

Understanding Test-Time Augmentation

Masanari Kimura¹[0000–0002–9953–3469]

Ridge-i Inc., Tokyo, Japan
mkimura@ridge-i.com

Abstract. Test-Time Augmentation (TTA) is a very powerful heuristic that takes advantage of data augmentation during testing to produce averaged output. Despite the experimental effectiveness of TTA, there is insufficient discussion of its theoretical aspects. In this paper, we aim to give theoretical guarantees for TTA and clarify its behavior.

Keywords: data augmentation, ensemble learning, machine learning

1 Introduction

The effectiveness of machine learning has been reported for a great variety of tasks [3, 11, 14, 15, 22]. However, satisfactory performance during testing is often not achieved due to the lack of training data or the complexity of the model.

One important concept to tackle such problems is data augmentation. The basic idea of data augmentation is to increase the training data by transforming the input data in some way to generate new data that resembles the original instance. Many data augmentations have been proposed [13, 25, 29, 37], ranging from simple ones, such as flipping input images [20, 26], to more complex ones, such as leveraging Generative Adversarial Networks (GANs) to automatically generate data [7, 8]. In addition, there are several studies on automatic data augmentation in the framework of AutoML [9, 18].

Another approach to improve the performance of machine learning models is ensemble learning [4, 27]. Ensemble learning generates multiple models from a single training dataset and combines their outputs, hoping to outperform a single model. The effectiveness of ensemble learning has also been reported in a number of domains [5, 6, 17].

Influenced by these approaches, a new paradigm called Test-Time Augmentation (TTA) [23, 34, 35] has been gaining attention in recent years. TTA is a very powerful heuristic that takes advantage of data augmentation during testing to produce averaged output. Despite the experimental effectiveness of TTA, there is insufficient discussion of its theoretical aspects. In this paper, we aim to give theoretical guarantees for TTA and clarify its behavior. Our contributions are summarized as follows:

- We prove that the expected error of the TTA is less than or equal to the average error of an original model. Furthermore, under some assumptions, the expected error of the TTA is strictly less than the average error of an original model;

- We introduce the generalized version of the TTA, and the optimal weights of it are given by the closed-form;
- We prove that the error of the TTA depends on the ambiguity term.

2 Preliminaries

Here, we first introduce the notations and problem formulation.

2.1 Problem formulation

Let $\mathcal{X} \in \mathbb{R}^d$ be the d -dimensional input space, $\mathcal{Y} \in \mathbb{R}$ be the output space, and $\mathcal{H} = \{h(\mathbf{x}; \boldsymbol{\theta}) : \mathcal{X} \rightarrow \mathcal{Y} \mid \boldsymbol{\theta} \in \Theta\}$ be a hypothesis class, where $\Theta \subset \mathbb{R}^p$ is the p -dimensional parameter space. In supervised learning, our goal is to obtain $h^* \in \mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}$ such that

$$h^* = \arg \min_{h \in \mathcal{H}} \mathcal{R}^\ell(h) = \arg \min_{h \in \mathcal{H}} \mathbb{E} [\ell(y, h(\mathbf{x}; \boldsymbol{\theta}))], \quad (1)$$

where

$$\mathcal{R}^\ell(h) := \mathbb{E} [\ell(y, h(\mathbf{x}; \boldsymbol{\theta}))] \quad (2)$$

is the expected error and $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ is some loss function. Since we can not access $\mathcal{R}^\ell(h)$ directly, we try to approximate $\mathcal{R}^\ell(h)$ from the limited sample $S = \{(y_i, \mathbf{x}_i)\}_{i=1}^N$ of size $N \in \mathbb{N}$. It is the ordinal empirical risk minimization (ERM) problem, and the minimizer of the empirical error $\hat{\mathcal{R}}_S^\ell := \frac{1}{N} \sum_{i=1}^N \ell(y_i, h(\mathbf{x}_i))$ can be calculated as

$$\hat{h} = \arg \min_{h \in \mathcal{H}} \hat{\mathcal{R}}_S^\ell(h) = \arg \min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N \ell(y_i, h(\mathbf{x}_i; \boldsymbol{\theta})). \quad (3)$$

It is known that when the hypothesis class is complex (e.g., a class of neural networks), learning by ERM can lead to overlearning [10]. To tackle this problem, many approaches have been proposed, such as data augmentation [26, 31, 36] and ensemble learning [4, 5, 27]. Among such methods, Test-Time Augmentation (TTA) [23, 34, 35] is an innovative paradigm that has attracted a great deal of attention in recent years.

