Better Aggregation in Test-Time Augmentation

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, John Guttag

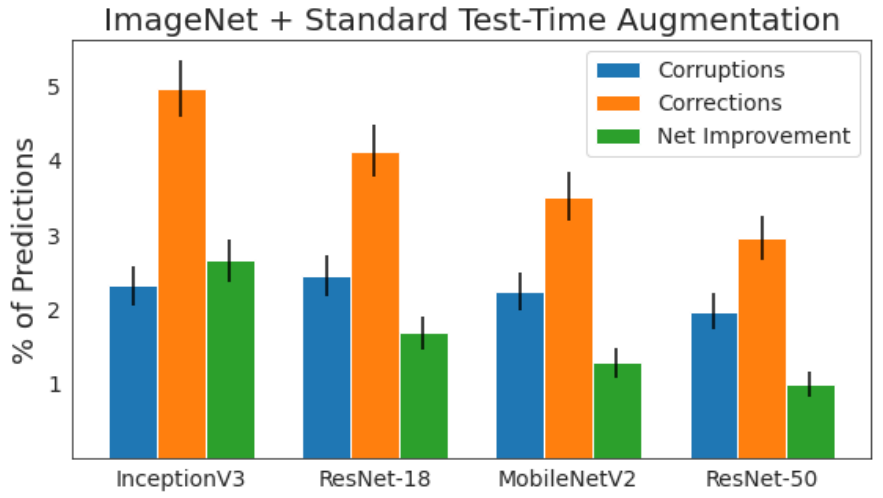
2021 ICCV OCTOBER 11-17 VIRTUAL

CSAIL

Motivation

TTA introduces many incorrect predictions.

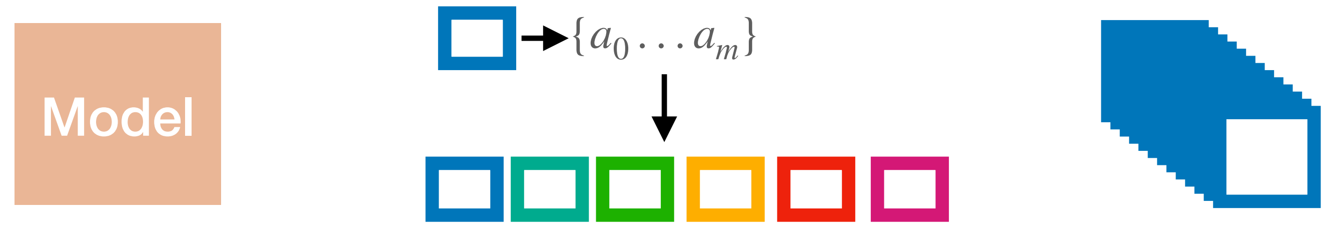
We aim to **characterize the errors introduced by TTA** and **develop a method to address these shortcomings.**



Method

Key idea: learn augmentation specific weights to aggregate predictions.

Using:



Probabilistic classifier

Augmentation Policy

Validation Set

Learn:

$$[\theta_1 \cdots \theta_M]\begin{bmatrix} a_{11} & \cdots & a_{1C} \\ \vdots & \ddots & \\ a_{M1} & & a_{MC} \end{bmatrix}$$

AugTTA

$$\begin{bmatrix} \theta_{11} & \cdots & \theta_{1C} \\ \vdots & \ddots & \\ \theta_{M1} & & \theta_{MC} \end{bmatrix}\begin{bmatrix} a_{11} & \cdots & a_{1C} \\ \vdots & \ddots & \\ a_{M1} & & a_{MC} \end{bmatrix}$$

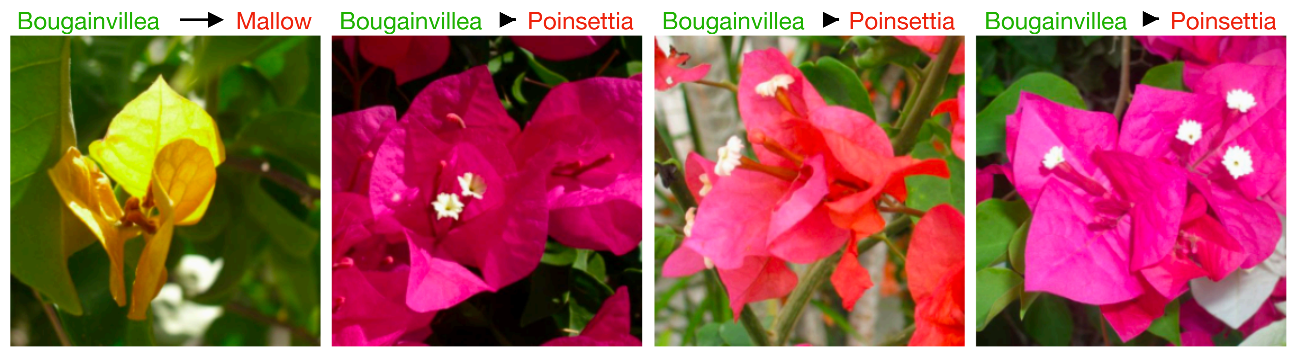
ClassTTA

Takeaway

We present a new TTA method that uses an **augmentation-specific approach to aggregation** and provides improvements in classification accuracy.

