Better Aggregation in Test-Time Augmentation

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, John Guttag

International Conference on Computer Vision (ICCV), 2021

TTA is the aggregation of predictions across transformations of an image.

Traditionally:

Model

“bakery”

TTA is the aggregation of predictions across transformations of an image.

Traditionally:

Model

“bakery” ✗

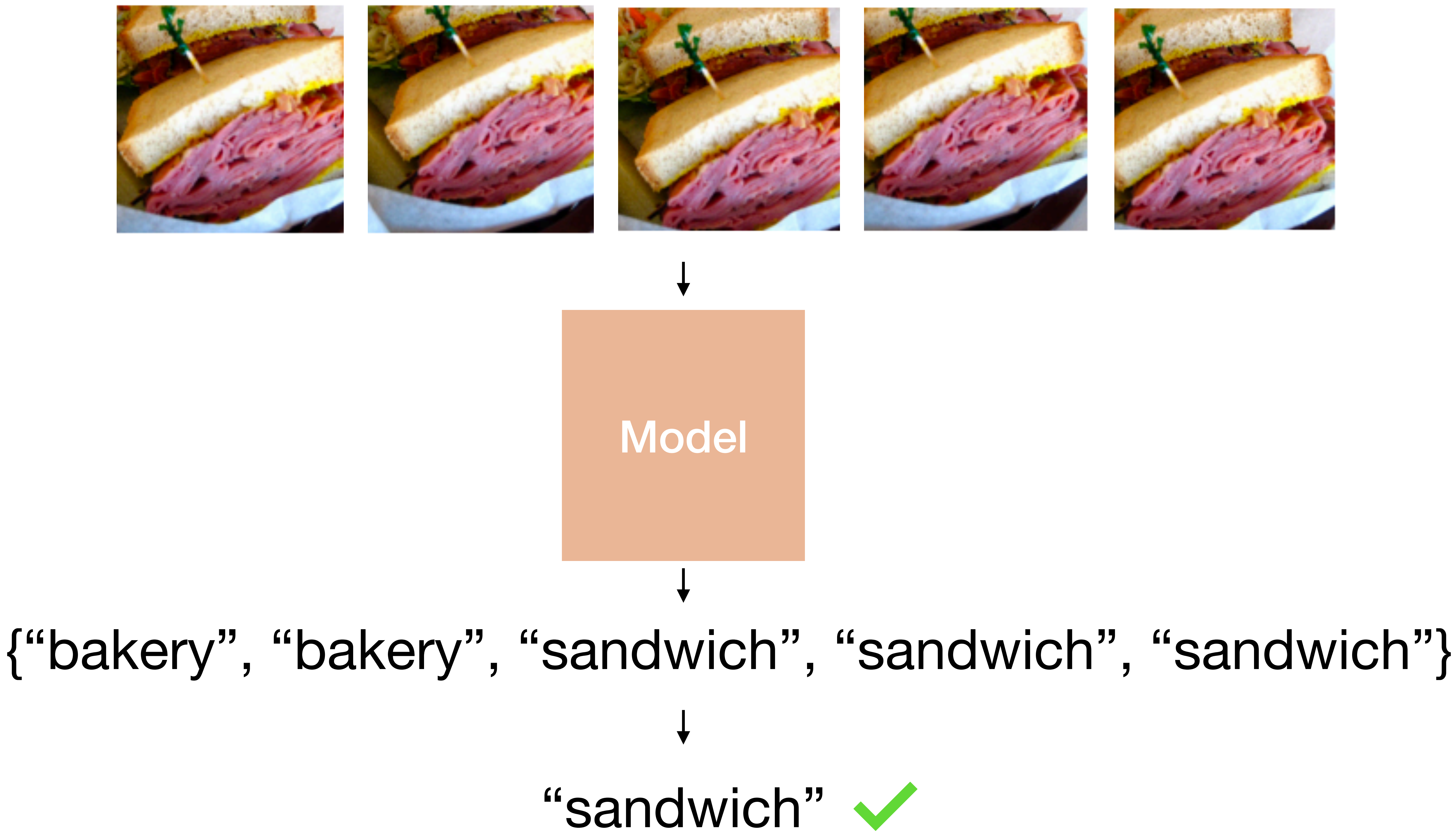
With TTA:

Model

{“bakery”, “bakery”, “sandwich”, “sandwich”, “sandwich”}

“sandwich” ✓

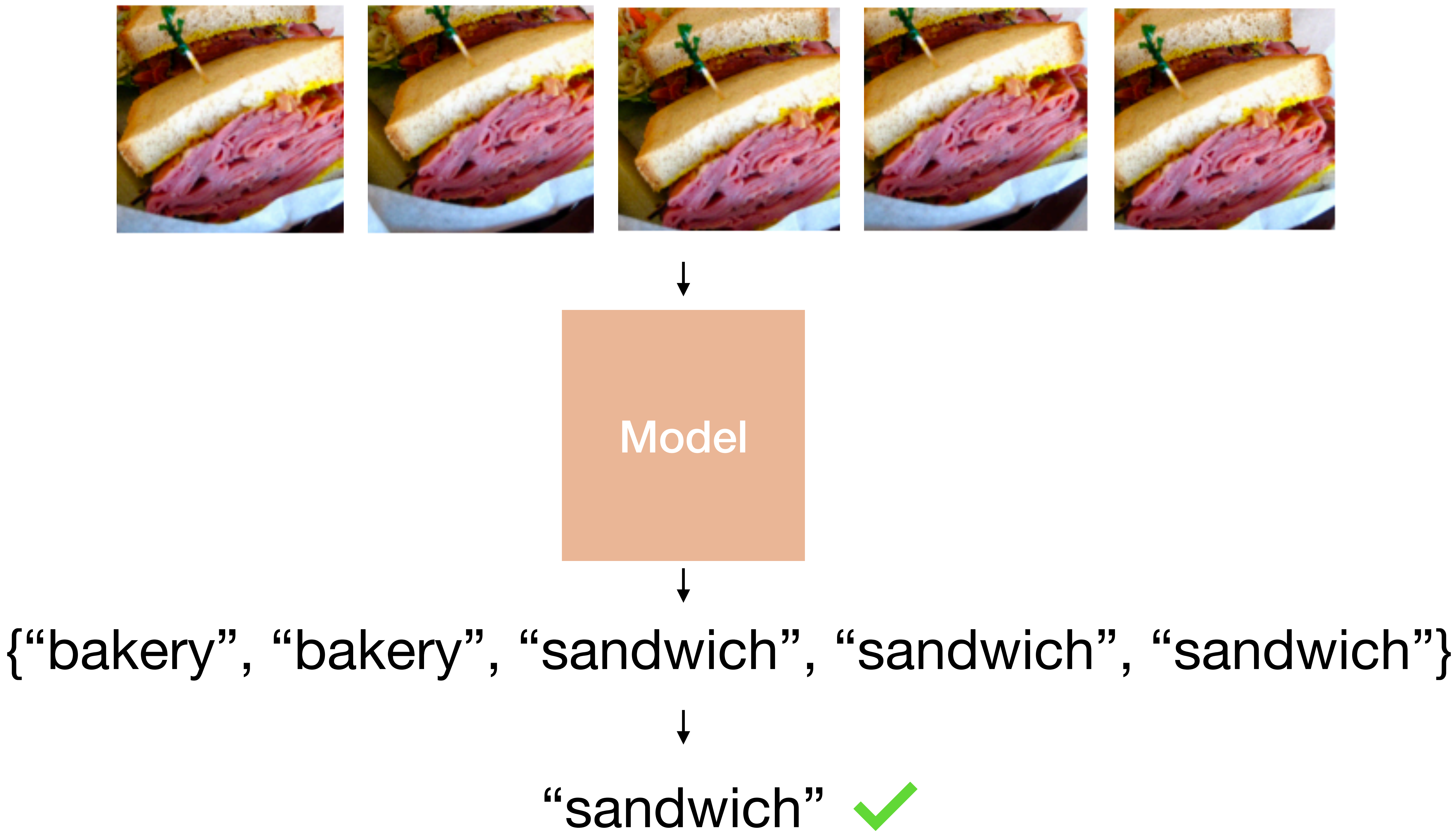
TTA produces more accurate and robust predictions than the original model *without retraining*



TTA produces more accurate and robust predictions than the original model *without retraining*

Two choices:

1. Selecting augmentations
2. Aggregating the resulting predictions

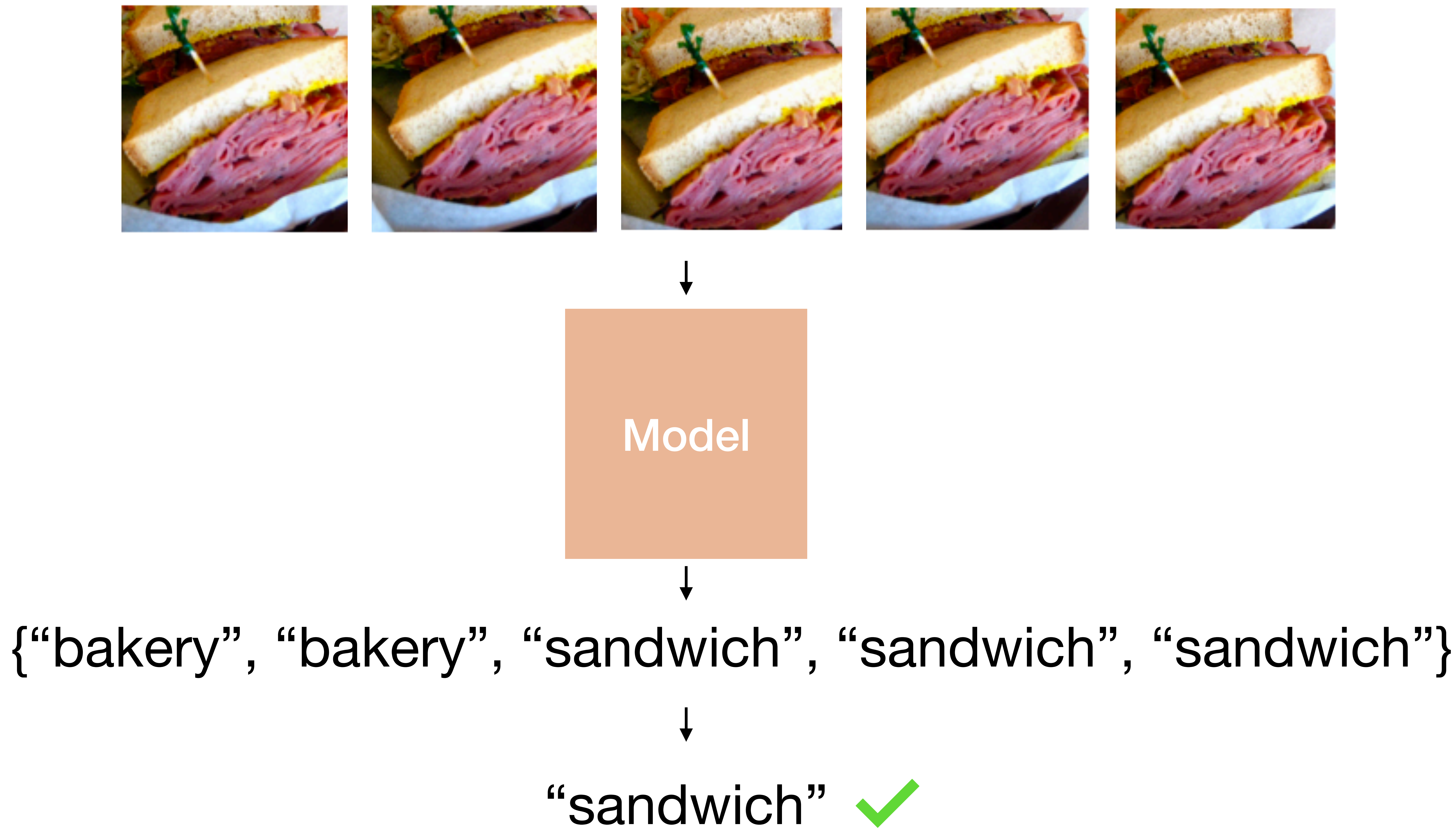


TTA produces more accurate and robust predictions than the original model *without retraining*

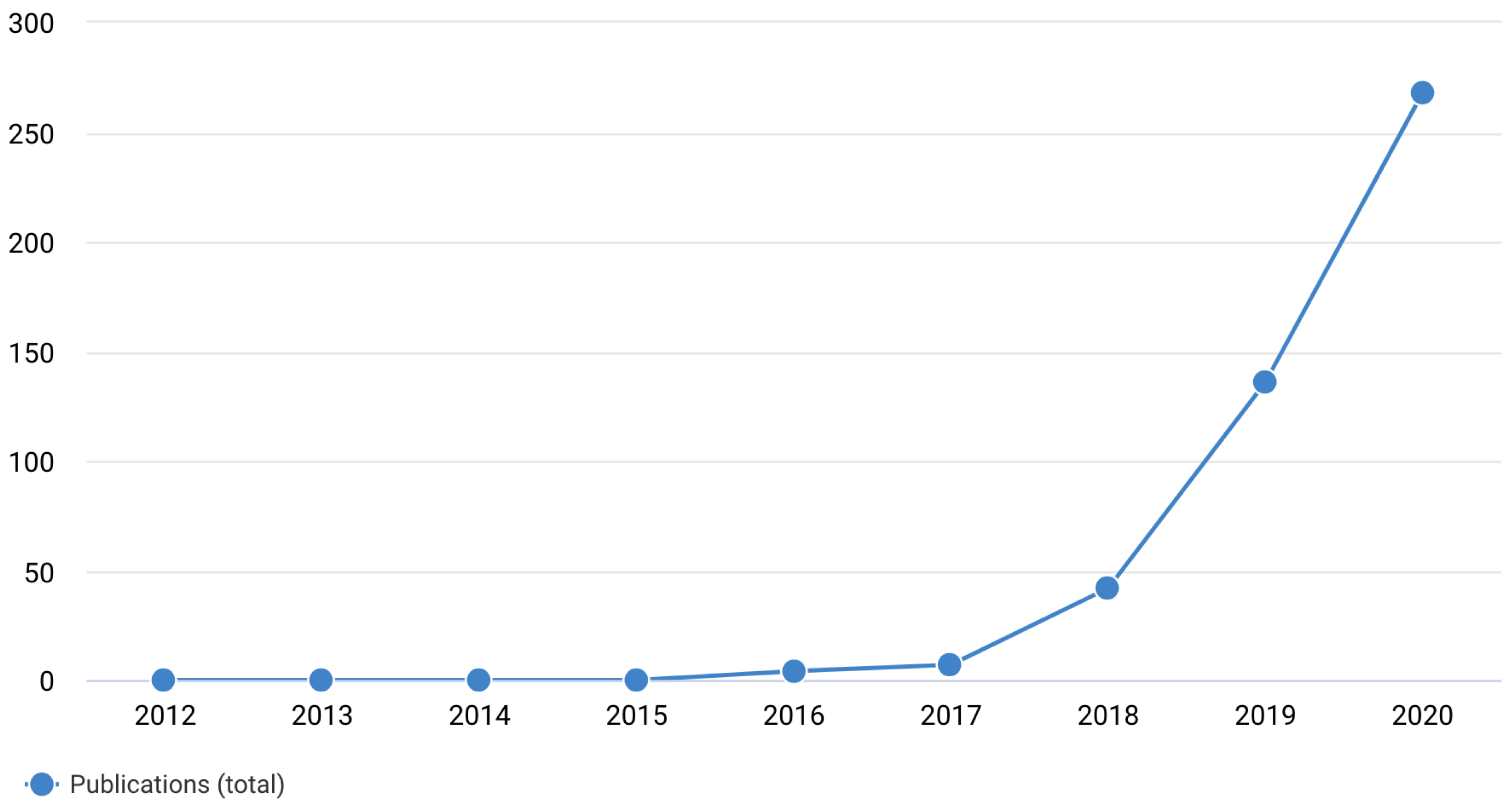
Two choices:

1. Selecting augmentations
2. Aggregating the resulting predictions

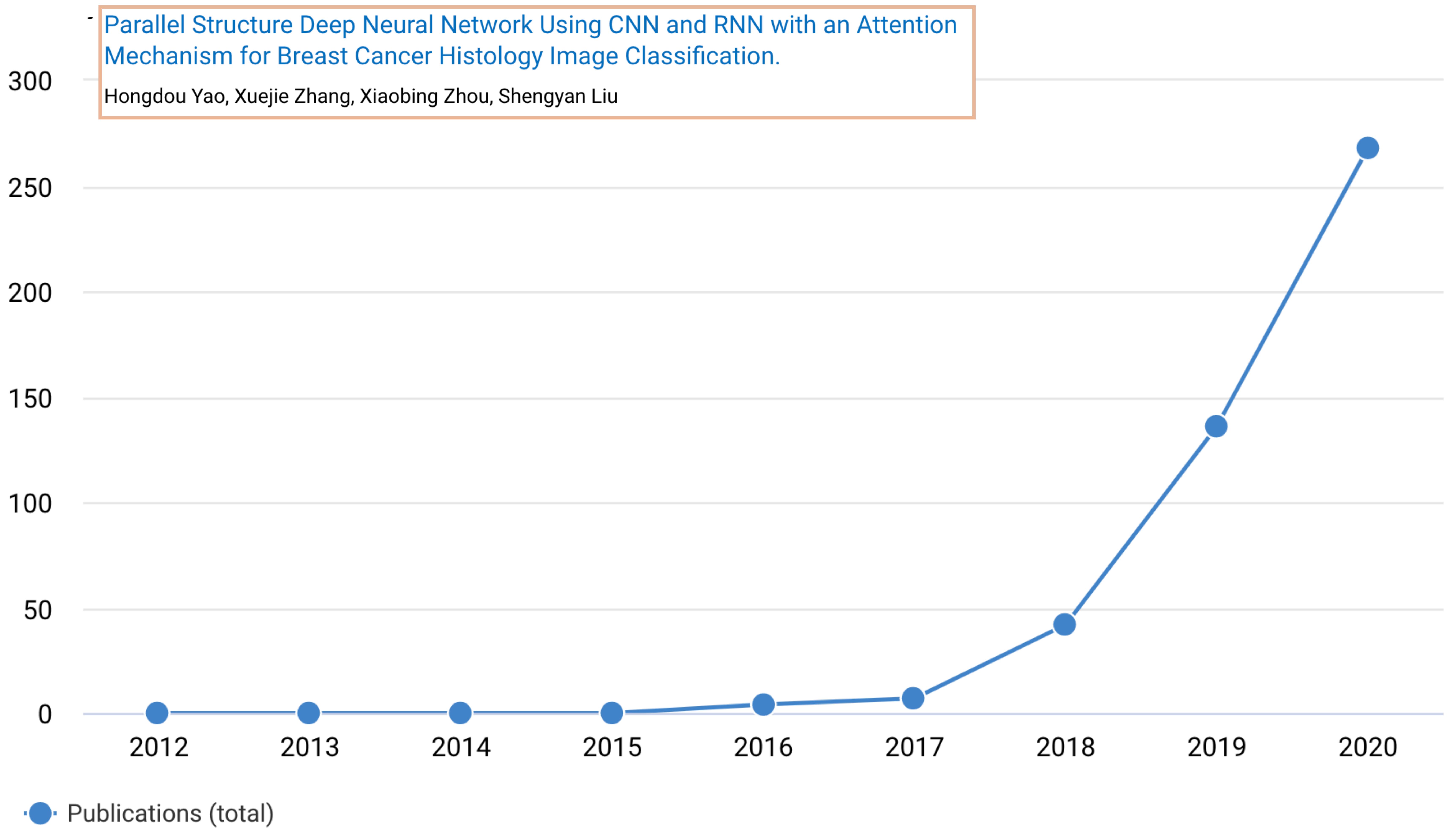
Common augmentations include **flips, crops, and scales**, and predictions are typically aggregated via a **simple average.**