2.2 TTA: Test-Time Augmentation

The TTA framework is generally described as follows: let $\mathbf{x} \in \mathcal{X}$ be the new input variable at test time. We now consider multiple data augmentations $\{\tilde{\mathbf{x}}_i\}_{i=1}^m$ for \mathbf{x} , where $\tilde{\mathbf{x}}_i \in \mathbb{R}^d$ is the i -th augmented data where \mathbf{x} is transformed and m is the number of strategies for data augmentation. Finally, we compute the output \tilde{y} for the original input \mathbf{x} as $\tilde{y} = \sum_{i=1}^m h(\tilde{\mathbf{x}}_i)$. Thus, intuitively, one would expect \tilde{y} to be a better predictor than y . TTA is a very powerful heuristic, and its effectiveness has been reported for many tasks [2, 23, 30, 34, 35]. Despite its

experimental usefulness, the theoretical analysis of TTA is insufficient. In this paper, we aim to theoretically analyze the behavior of TTA. In addition, at the end of the manuscript, we provide directions for future works [28] on the theoretical analysis of TTA in light of the empirical observations given in existing studies.

3 Theoretical results for the Test-Time Augmentation

In this section, we give several theoretical results for the TTA procedure.

3.1 Re-formalization of TTA

First of all, we reformulate the TTA procedure as follows.

Definition 1. (*Augmented input space*) For the transformation class \mathcal{G} , we define the augmented input space $\bar{\mathcal{X}}$ as

$$\bar{\mathcal{X}} := \mathcal{X} \cup \left(\bigcup_{i=1}^{\infty} g(\mathcal{X}; \boldsymbol{\xi}_i) \right) = \mathcal{X} \cup \left(\bigcup_{i=1}^{\infty} \bigcup_{j=1}^{\infty} g(\mathbf{x}_j; \boldsymbol{\xi}_i) \right). \quad (4)$$

Definition 2. (*TTA as the function composition*) Let $\mathcal{F} = \{f(\mathbf{x}; \boldsymbol{\theta}_{\mathcal{F}}) \mid \boldsymbol{\theta}_{\mathcal{F}} \in \Theta_{\mathcal{F}} \subset \Theta\} \subset \mathcal{H}$ be a subset of the hypothesis class and $\mathcal{G} = \{g(\mathbf{x}; \boldsymbol{\xi}) : \mathcal{X} \rightarrow \bar{\mathcal{X}} \mid \boldsymbol{\xi} \in \Xi\}$ be the transformation class. We assume that $\{g_i = g(\mathbf{x}; \boldsymbol{\xi}_i)\}_{i=1}^m$ is a set of the data augmentation strategies, and the TTA output \tilde{y} for the input \mathbf{x} is calculated as

$$\tilde{y}(\mathbf{x}, \{\boldsymbol{\xi}_{i=1}^m\}) := \sum_{i=1}^m f \circ g_i(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m f(g(\mathbf{x}; \boldsymbol{\xi}_i); \boldsymbol{\theta}_{\mathcal{F}}). \quad (5)$$

From these definitions, we have the expected error for the TTA procedure as follows.

Definition 3. (*Expected error with TTA*) The empirical error $\mathcal{R}^{\ell, \mathcal{G}}$ of the hypothesis $h \in \mathcal{H}$ obtained by the TTA with transformation class \mathcal{G} is calculated as follows:

$$\mathcal{R}^{\ell, \mathcal{G}}(h) := \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, \tilde{y}(\mathbf{x}, \{\boldsymbol{\xi}_{i=1}^m\})) p(\mathbf{x}, y) d\mathbf{x} dy. \quad (6)$$

The next question is, whether $\mathcal{R}^{\ell, \mathcal{G}}(h)$ is less than \mathcal{R}^{ℓ} or not. In addition, if $\mathcal{R}^{\ell, \mathcal{G}}(h)$ is strictly less than \mathcal{R}^{ℓ} , it is interesting to show the required assumptions.

3.2 Upper bounds for the TTA

Next we derive the upper bounds for the TTA. For the sake of argument, we assume that $\ell(a, b) = (a - b)^2$ and we decompose the output of the hypothesis for (\mathbf{x}, y) as follows:

$$h(\mathbf{x}; \boldsymbol{\theta}) = y + \epsilon(\mathbf{x}, y; \boldsymbol{\theta}) \quad (\forall h \in \mathcal{H}). \quad (7)$$

Then, the following theorem holds.

