Better Aggregation in Test-Time Augmentation: Supplementary Material

August 24, 2022

1 Augmentation Policies

1.1 Standard TTA Policy

The standard TTA policy produces 30 transformations of the original image. The 30 perturbations are produced by all possible combinations of flips, five crops and scales. We list the specific parameters in Table 1.

Augmentation	# Parameters	Description
hflip	2	Produces the original image or its horizontal flip
five_crops	5	Produces crops from the center or one of four corners of the image. Crops can be 224x224 (Flowers-102, ImageNet), 96x96 (STL-10), or 32x32 (CIFAR-100).
scale	3	Rescales image by 1.0, 1.04, or 1.10.

Table 1: Description of the three types of augmentations used in the standard test-time augmentation policy. The cross-product of these parameters produces 30 unique perturbations (including the identity).

1.2 Expanded TTA Policy

The expanded TTA policy produces 128 transformations of the original image. The 128 transformations correspond to a individual parameter sweep for 20 augmentations. We list the specific augmentations and the parameters in Table 1. We borrow this set directly from [1].

Augmentation	# Parameters	Description
FlipLR	1	Produces the horizontal flip of an image
FlipUD	1	Produces the vertical flip of an image
Invert	1	Invert (negate) the image.
Rotate	10	Rotates the image by p degrees.
Posterize	10	Reduces the number of bits for each color channel.
CropBilinear	10	Crops image and resizes using bilinear interpolation.
Solarize	10	Inverts all pixel values above a threshold.
Color	10	Adjust the color balance of the image.
Contrast	1	Control the contrast of the image.
Brightness	10	R Adjust the brightness of the image.
Sharpness	10	Adjust the sharpness of the image.
ShearX	10	Shear the image along the horizontal axis by parameter p.
ShearY	10	Shear the image along the vertical axis by parameter p.
TranslateX	10	Translate the image in the horizontal direction by p pixels.
TranslateY	10	Translate the image in the vertical direction by p pixels.
Cutout	1	Set a random square patch of side-length p pixels to gray.
Blur	10	Applies Gaussian Blur filter to image.
Smooth	1	Applies Python Smooth filter to image.
Equalize	1	Applies a non-linear mapping to the image to produce a uniform distribution of grayscale values.
AutoContrast	1	Calculates a histogram of the input image, removes cutoff percent of the lightest and darkest pixels from the histogram, and remaps image so the darkest pixel becomes black (0), and the lightest becomes white (255).

Table 2: Description of the 20 augmentations used in the expanded test-time augmentation policy. In total, 128 perturbations are produced, representing one perturbation for each augmentation included above. We do not consider the cross-product of the augmentations because it requires a prohibitive amount of memory.

2 Results:*AugTTA* vs *ClassTTA*

Our method selects between two parameterizations: *AugTTA* and *ClassTTA*. *AugTTA* learns a weight per augmentation, while *ClassTTA* learns a weight per class-augmentation pair. Here, we include results for each parameterization given a standard TTA policy (Table 2) and a expanded TTA policy (Table **??**). Earlier, we note that our method chooses *AugTTA* on all datasets except Flowers-102. Our experiments suggest that this is due to the abundance of validation data, paired with the class-specific relationships to specific test-time augmentations.