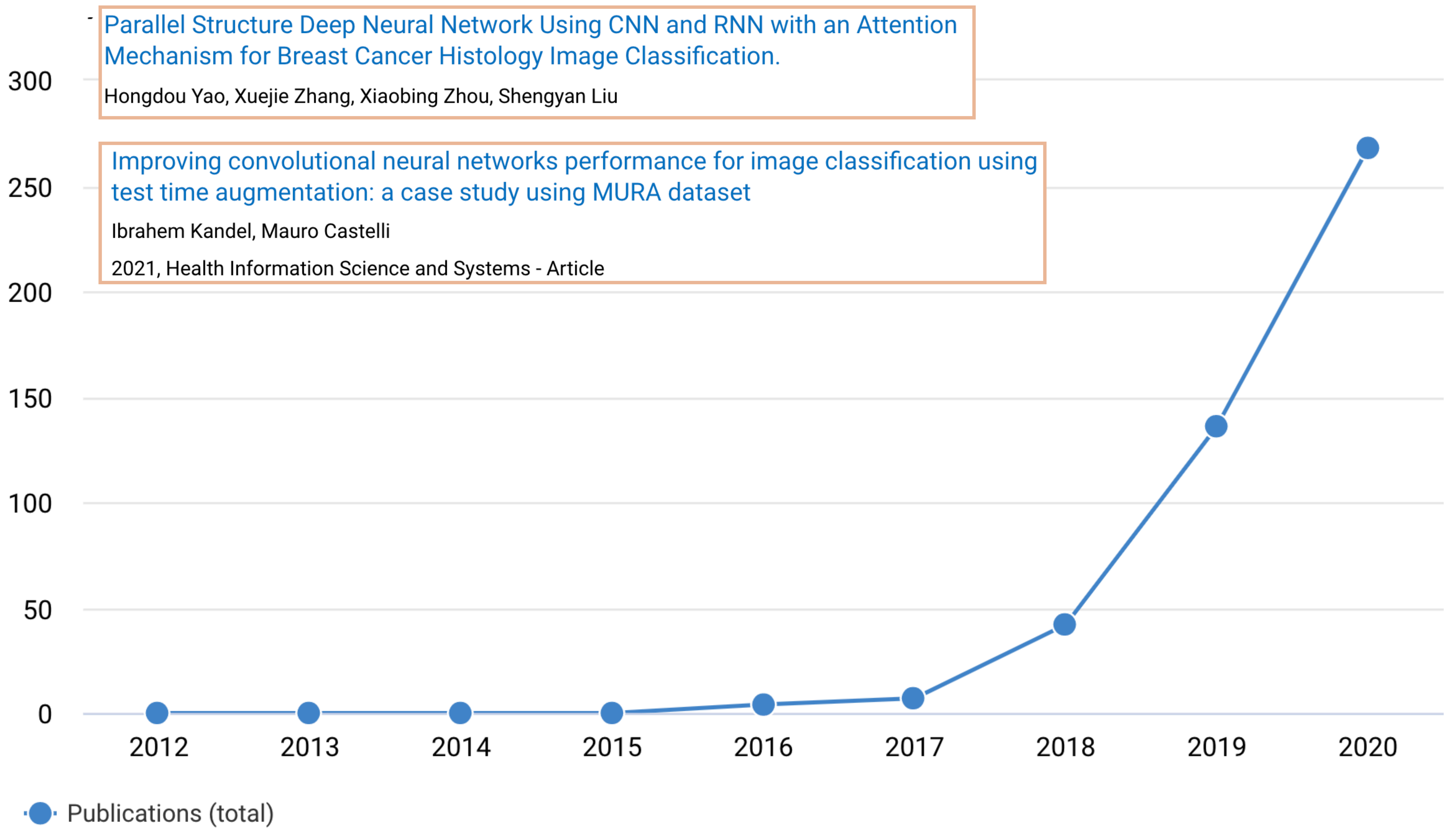
TTA is widely applied.



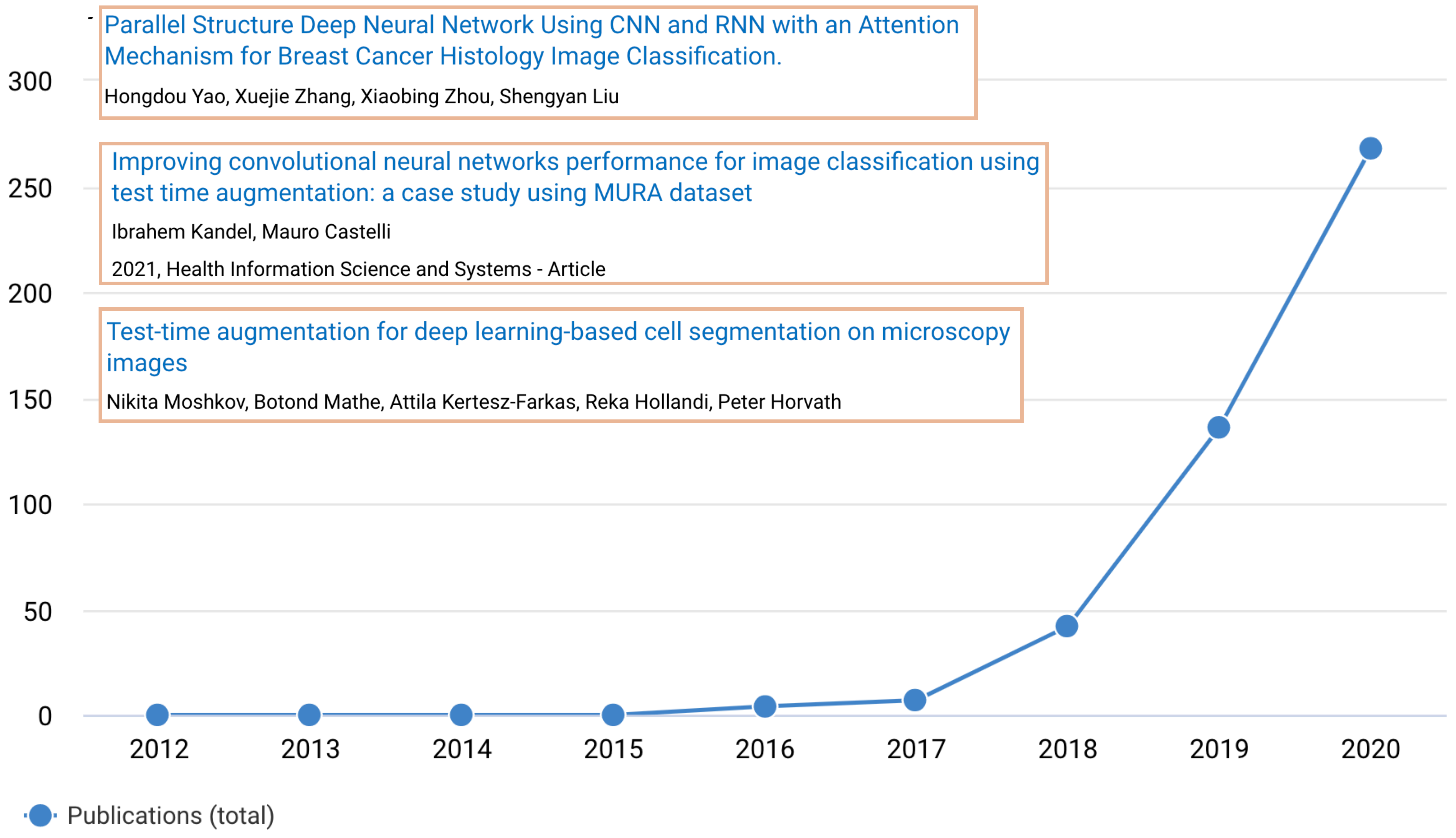
TTA is widely applied.