Theorem 1. Assume that $f \circ g \in \mathcal{H}$ for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$, and \mathcal{G} contains the identity transformation $g : \mathbf{x} \mapsto \mathbf{x}$. Then, the expected error obtained by TTA is bounded from above by the average error of single hypotheses:

$$\mathcal{R}^{\ell, \mathcal{G}}(h) \leq \bar{\mathcal{R}}^\ell(h) := \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^m \ell(y, h(\mathbf{x}; \boldsymbol{\theta}_i)) \right]. \quad (8)$$

Proof. From the definition, the ordinal expected average error is calculated as

$$\bar{\mathcal{R}}^\ell(h) = \int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{m} \sum_{i=1}^m (y - h(\mathbf{x}; \boldsymbol{\theta}_i))^2 p(\mathbf{x}, y) d\mathbf{x} dy \quad (9)$$

$$= \int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{m} \sum_{i=1}^m \epsilon(\mathbf{x}, y; \boldsymbol{\theta}_i)^2 p(\mathbf{x}, y) d\mathbf{x} dy. \quad (10)$$

On the other hand, the expected error of TTA is

$$\begin{aligned} \mathcal{R}^{\ell, \mathcal{G}}(h) &= \int_{\mathcal{X} \times \mathcal{Y}} \left(y - \frac{1}{m} \sum_{i=1}^m f \circ g_i(\mathbf{x}) \right)^2 p(\mathbf{x}, y) d\mathbf{x} dy \\ &= \int_{\mathcal{X} \times \mathcal{Y}} \left(\frac{1}{m} \sum_{i=1}^m (y - f \circ g_i(\mathbf{x})) \right)^2 p(\mathbf{x}, y) d\mathbf{x} dy \end{aligned} \quad (11)$$

$$= \int_{\mathcal{X} \times \mathcal{Y}} \left(\frac{1}{m} \sum_{i=1}^m \epsilon(\mathbf{x}, y; \boldsymbol{\theta}_i) \right)^2 p(\mathbf{x}, y) d\mathbf{x} dy. \quad (12)$$

Then, from Eq. (10) and (12), we have the proof of the theorem.

By making further assumptions, we also have the following theorem.

Theorem 2. Assume that $f \circ g \in \mathcal{H}$ for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$, and \mathcal{G} contains the identity transformation $g : \mathbf{x} \mapsto \mathbf{x}$. Assume also that each ϵ has mean zero and is uncorrelated with each other:

$$\int_{\mathcal{X} \times \mathcal{Y}} \epsilon(\mathbf{x}, y; \boldsymbol{\theta}_i) p(\mathbf{x}, y) d\mathbf{x} dy = 0 \quad (\forall i \in \{1, \dots, m\}), \quad (13)$$

$$\int_{\mathcal{X} \times \mathcal{Y}} \epsilon(\mathbf{x}, y; \boldsymbol{\theta}_i) \epsilon(\mathbf{x}, y; \boldsymbol{\theta}_j) p(\mathbf{x}, y) d\mathbf{x} dy = 0 \quad (i \neq j). \quad (14)$$

In this case, the following relationship holds

$$\mathcal{R}^{\ell, \mathcal{G}}(h) = \frac{1}{m} \bar{\mathcal{R}}^\ell(h) < \bar{\mathcal{R}}^\ell(h). \quad (15)$$

Proof. From the assumptions (13), (14) and Eq. (10) and (12), the proof of the theorem can be obtained immediately.

3.3 Weighted averaging for the TTA

We consider the generalization of TTA as follows.

Definition 4. (*Weighted averaging for the TTA*) Let $\mathcal{F} = \{f(\mathbf{x}; \boldsymbol{\theta}_{\mathcal{F}}) \mid \boldsymbol{\theta}_{\mathcal{F}} \in \Theta_{\mathcal{F}} \subset \Theta\} \subset \mathcal{H}$ be a subset of the hypothesis class and $\mathcal{G} = \{g(\mathbf{x}; \boldsymbol{\xi}) : \mathcal{X} \rightarrow \bar{\mathcal{X}} \mid \boldsymbol{\xi} \in \Xi\}$ be the transformation class. We assume that $\{g_i = g(\mathbf{x}; \boldsymbol{\xi}_i)\}_{i=1}^m$ is a set of the data augmentation strategies, and the TTA output \tilde{y} for the input \mathbf{x} is calculated as

$$\tilde{y}_w(\mathbf{x}, \{\boldsymbol{\xi}_i\}) := \sum_{i=1}^m w_i f \circ g_i(\mathbf{x}) = \sum_{i=1}^m w(\boldsymbol{\xi}_i) f(g(\mathbf{x}; \boldsymbol{\xi}_i); \boldsymbol{\theta}_{\mathcal{F}}), \quad (16)$$

where $w_i = w(\boldsymbol{\xi}_i) : \Xi \rightarrow \mathbb{R}_+$ is the weighting function:

$$w_i \geq 0 \ (\forall i \in \{1, \dots, m\}), \quad \sum_{i=1}^m w_i = 1 \quad (17)$$

Then, we can obtain the expected error of Eq. (16) as follows.

