Zero-Shot Attribute Attacks on Fine-Grained Recognition Models

Nasim Shafiee
Shafiee.n@northeastern.edu

Ehsan Elhamifar
e.elhamifar@northeastern.edu

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Motivation

- Generating Perturbations to adversarially attack fine-grained data-driven models
- Extending the attack for images from Unseen(Zero-Shot) classes

Adversarial Attack

- Customizing perturbation per image
- Attacking needs time-consuming perturbations generation
- Not expandable to new images

Universal Attack

- Employing a common perturbation(s) learned during training
- Attacking in real-time
- Easily expanding to new images

Contributions

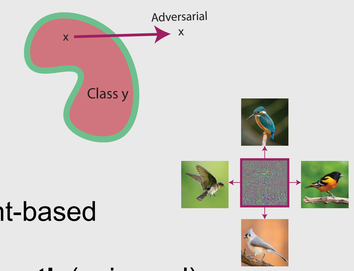
Developing a **compositional attribute-based attack** that:

- Captures **minor differences** in fine-grained classes
- Generalizes well over well to **unseen classes** without training images
- Generalizes across well-known fine-grained **models and datasets**
- Can be applied in **real-time**

Introducing a **novel utility loss** to ensure training of all attribute-based universal perturbations

Prior Works

Properties of an attack can be :



- **Non-targeted** vs Targeted
- Blackbox vs **Whitebox**
- **Optimization-based** vs Gradient-based
- Image-dependent vs **Image-agnostic**(universal)

Limitations

- Conventional adversarial attacks for coarse-level classification **do not work well for fine-grained** recognition.
- There is no principled method for crafting effective adversarial attacks for **unseen classes** that do not have training images

Problem Setting

To generate an attack on an **image I** that belongs to a **class y**, we need to generate **a perturbation e** such that

$$\exists\, c \in C \backslash y \;\; s.t. \;\; S^c(I+e) > S^y(I+e)$$

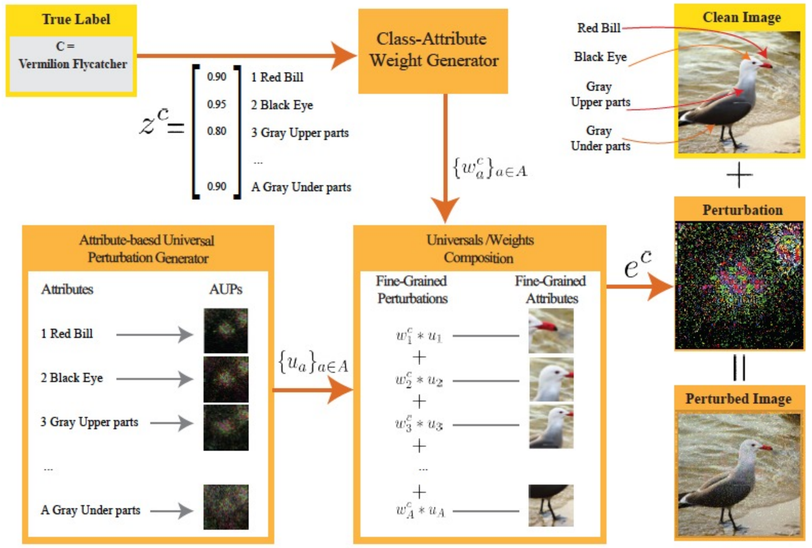
Proposed Framework

- **Attribute-based Universal Perturbations(AUPs):** $\{u_a\}_{a\in A}$
- **Class-level Perturbations:**
  composing attributed-based universal as following

$$e^c = \sum_a \omega_a^c \, u_a$$

- **Compositional Weights:**
  composing attributed-based universal as following

$$\omega_a^c = \tanh(v_a^T W_a z^c)$$



Training

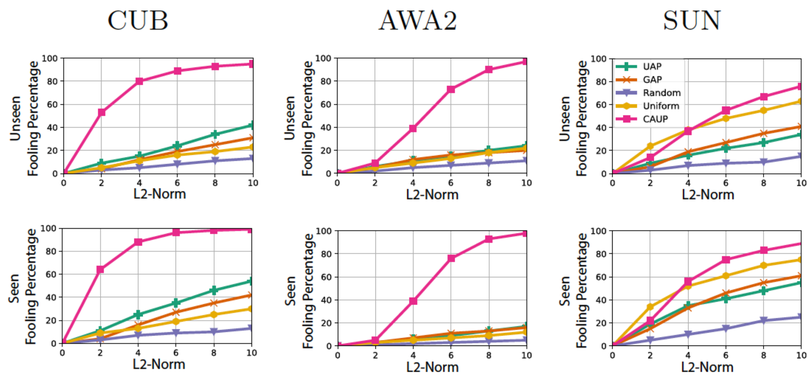
End-to-End training using the following loss function

$$L_{rank} + \lambda_{reg} L_{reg} + \lambda_{util} L_{util}$$

- **Ranking** Loss $L_{rank} = \sum_I \max\{0, \delta + S^y(I+e^y) - \max S^c(I+e^y)\}$
- **Regularization** Loss $L_{reg} = \begin{cases} \sum_a \|u_a\|_2^2 & L2\ attack \\ \sum_a \sum_j (u_{a,j} - k)^+ & L\infty\ attack \end{cases}$
- **Utility** Loss $L_{util} = \sum_a \tau_a^2$

Experiments

The performance of our attack on both **seen** and **unseen** classes with different perturbation magnitudes



Comparing our attack with other methods on three datasets and four different fine-grained models:

Fooling Percentage	CUB			AWA2			SUN		
Seen/Unseen	UAP	GAP	CAUP	UAP	GAP	CAUP	UAP	GAP	CAUP
DAZLE	35/ 24	27/ 19	**96/ 89**	09/ 15	11/ 16	**89/ 76**	41/ 22	46/ 27	**75/ 55**
DCN	51/ 46	21/ 21	**70/ 70**	02/ 05	07/ 10	**08/ 15**	41/ 21	30/ 15	**59/ 33**
CNZSL	26/ 25	18/ 21	**96/ 91**	13/ 12	09/ 12	**54/ 42**	**29/ 16**	27/ 20	22/ 12
CEZSL	39/ 40	35/ 35	**99/ 95**	24/ 21	12/ 15	**81/ 75**	29/ 28	26/ 26	**92/ 89**

L2 Norm Attack

Fooling Percentage	CUB			AWA2			SUN		
Seen/Unseen	UAP	GAP	CAUP	UAP	GAP	CAUP	UAP	GAP	CAUP
DAZLE	14/ 11	89/ 77	**98/ 91**	04/ 08	43/ 38	**82/ 77**	21/ 10	**90/ 78**	85/ 71
DCN	05/ 06	**74/ 73**	70/ 66	01/ 02	15/ 19	**47/ 48**	09/ 05	**81/ 63**	66/ 41
CNZSL	16/ 18	61/ 55	**97/ 93**	06/ 08	57/ 73	**86/ 79**	20/ 11	**75/ 55**	48/ 19
CEZSL	18/ 24	79/ 77	**99/ 95**	02/ 04	55/ 43	**97/ 96**	26/ 15	81/ 72	**92/ 86**

L∞ Norm Attack

Visualization of attribute-based universal perturbations (left) and the adversarial examples generated by our attack (right):