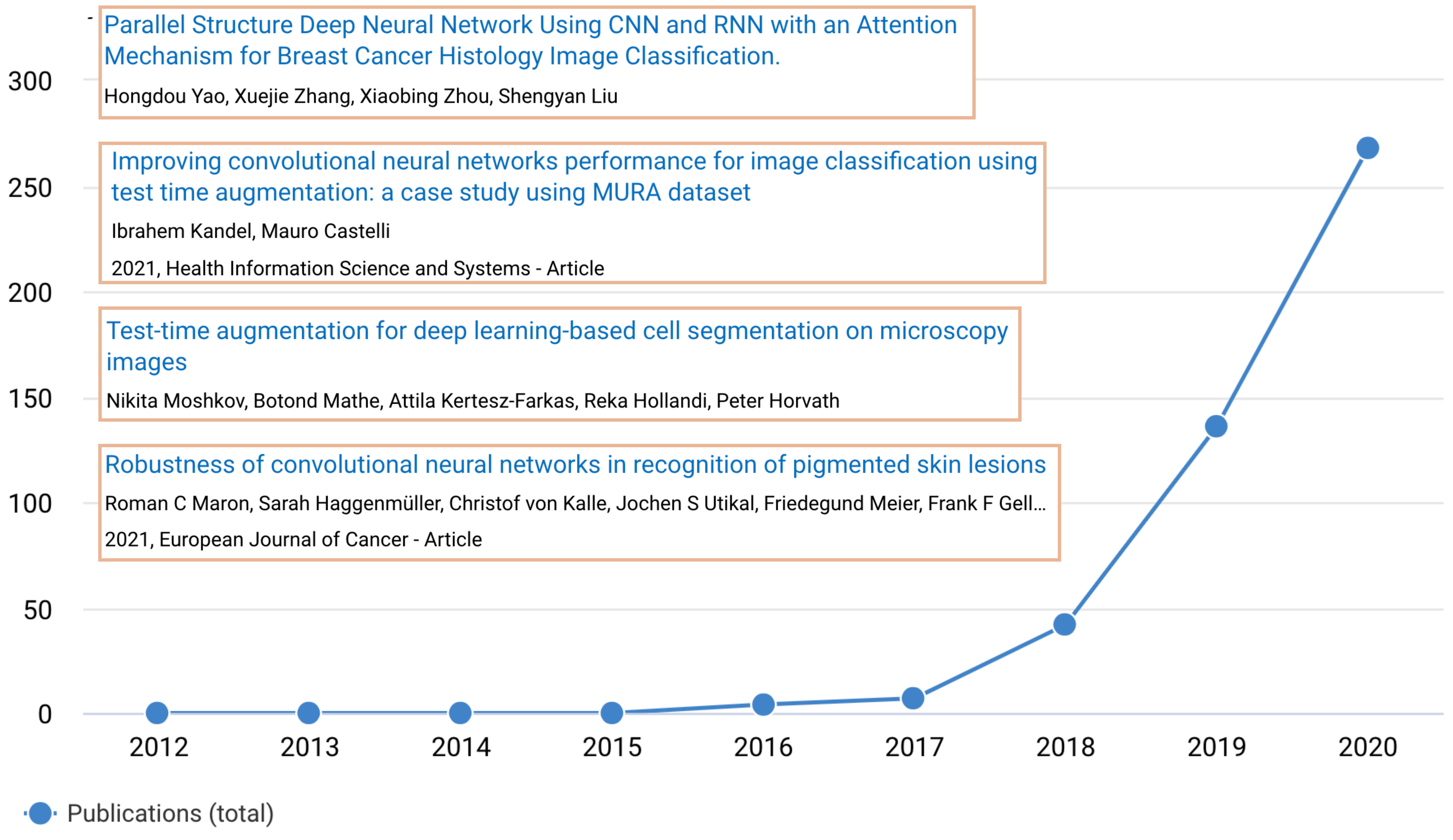
TTA is widely applied.



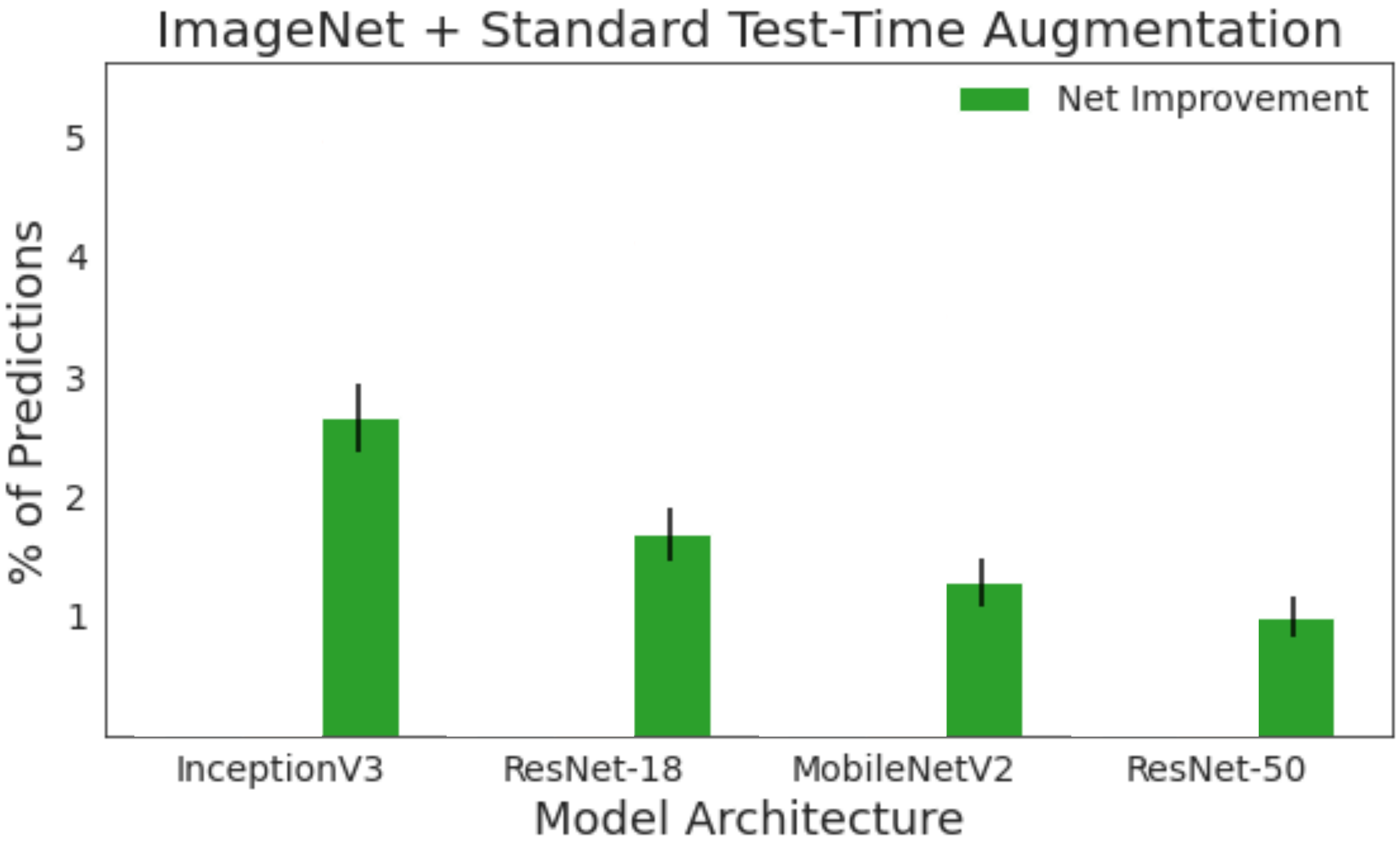
TTA is widely applied.