Dataset	Model	Original	Max	Mean	GPS	AugTTA	ClassTTA
Flowers102	MobileNetV2	90.28 ± 0.10	90.17 ± 0.25	90.47 ± 0.20	88.28 ± 0.17	90.71 ± 0.14	**92.48 ± 0.15**
Flowers102	InceptionV3	89.28 ± 0.08	89.59 ± 0.15	90.07 ± 0.22	89.93 ± 0.16	90.30 ± 0.15	**91.31 ± 0.21**
Flowers102	ResNet-18	89.78 ± 0.17	89.47 ± 0.11	90.21 ± 0.23	90.01 ± 0.22	90.42 ± 0.16	**90.78 ± 0.12**
Flowers102	ResNet-50	91.72 ± 0.18	91.61 ± 0.08	91.96 ± 0.27	92.03 ± 0.09	91.96 ± 0.25	**92.85 ± 0.20**
ImageNet	MobileNetV2	71.38 ± 0.06	72.50 ± 0.13	**72.69 ± 0.06**	72.50 ± 0.11	**72.74 ± 0.08**	72.42 ± 0.07
ImageNet	InceptionV3	69.66 ± 0.12	71.80 ± 0.09	72.45 ± 0.13	71.57 ± 0.10	**72.74 ± 0.08**	**72.87 ± 0.07**
ImageNet	ResNet-18	69.37 ± 0.10	70.26 ± 0.13	**71.02 ± 0.13**	70.80 ± 0.10	**71.12 ± 0.09**	70.74 ± 0.12
ImageNet	ResNet-50	75.78 ± 0.08	76.62 ± 0.08	**76.91 ± 0.09**	**76.73 ± 0.11**	**76.81 ± 0.13**	76.67 ± 0.09
CIFAR100	CNN-7	74.38 ± 0.18	75.04 ± 0.17	**75.55 ± 0.25**	75.39 ± 0.11	**75.95 ± 0.24**	74.29 ± 0.23
STL10	CNN-5	77.92 ± 0.19	77.76 ± 0.22	**78.58 ± 0.25**	**78.32 ± 0.17**	**78.60 ± 0.36**	78.51 ± 0.27

Table 3: Results for *AugTTA* and *ClassTTA* given the *standard* TTA policy.

Dataset	Model	Original	Max	Mean	GPS	AugTTA	ClassTTA
Flowers102	MobileNetV2	90.94 ± 0.16	86.85 ± 0.24	91.14 ± 0.08	91.34 ± 0.16	91.11 ± 0.24	**92.76 ± 0.14**
Flowers102	InceptionV3	89.17 ± 0.33	87.89 ± 0.20	89.20 ± 0.23	89.43 ± 0.16	89.75 ± 0.15	**91.17 ± 0.22**
Flowers102	ResNet-18	89.20 ± 0.10	83.30 ± 0.19	89.47 ± 0.09	89.90 ± 0.24	89.83 ± 0.05	**91.28 ± 0.19**
Flowers102	ResNet-50	92.37 ± 0.13	89.39 ± 0.19	92.48 ± 0.11	92.57 ± 0.21	92.67 ± 0.28	**93.32 ± 0.11**
ImageNet	MobileNetV2	71.18 ± 0.05	67.65 ± 0.08	71.84 ± 0.12	**72.49 ± 0.09**	**72.58 ± 0.06**	64.58 ± 0.09
ImageNet	InceptionV3	69.51 ± 0.08	66.00 ± 0.13	70.85 ± 0.11	**71.05 ± 0.08**	**71.15 ± 0.11**	67.95 ± 0.13
ImageNet	ResNet-18	69.62 ± 0.15	66.56 ± 0.12	70.11 ± 0.13	**70.91 ± 0.05**	**70.83 ± 0.08**	63.33 ± 0.10
ImageNet	ResNet-50	75.53 ± 0.06	71.99 ± 0.15	75.87 ± 0.17	76.12 ± 0.08	**76.33 ± 0.09**	70.35 ± 0.12
CIFAR100	CNN-7	74.17 ± 0.18	64.05 ± 0.16	73.33 ± 0.13	**75.06 ± 0.25**	73.05 ± 0.33	73.24 ± 0.05
STL10	CNN-5	78.04 ± 0.18	74.77 ± 0.12	**79.02 ± 0.21**	78.81 ± 0.27	**79.09 ± 0.19**	**79.29 ± 0.23**

Table 4: Results for *AugTTA* and *ClassTTA* given the *expanded* TTA policy.

3 Results: Top-5 Accuracies

We include the results of our method and the baselines in terms of Top-5 classification accuracy given a standard TTA policy (Table 3) and an expanded TTA policy (Table 3). These results align with trends we note with respect to Top-1 accuracy. In particular, our method exceeds the performance of existing approaches in 10 out of the 20 included comparisons.

