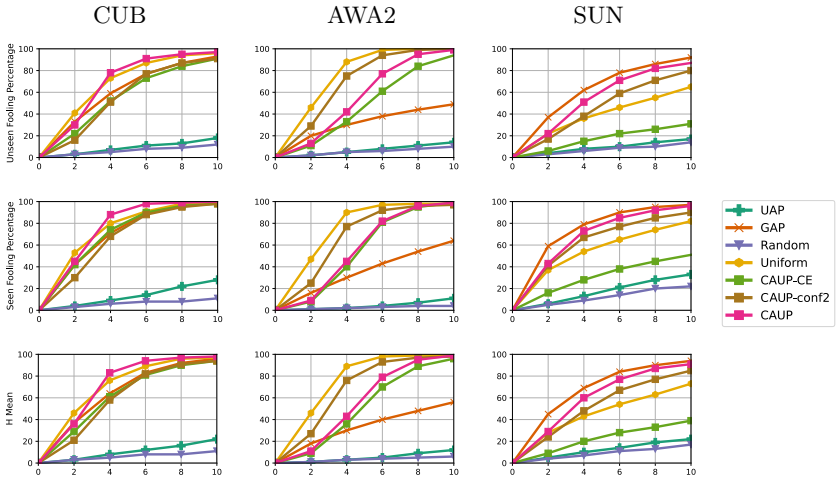


Fig. 2: Unseen (top) and seen (bottom row) fooling percentage of different universal attacks on DAZLE as a function of the magnitude (ℓ_∞-norm $\times 0.01$) of the perturbation.

As Figure 1 shows, harmonic mean mimics the same trend as seen/unseen figures. CAUP achieves a higher H mean for different values of the perturbation magnitude on CUB and SUN. On the AWA2 dataset, CAUP and CAUP-conf2 perform similarly and better than other attacks. This supports CAUP as a more intelligent attack, with a proper direction of perturbation vector, to deceive zero-shot fine-grained models.

Figure 2 presents the performance evaluation of ℓ_∞-norm attack using UAP [2], GAP [1] and different settings of CAUP. On CUB and AWA2, CAUP outper-

forms UAP and GAP both on seen and unseen scenarios. This indicates that CAUP as an ℓ_∞-norm attack is a stronger tool to mislead zero-shot fine-grained models. On SUN, GAP outperforms CAUP, but CAUP still outperforms UAP. The reason is SUN dataset includes more abstract attributes, which makes it harder for our attack to perform. Comparing different settings of CAUP attack, on CUB and SUN dataset CAUP attack, trained on ranking loss with $\delta = 8$, outperforms the other two CAUP settings. On AWA2, Uniform(CAUP attack with uniform weights) and CAUP-conf2 are fine-grained attacks that outperform CAUP. We can conclude that composing attribute-based perturbation to attack zero-shot fine-grained models performs better than just directly learning an image-level universal perturbation.

1.1 Accuracies of Fine-Grained Models

The success of an adversarial attack can originate from either the universal attack's strength or the model's weakness. To investigate this, we have to keep one of the factors constant and compare all the universal attacks on the same model. In this section, we discuss more details about the accuracy of each model and why our comparison of attacks represents their effectiveness. Table 1 shows the clean accuracies (accuracy on unperturbed images) of DAZLE, DCN, CNZSL, and CEZSL models on the test set of CUB, AWA2, and SUN datasets. The seen column corresponds to the accuracy of seen classes whose samples are present in the training. On the other hand, the unseen column corresponds to zero-shot classes where the model is not trained on any samples from these classes. Comparing all the models, CEZSL outperforms other models in five out of six cases and is our most accurate candidate to attack. In addition, Figure 3 shows the training and validation results for these models on the CUB dataset. The y-axis and x-axis represent the accuracies on each set, and the epoch number during the training, respectively. As the figure demonstrates, there is no case in which the training accuracy (yellow) increases while testing accuracies (blue and red) decrease, which implies the models do not overfit. It is worth mentioning that the overfitting of the model could potentially benefit image-specific attacks like PGD and MIM. However, this will not be likely for a universal attack (a single perturbation direction); therefore, increasing the nonlinearity of the classifier will not help these attacks. In addition, comparing different universal attacks is fair since overfitting and instability of the target models would affect all the attack methods to have a better/worse performance. Consequently, UAP and GAP do not perform well on the fine-grained models for the three datasets compared to our attack as shown in figure 1, 2.

2 CAUP Hyperparameters

To select hyperparameters $\{\lambda_{reg}, \lambda_{util}\}$, we perform a search over 5 values in $[2^{-7}, 2^1]$ for λ_{reg} and 5 values in $[2^{-7}, 2^1]$ for λ_{util} and select the pair that achieves the best accuracy drop using CAUP attack over validation sets. Also, the margin

Table 1: Accuracy of Models on Clean Images

Accuracy %	CUB		AWA2		SUN	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
DAZLE	59.6	56.7	75.7	60.3	24.3	52.3
DCN	60.7	28.4	37.0	25.5	37.0	25.5
CNZSL	50.7	49.9	77.q	60.2	41.6	44.7
CEZSL	66.8	63.9	78.6	63.1	36.8	48.8