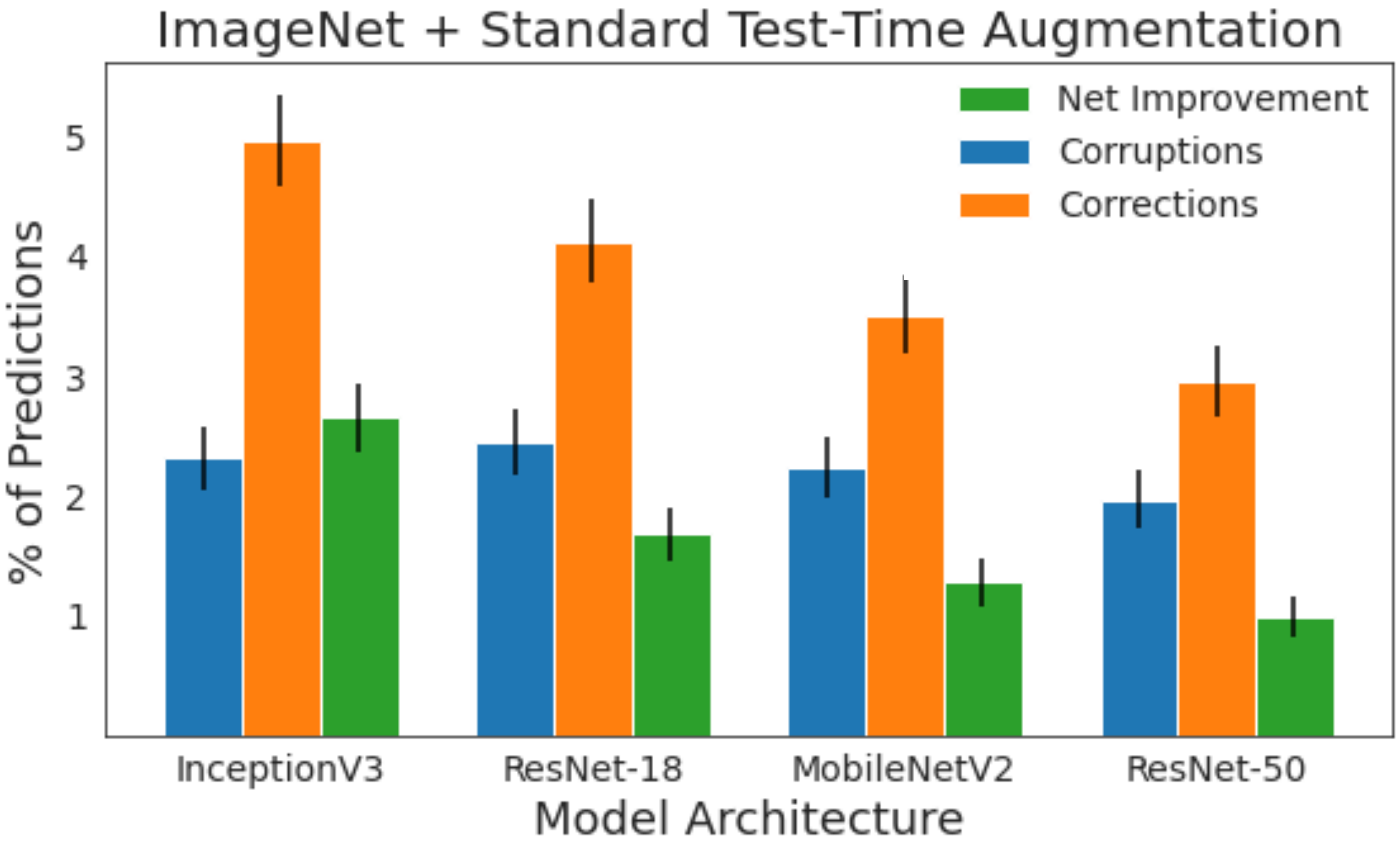
TTA is widely applied.



Standard approaches to TTA work consistently improve network performance.



Standard approaches to TTA change many predictions from correct to incorrect.



Our plan

Characterize the errors introduced by TTA.

Present a new TTA method that addresses these shortcomings.

Datasets we considered:

ImageNet: 1000 classes, 1.2 million images

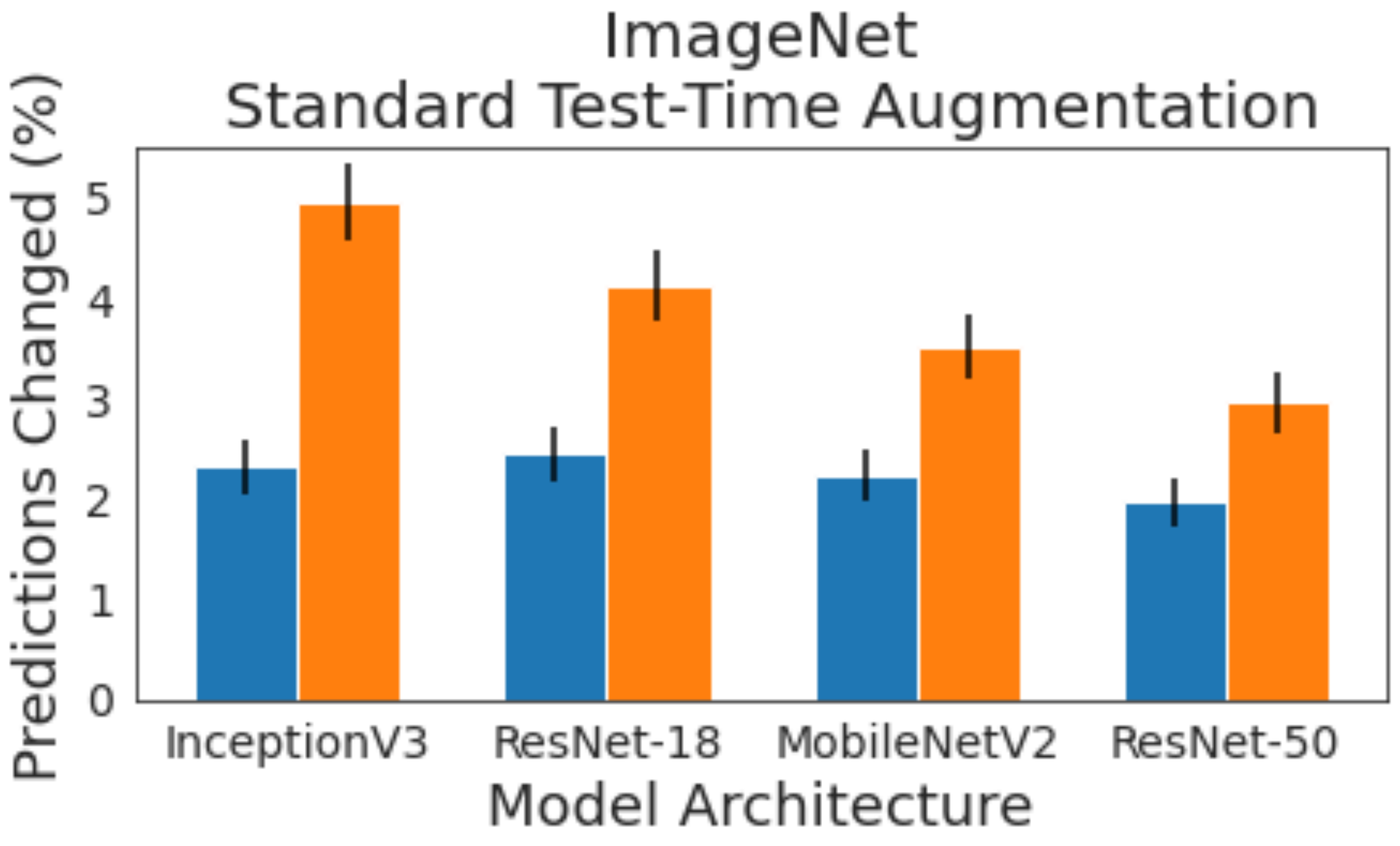
Datasets we considered:

ImageNet: 1000 classes, 1.2 million images

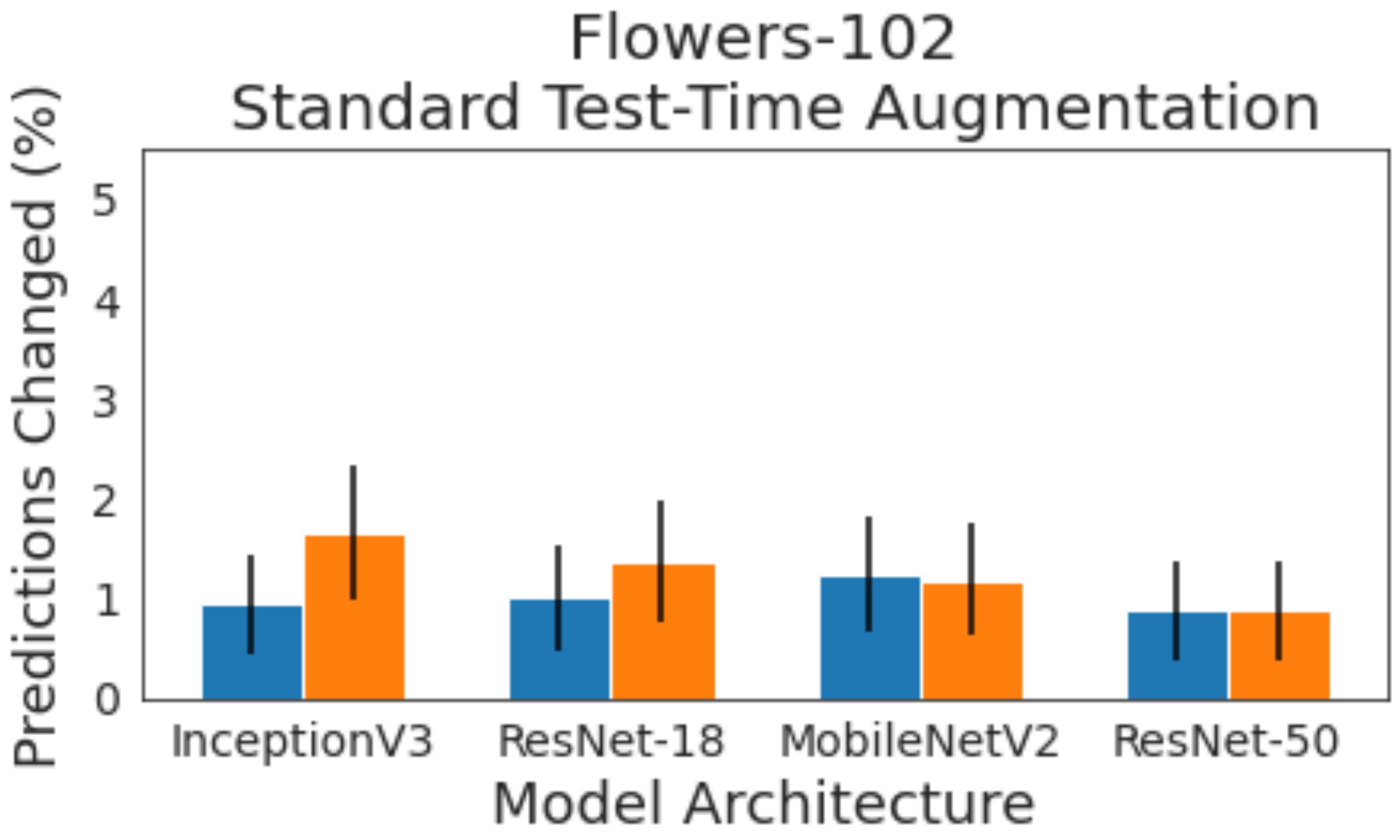
Flowers-102: 102 classes, 1020 images

Datasets we considered:

ImageNet: 1000 classes, 1.2 million images



Flowers-102: 102 classes, 1020 images



Understanding why corruptions occur

True Label:
Ibizan Hound

Zooming in on images with **multiple classes** favors classes that appear smaller.

True Label:
Ibizan Hound

Test-Time Augmentations of Original Image
(Flips, Crops, and Scales)

TTA Label:
Tennis Ball

TTA can also benefit classes differently because of **class-dependent variation.**

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

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

Class-specific and dataset-specific attributes can affect the performance of traditional TTA.

Key idea: Learn augmentation-specific weights for aggregating predictions.

Key idea: Learn augmentation-specific weights for aggregating predictions.

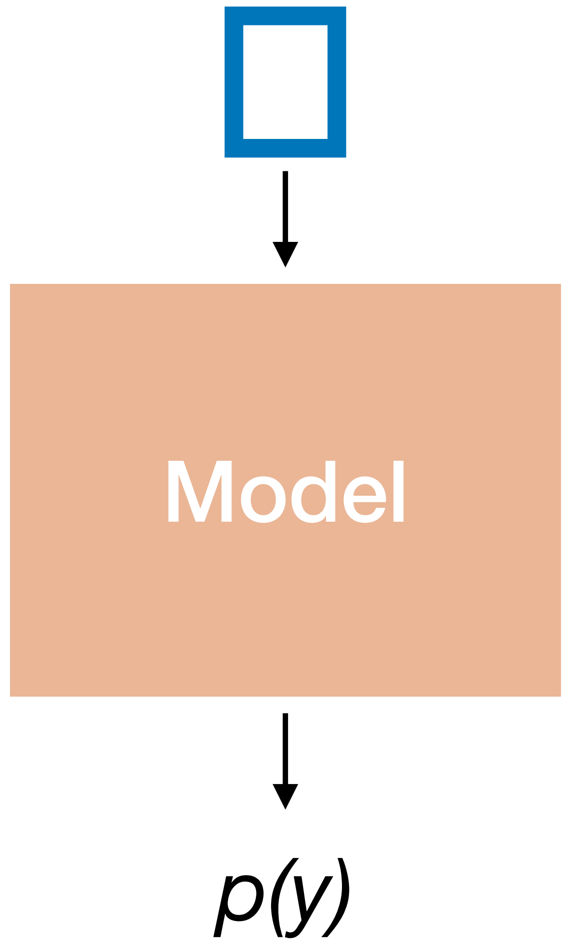
We assume three inputs:

Key idea: Learn augmentation-specific weights for aggregating predictions.

We assume three inputs:

1

Black box classifier

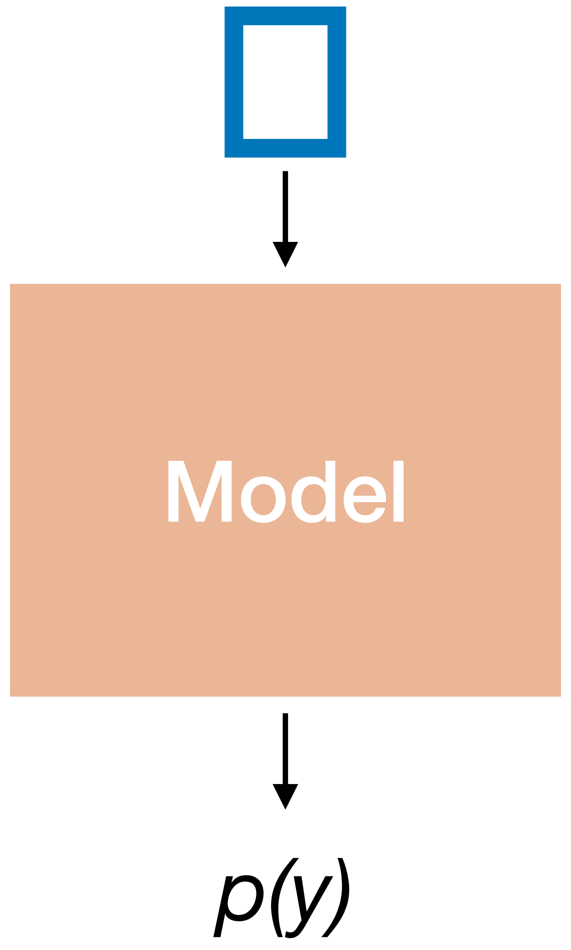


Key idea: Learn augmentation-specific weights for aggregating predictions.

We assume three inputs:

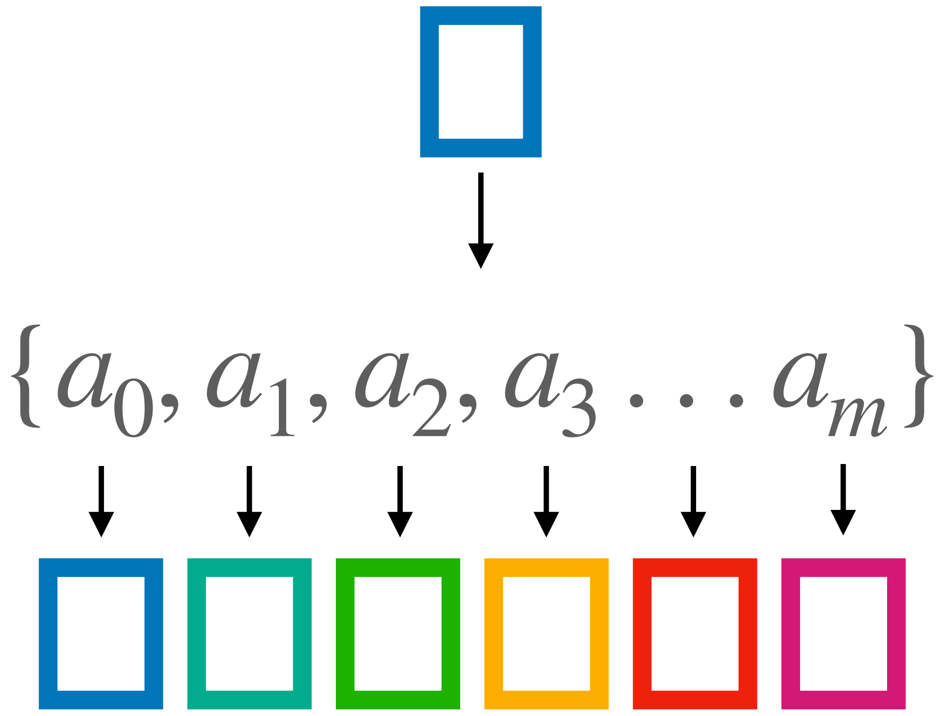
1

Black box classifier



2

Augmentation policy

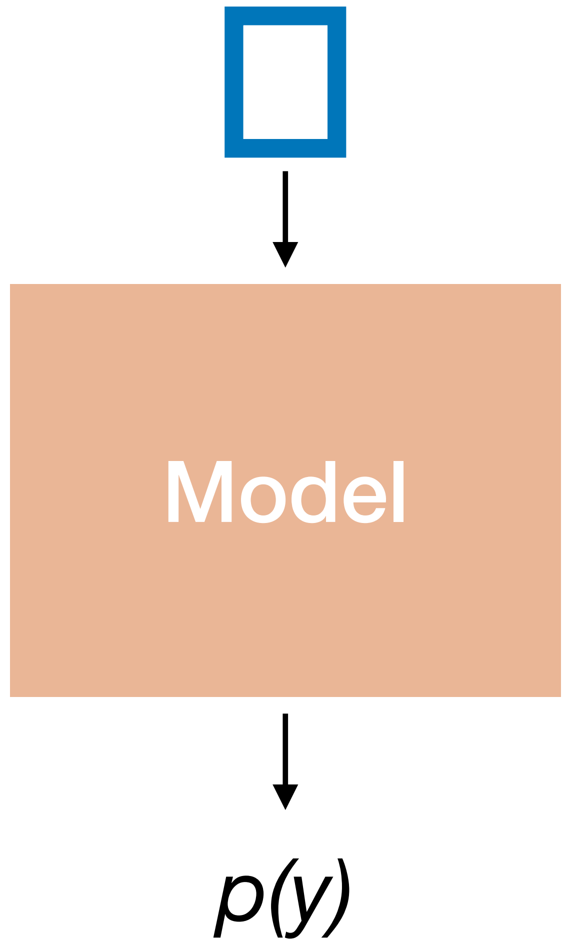


Key idea: Learn augmentation-specific weights for aggregating predictions.