Dataset	Model	Original	Max	Mean	GPS	Ours
Flowers102	MobileNetV2	97.91 ± 0.08	97.04 ± 0.13	97.78 ± 0.10	97.10 ± 0.07	**98.67 ± 0.11**
Flowers102	InceptionV3	97.29 ± 0.09	96.95 ± 0.12	97.67 ± 0.13	97.34 ± 0.05	**98.20 ± 0.05**
Flowers102	ResNet-18	97.56 ± 0.06	97.09 ± 0.09	97.57 ± 0.09	97.67 ± 0.12	**98.00 ± 0.08**
Flowers102	ResNet-50	97.81 ± 0.09	97.69 ± 0.10	**98.05 ± 0.11**	**98.05 ± 0.10**	**98.07 ± 0.09**
ImageNet	MobileNetV2	90.11 ± 0.07	90.13 ± 0.08	90.99 ± 0.05	90.84 ± 0.08	**91.15 ± 0.03**
ImageNet	InceptionV3	89.06 ± 0.06	88.80 ± 0.05	90.71 ± 0.06	90.17 ± 0.06	**90.99 ± 0.08**
ImageNet	ResNet-18	89.03 ± 0.06	88.90 ± 0.12	**89.98 ± 0.05**	**89.90 ± 0.05**	**90.01 ± 0.09**
ImageNet	ResNet-50	92.87 ± 0.05	92.74 ± 0.07	**93.41 ± 0.05**	**93.28 ± 0.05**	**93.34 ± 0.11**
CIFAR100	CNN-7	92.85 ± 0.09	92.63 ± 0.16	**93.61 ± 0.14**	**93.56 ± 0.04**	**93.68 ± 0.19**
STL10	CNN-5	**98.00 ± 0.07**	**98.09 ± 0.10**	98.14 ± 0.08	**98.24 ± 0.07**	**98.19 ± 0.04**

Table 5: Top-5 Accuracy of our method and baselines given the standard TTA policy.

4 Results: Weights learned for Expanded Policy

In the body of the paper, we discuss the weights learned given the standard TTA policy. Here, we devote space to the augmentations with non-zero weights in the expanded policy. In particular, we focus on the application of our method to STL-10 and the CNN-5 architecture, along with ImageNet and ResNet-50. We focus on these since they are cases in which our method outperforms the baselines by a large margin.

For STL-10, there are three types of augmentations with non-zero weights. There are augmentations that produce small transformations, that nearly produce the original image (weight of .30). There are the traditional standard augmentations - horizontal flips (weight of .18), and to a lesser extent, vertical flips (weight of .04). Lastly, and most interestingly, there is a third group of augmentations that modify the intensities within the image— AutoContrast and Equalizat—that are also also contribute to the final prediction (weights of .11 and .10 respectively).

Dataset	Model	Original	Max	Mean	GPS	Ours
Flowers102	MobileNetV2	97.65 ± 0.12	93.34 ± 0.06	97.58 ± 0.10	97.75 ± 0.10	**98.62 ± 0.05**
Flowers102	InceptionV3	97.31 ± 0.16	95.04 ± 0.17	97.65 ± 0.10	97.73 ± 0.10	**98.77 ± 0.06**
Flowers102	ResNet-18	97.56 ± 0.10	91.28 ± 0.06	97.48 ± 0.10	**97.74 ± 0.11**	**97.66 ± 0.09**
Flowers102	ResNet-50	97.89 ± 0.10	95.69 ± 0.08	98.16 ± 0.10	98.12 ± 0.05	**99.18 ± 0.06**
ImageNet	MobileNetV2	90.24 ± 0.05	84.64 ± 0.08	90.32 ± 0.05	**90.73 ± 0.08**	**90.69 ± 0.05**
ImageNet	InceptionV3	88.52 ± 0.07	83.28 ± 0.08	**89.43 ± 0.02**	**89.60 ± 0.03**	89.60 ± 0.05
ImageNet	ResNet-18	89.02 ± 0.07	84.22 ± 0.07	89.20 ± 0.05	89.55 ± 0.06	**89.69 ± 0.07**
ImageNet	ResNet-50	92.65 ± 0.05	87.65 ± 0.06	92.65 ± 0.06	**92.89 ± 0.04**	**92.92 ± 0.07**
CIFAR100	CNN-7	92.84 ± 0.11	76.37 ± 0.19	93.05 ± 0.12	**93.67 ± 0.15**	93.09 ± 0.16
STL10	CNN-5	97.97 ± 0.15	94.85 ± 0.09	**98.28 ± 0.07**	**98.24 ± 0.08**	**98.21 ± 0.14**