Analysis

1. Standard TTA changes predictions for classes with smaller distinguishing features, and classes that vary in scale.

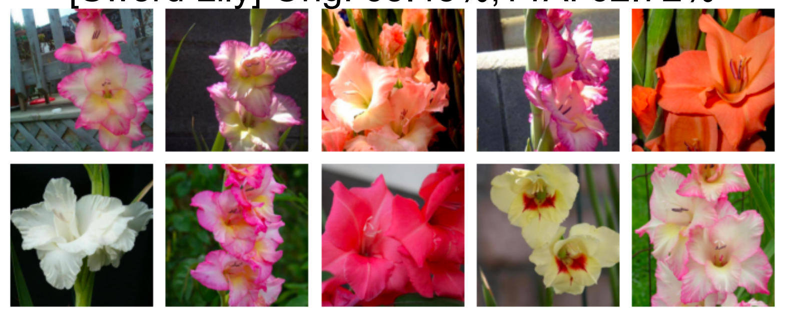

Bougainvillea → Mallow
Bougainvillea ► Poinsettia
Bougainvillea ► Poinsettia
Bougainvillea ► Poinsettia

2. TTA harms classification accuracy for classes that exhibit higher variation in the training data.

3. The value of TTA is significantly correlated with the number of examples per class.

[Primula] Orig: 65.75%, TTA: 69.86%

[Sword Lily] Orig: 65.45%,TTA: 62.72%



Results

1. Our method outperforms others across four datasets, four architectures, and two test-time augmentation policies.

Standard TTA Policy.

Dataset	Model	Original	Max	Mean	GPS	Ours
Flowers102	MobileNetV2	90.28 ± 0.10	90.17 ± 0.25	90.47 ± 0.20	88.28 ± 0.17	**92.62 ± 0.10**
Flowers102	InceptionV3	89.28 ± 0.08	89.59 ± 0.15	90.07 ± 0.22	89.93 ± 0.16	**91.16 ± 0.21**
Flowers102	ResNet-18	89.78 ± 0.17	89.47 ± 0.11	90.21 ± 0.23	90.01 ± 0.22	**91.02 ± 0.17**
Flowers102	ResNet-50	**91.72 ± 0.18**	91.61 ± 0.08	**91.96 ± 0.27**	**92.03 ± 0.09**	**92.02 ± 0.16**
ImageNet	MobileNetV2	71.38 ± 0.06	72.50 ± 0.13	**72.69 ± 0.06**	72.50 ± 0.11	72.43 ± 0.08
ImageNet	InceptionV3	69.66 ± 0.12	71.8 ± 0.09	72.45 ± 0.13	71.57 ± 0.10	**72.79 ± 0.02**

Expanded TTA Policy.

Dataset	Model	Original	Max	Mean	GPS	Ours
Flowers102	MobileNetV2	90.94 ± 0.16	86.85 ± 0.24	91.14 ± 0.08	91.34 ± 0.16	**92.49 ± 0.20**
Flowers102	InceptionV3	89.17 ± 0.33	87.89 ± 0.20	89.20 ± 0.23	89.43 ± 0.16	**91.02 ± 0.26**
Flowers102	ResNet-18	89.20 ± 0.10	83.30 ± 0.19	89.47 ± 0.09	**89.90 ± 0.24**	**89.78 ± 0.16**
Flowers102	ResNet-50	92.37 ± 0.13	89.39 ± 0.19	92.48 ± 0.11	92.57 ± 0.21	**93.29 ± 0.21**
ImageNet	MobileNetV2	71.18 ± 0.05	67.65 ± 0.08	71.84 ± 0.12	**72.49 ± 0.09**	**72.57 ± 0.09**
ImageNet	InceptionV3	69.51 ± 0.08	66.00 ± 0.13	70.85 ± 0.11	**71.05 ± 0.08**	**71.02 ± 0.06**

2. Learned weights confirm qualitative results, and demonstrate higher variance for classes that exhibit higher variation in the training data.

Black-Eyed Susan
Low Variance in Aug. Weights

Columbine
High Variance in Aug. Weights

Read the paper for more test-time augmentation insights and instructions to reproduce experiments!