We assume three inputs:

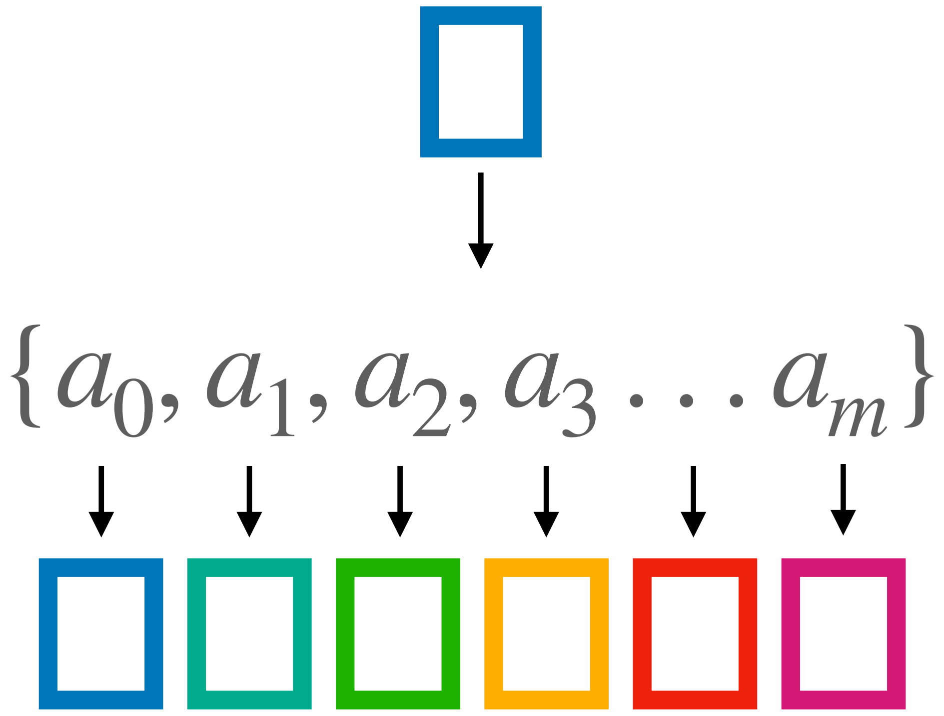
1

Black box classifier



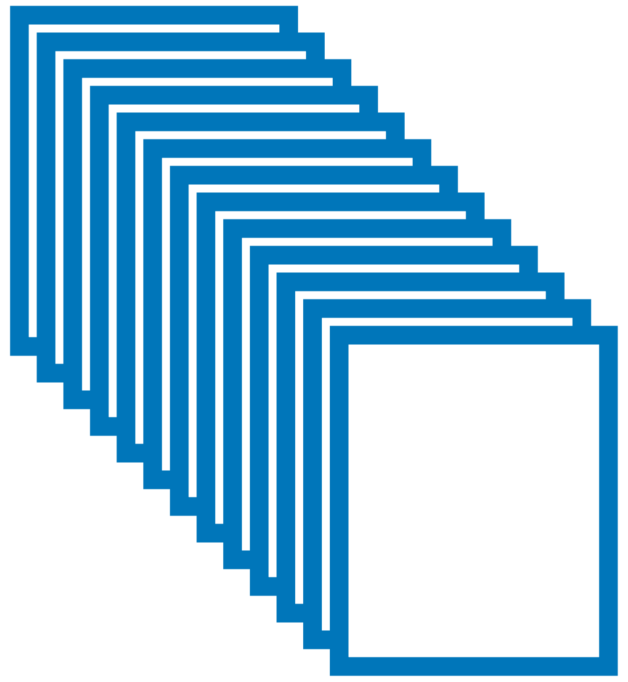
2

Augmentation policy



3

Labeled set of images



Key idea: Learn augmentation-specific weights for aggregating predictions.

Two models:

1) Learn a weight parameter for each augmentation
2) Learn a weight parameter for each augmentation-class pair

Key idea: Learn augmentation-specific weights for aggregating predictions.

Two models:
1) Learn a weight parameter for each augmentation
2) Learn a weight parameter for each augmentation-class pair

AugTTA

$$\underbrace{\begin{bmatrix} \theta_1 & \dots & \theta_M \end{bmatrix}}_{\text{Weights for each augmentation}} \begin{bmatrix} a_{11} & \dots & a_{1C} \\ \vdots & \ddots & \\ a_{M1} & & a_{MC} \end{bmatrix}$$

Key idea: Learn augmentation-specific weights for aggregating predictions.

Two models:
1) Learn a weight parameter for each augmentation
2) Learn a weight parameter for each augmentation-class pair

AugTTA

$$\underbrace{\begin{bmatrix}\theta_1 & \dots & \theta_M\end{bmatrix}}_{\text{Weights for each augmentation}} \begin{bmatrix} a_{11} & \dots & a_{1C} \\ \vdots & \ddots & \\ a_{M1} & & a_{MC} \end{bmatrix}$$

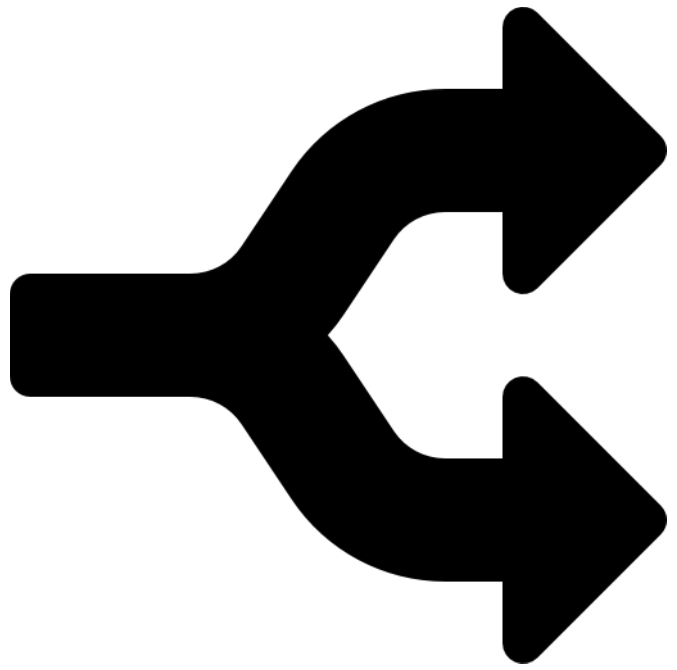
ClassTTA

$$\mathbf{1}^T \underbrace{\begin{bmatrix} \theta_{11} & \dots & \theta_{1C} \\ \vdots & \ddots & \\ \theta_{M1} & & \theta_{MC} \end{bmatrix}}_{\text{Weights for each class-augmentation pair}} \begin{bmatrix} a_{11} & \dots & a_{1C} \\ \vdots & \ddots & \\ a_{M1} & & a_{MC} \end{bmatrix}$$

Our method in three steps:

1

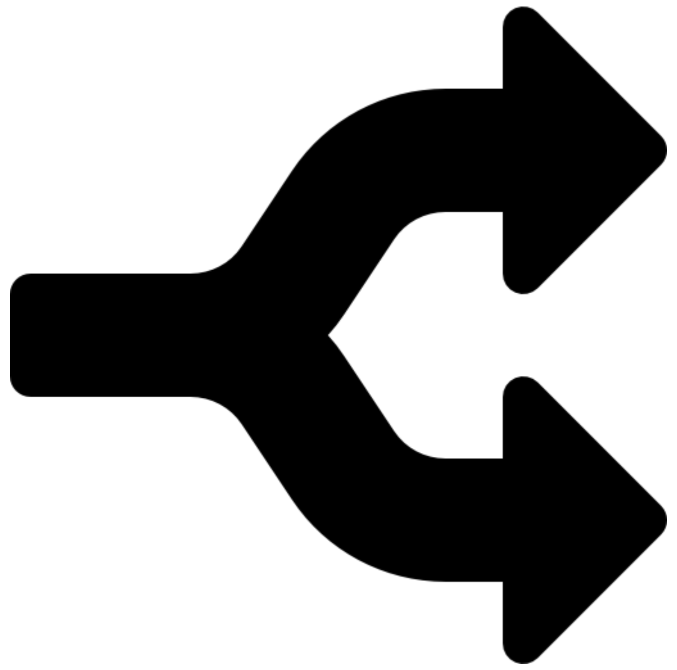
Split the labeled data into data for training the models, and data for deciding which model to use.