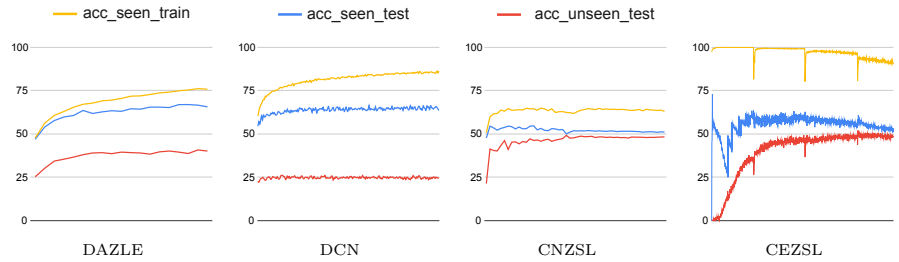


Fig. 3: Accuracy vs Epoch for training (yellow) and testing-seen (blue) and testing-unseen (red) of the CUB dataset.

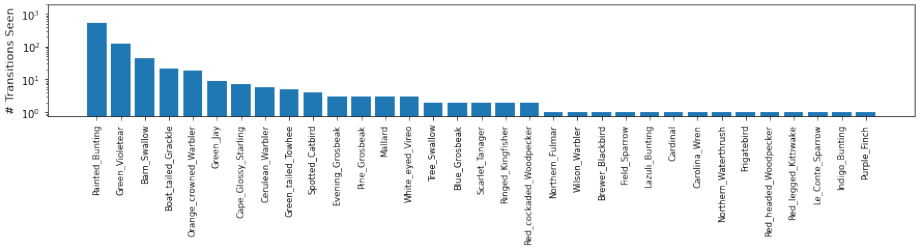
CAUP hyperparameters		ℓ_2		ℓ_∞	
Model	Dataset	λ_{reg}	λ_{util}	λ_{reg}	λ_{util}
DAZLE	CUB	2^{-1}	2^{+1}	2^{-5}	2^{+1}
	AWA	2^{-3}	2^{-1}	2^{-3}	2^{-1}
	SUN	2^{-5}	2^{-7}	2^{-1}	2^{-7}
CEZSL	CUB	2^{-3}	2^{+1}	2^{-5}	2^{+1}
	AWA	2^{-3}	2^{+1}	2^{-5}	2^{+1}
	SUN	2^{-1}	2^{-7}	2^{-5}	2^{-5}
CNZSL	CUB	2^{-7}	2^{-3}	2^{-3}	2^{-3}
	AWA	2^{-3}	2^{-5}	2^{-5}	2^{-1}
	SUN	2^{-3}	2^{-7}	2^{-5}	2^{-7}
DCN	CUB	2^{-1}	2^{-1}	2^{+1}	2^{-1}
	AWA	2^{-5}	2^{-7}	2^{-1}	2^{+1}
	SUN	2^{+1}	2^{+1}	2^{+1}	2^{+1}

Table 2: Hyperparameters values λ_{reg} and λ_{util} used in our experiments. These values are obtained by grid search over the validation set.

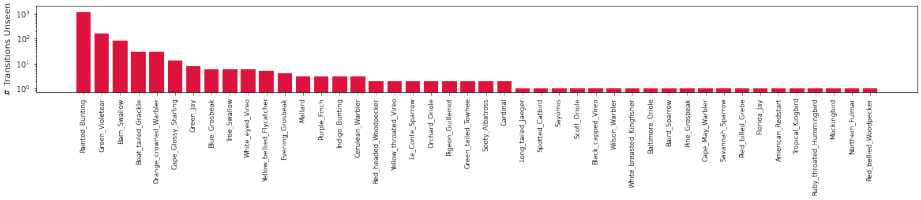
parameter δ is set to 8. The optimal values of these hyperparameters are reported in Table 2. Each row contains optimized values for each model and each dataset. Note that we used one-third of the validation set for hyperparameter tuning to be consistent with all other experiments.

3 Dominant Adversarial Classes

In Section 4.2.2 of the main paper, we investigated the existence of dominant adversarial classes. To better demonstrate the generalization of dominant adversarial classes, on CUB dataset, we attack all images from seen and unseen classes of the test set and count the number of images deceived into each class. In Figure 4, each bar corresponds to the number of images that has been fooled



(a) Seen



(b) Unseen

Fig. 4: Top dominant adversarial classes on seen (top) and unseen (bottom) test set. Each plot is the distribution of all test images that have been fooled into adversarial classes. For better visualization, we removed the tail of classes with zero fooled images.

into a particular adversarial class where the class name is shown on the horizontal axis. Notice that CUB includes 200 classes in total. Nevertheless, only 33 and 44 of classes will be adversarial classes when we attack all test images from seen and unseen splits, respectively. This shows the dominancy of a small subset of adversarial classes in the fine-grained setting. As we mentioned before, the dominancy of adversarial classes happens because of the major contribution of attributes that are more specific to some species of birds. Employing these characteristics can lead to defense mechanisms for fine-grained recognition and we leave it for future studies.

References

1. Hayes, J., Danezis, G.: Learning universal adversarial perturbations with generative models. IEEE Security and Privacy Workshops (2018) 1, 2
2. Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. IEEE Conference on Computer Vision and Pattern Recognition (2017) 1, 2
3. Zhang, C., Benz, P., Imtiaz, T., Kweon, I.S.: Cd-uap: Class discriminative universal adversarial perturbation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6754–6761 (2020) 1