Table 6: Top-5 Accuracy of our method and baselines given the expanded TTA policy.

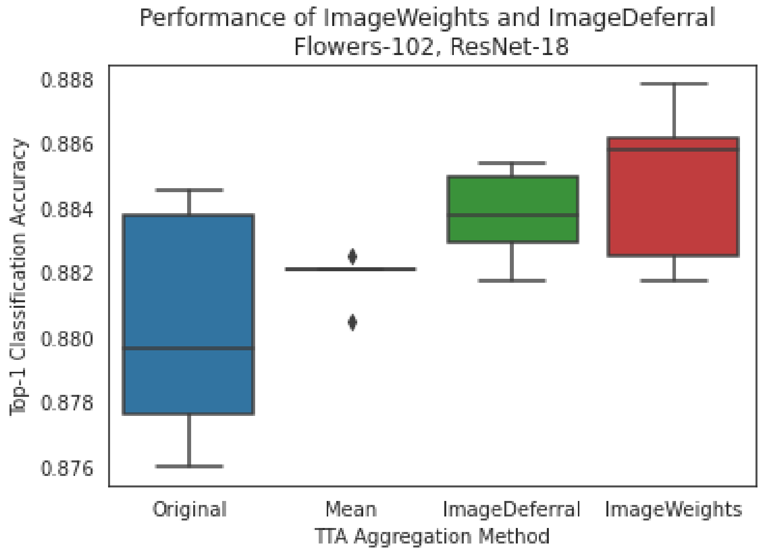

Figure 1: Performance of *ImageDeferral* and *ImageWeights* compared to taking the simple average and the original model predictions. Neither perform significantly better than either baseline.

For ImageNet, a similar trend holds, albeit with slightly different augmentations. Horizontal flips are weighted to be 30% of the final prediction. Interestingly, small translations appear to be useful at test-time. AutoContrast, Smooth, Blur, and Invert are also weighted higher with our method compared to the simple average. The non-traditional TTAs with non-zero weights differ in this case, and suggest that such a method could surface domain-specific TTAs that are not typically considered.

5 Results: Comparing to Other Baselines

We include this section primarily for researchers interested in similar questions. We tried out two additional models: 1) learning the optimal mixture of the simple average and the original model using the validation set and 2) learning augmentation weights directly from an image, rather than statically for an entire dataset. The first approach equates to learning to defer, where deferral equates to choosing the original model's prediction. The second approach models our hypothesis that useful TTAs depend on the input more directly. We determined that both approaches require a prohibitive amount of labelled data and thus, are not useful in practice. Our experiments compare the performance of both to the simple average on Flowers-102, using ResNet-18. Note that the accuracies are not identical to previous experiments, because the splits are re-randomized. We train each of the baselines for 30 epochs and code to reproduce these experiments is included.

Results are plotted in Figure 1. Neither method performs significantly better than the remaining baselines. Our experience suggests this is because the amount of data in the validation set is not sufficient to learn the function that maps images to the optimal augmentation weights, or to a deferral decision. While one could argue that these methods present some utility in the presence of abundant labeled data, it is likely more advantageous to train your model further on that labelled dataset, rather than learn a test-time augmentation approach.

References

[1] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.