Our method in three steps:

1

Split the labeled data into data for training the models, and data for deciding which model to use.

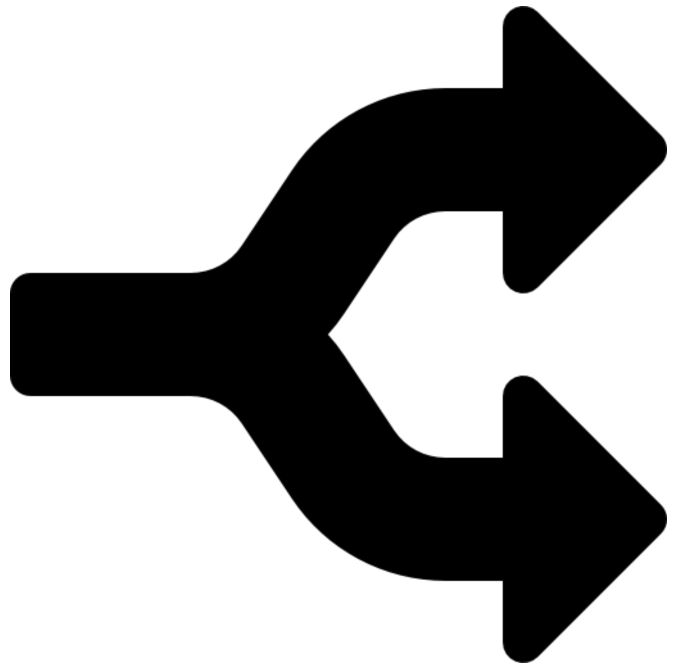


2

Learn the parameters for AugTTA and ClassTTA using projected gradient descent to ensure learned weights are non-negative.

Our method in three steps:

1. Split the labeled data into data for training the models, and data for deciding which model to use.



2. Learn the parameters for AugTTA and ClassTTA using projected gradient descent to ensure learned weights are non-negative.

3. Choose AugTTA or ClassTTA based on performance on the held-out data.

Our method produces higher Top-1 classification accuracy than existing work.

Standard TTA Policy.

Dataset	Model	Original	Max	Mean	GPS	Ours
Flowers102	MobileNetV2	90.28 ± 0.10				

Our method produces higher Top-1 classification accuracy than existing work.

Standard TTA Policy.

Dataset	Model	Original	Max	Mean	GPS	Ours
Flowers102	MobileNetV2	90.28 ± 0.10	90.17 ± 0.25			

Our method produces higher Top-1 classification accuracy than existing work.

Standard TTA Policy.

Dataset	Model	Original	Max	Mean	GPS	Ours
Flowers102	MobileNetV2	90.28 ± 0.10	90.17 ± 0.25	90.47 ± 0.20		

Our method produces higher Top-1 classification accuracy than existing work.

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	

Our method produces higher Top-1 classification accuracy than existing work.

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	$\mathbf{92.62} \pm \mathbf{0.10}$

Our method produces higher Top-1 classification accuracy than existing work.

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	$\mathbf{92.62 \pm 0.10}$
Flowers102	InceptionV3	89.28 ± 0.08	89.59 ± 0.15	90.07 ± 0.22	89.93 ± 0.16	$\mathbf{91.16 \pm 0.21}$
Flowers102	ResNet-18	89.78 ± 0.17	89.47 ± 0.11	90.21 ± 0.23	90.01 ± 0.22	$\mathbf{91.02 \pm 0.17}$
Flowers102	ResNet-50	$\mathbf{91.72 \pm 0.18}$	91.61 ± 0.08	$\mathbf{91.96 \pm 0.27}$	$\mathbf{92.03 \pm 0.09}$	$\mathbf{92.02 \pm 0.16}$
ImageNet	MobileNetV2	71.38 ± 0.06	72.50 ± 0.13	$\mathbf{72.69 \pm 0.06}$	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	$\mathbf{72.79 \pm 0.02}$
ImageNet	ResNet-18	69.37 ± 0.1	70.26 ± 0.13	$\mathbf{71.02 \pm 0.13}$	70.8 ± 0.1	$\mathbf{71.06 \pm 0.10}$
ImageNet	ResNet-50	75.78 ± 0.08	76.62 ± 0.08	$\mathbf{76.91 \pm 0.09}$	$\mathbf{76.73 \pm 0.11}$	$\mathbf{76.75 \pm 0.14}$
CIFAR100	CNN-7	74.15 ± 0.18	75.00 ± 0.31	75.48 ± 0.11	75.45 ± 0.21	$\mathbf{75.92 \pm 0.20}$
STL10	CNN-5	77.92 ± 0.19	77.76 ± 0.22	$\mathbf{78.58 \pm 0.25}$	$\mathbf{78.32 \pm 0.17}$	$\mathbf{78.52 \pm 0.31}$

TTA + smaller networks can exceed original performance of larger networks.

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	$\mathbf{92.62 \pm 0.10}$
Flowers102	InceptionV3	89.28 ± 0.08	89.59 ± 0.15	90.07 ± 0.22	89.93 ± 0.16	$\mathbf{91.16 \pm 0.21}$
Flowers102	ResNet-18	89.78 ± 0.17	89.47 ± 0.11	90.21 ± 0.23	90.01 ± 0.22	$\mathbf{91.02 \pm 0.17}$
Flowers102	ResNet-50	$\mathbf{91.72 \pm 0.18}$	91.61 ± 0.08	$\mathbf{91.96 \pm 0.27}$	$\mathbf{92.03 \pm 0.09}$	$\mathbf{92.02 \pm 0.16}$
ImageNet	MobileNetV2	71.38 ± 0.06	72.50 ± 0.13	$\mathbf{72.69 \pm 0.06}$	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	$\mathbf{72.79 \pm 0.02}$
ImageNet	ResNet-18	69.37 ± 0.1	70.26 ± 0.13	$\mathbf{71.02 \pm 0.13}$	70.8 ± 0.1	$\mathbf{71.06 \pm 0.10}$
ImageNet	ResNet-50	75.78 ± 0.08	76.62 ± 0.08	$\mathbf{76.91 \pm 0.09}$	$\mathbf{76.73 \pm 0.11}$	$\mathbf{76.75 \pm 0.14}$
CIFAR100	CNN-7	74.15 ± 0.18	75.00 ± 0.31	75.48 ± 0.11	75.45 ± 0.21	$\mathbf{75.92 \pm 0.20}$
STL10	CNN-5	77.92 ± 0.19	77.76 ± 0.22	$\mathbf{78.58 \pm 0.25}$	$\mathbf{78.32 \pm 0.17}$	$\mathbf{78.52 \pm 0.31}$

TTA + smaller networks can exceed original performance of larger networks.

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	$\mathbf{92.62 \pm 0.10}$
Flowers102	InceptionV3	89.28 ± 0.08	89.59 ± 0.15	90.07 ± 0.22	89.93 ± 0.16	$\mathbf{91.16 \pm 0.21}$
Flowers102	ResNet-18	89.78 ± 0.17	89.47 ± 0.11	90.21 ± 0.23	90.01 ± 0.22	$\mathbf{91.02 \pm 0.17}$
Flowers102	ResNet-50	$\mathbf{91.72 \pm 0.18}$	91.61 ± 0.08	$\mathbf{91.96 \pm 0.27}$	$\mathbf{92.03 \pm 0.09}$	$\mathbf{92.02 \pm 0.16}$
ImageNet	MobileNetV2	71.38 ± 0.06	72.50 ± 0.13	$\mathbf{72.69 \pm 0.06}$	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	$\mathbf{72.79 \pm 0.02}$
ImageNet	ResNet-18	69.37 ± 0.1	70.26 ± 0.13	$\mathbf{71.02 \pm 0.13}$	70.8 ± 0.1	$\mathbf{71.06 \pm 0.10}$
ImageNet	ResNet-50	75.78 ± 0.08	76.62 ± 0.08	$\mathbf{76.91 \pm 0.09}$	$\mathbf{76.73 \pm 0.11}$	$\mathbf{76.75 \pm 0.14}$
CIFAR100	CNN-7	74.15 ± 0.18	75.00 ± 0.31	75.48 ± 0.11	75.45 ± 0.21	$\mathbf{75.92 \pm 0.20}$
STL10	CNN-5	77.92 ± 0.19	77.76 ± 0.22	$\mathbf{78.58 \pm 0.25}$	$\mathbf{78.32 \pm 0.17}$	$\mathbf{78.52 \pm 0.31}$

The weights learned by ClassTTA reflect variation in the training data.

Low Variance in Augmentation Weights

High Variance in Augmentation Weights

Our method improves classification accuracy and is **nearly free** in terms of model size, training time, and implementation burden.

The learned weights shed light on 1) dataset-specific and class-specific robustness to specific augmentations and 2) which classes exhibit higher variation in the training data.

In summary:

* Class-specific and dataset-specific attributes have systematic effects on the performance of common approaches to TTA.
* We share insights on when TTA is likely to be successful and which classes are negatively affected by the use of TTA.
* We develop a method that increases the classification accuracy of a pre-trained network.

Visit our poster to learn more!

(or email me at divyas@mit.edu)