Proposition 1. *The expected error of the weighted TTA is*

$$\mathcal{R}^{\ell, \mathcal{G}, w}(h) = \sum_{i=1}^m \sum_{j=i}^m w_i w_j \Gamma_{ij}, \quad (18)$$

where

$$\Gamma_{ij} = \int_{\mathcal{X} \times \mathcal{Y}} \left(y - f \circ g_i(\mathbf{x}) \right) \left(y - f \circ g_j(\mathbf{x}) \right) p(\mathbf{x}, y) d\mathbf{x} dy. \quad (19)$$

Proof. We can calculate as

$$\begin{aligned} \mathcal{R}^{\ell, \mathcal{G}, w}(h) &= \int_{\mathcal{X} \times \mathcal{Y}} \left(y - \sum_{i=1}^m w_i f \circ g_i(\mathbf{x}) \right)^2 p(\mathbf{x}, y) d\mathbf{x} dy \\ &= \int_{\mathcal{X} \times \mathcal{Y}} \left(y - \sum_{i=1}^m w_i f \circ g_i(\mathbf{x}) \right) \left(y - \sum_{j=1}^m w_j f \circ g_j(\mathbf{x}) \right) p(\mathbf{x}, y) d\mathbf{x} dy \\ &= \sum_{i=1}^m \sum_{j=i}^m w_i w_j \Gamma_{ij}. \end{aligned} \quad (20)$$

Proposition 18 implies that the expected error of the weighted TTA is highly depending on the correlations of $\{g_1, \dots, g_m\}_{i=1}^m$.

Theorem 3. (*Optimal weights for the weighted TTA*) We can obtain the optimal weights $\mathbf{w} = \{w_1, \dots, w_m\}$ for the weighted TTA as follows:

$$w_i = \frac{\sum_{j=1}^m \Gamma_{ij}^{-1}}{\sum_{k=1}^m \sum_{j=1}^m \Gamma_{kj}^{-1}}, \quad (21)$$

where Γ_{ij}^{-1} is the (i, j) -element of the inverse matrix of (Γ_{ij}) .

Proof. The optimal weights can be obtained by solving

$$\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{i=1}^m \sum_{j=1}^m w_i w_j \Gamma_{ij}. \quad (22)$$

Then, from the method of Lagrange multiplier,

$$\frac{\partial}{\partial w_k} \left\{ \sum_{i=1}^m \sum_{j=1}^m w_i w_j \Gamma_{ij} - 2\lambda \left(\sum_{i=1}^m w_i - 1 \right) \right\} = 0 \quad (23)$$

$$2 \sum_{j=1}^m w_k \Gamma_{kj} = 2\lambda \quad (24)$$

$$\sum_{j=1}^m w_k \Gamma_{kj} = \lambda. \quad (25)$$

From the condition (17), since $\sum_{i=1}^m w_i = 1$ and then, we have

$$w_i = \frac{\sum_{j=1}^m \Gamma_{ij}^{-1}}{\sum_{k=1}^m \sum_{j=1}^m \Gamma_{kj}^{-1}}. \quad (26)$$

From Theorem 3, we obtain a closed-form expression for the optimal weights of the weighted TTA. Furthermore, we see that this solution requires an invertible correlation matrix Γ . However, in TTA we consider the set of $\{f \circ g_i\}_{i=1}^m$, and all elements depend on $f \in \mathcal{F}$ in common. This means that the correlations among $\{f \circ g_i\}_{i=1}^m$ will be very high, and such correlation matrix is generally known to be singular or ill-conditioned.

3.4 Existence of the unnecessary transformation functions

To simplify the discussion, we assume that all weights are equal. Then, from Eq. (20), we have

$$\mathcal{R}^{\ell, \mathcal{G}, w}(h) = \sum_{i=1}^m \sum_{j=1}^m \Gamma_{ij} / m^2. \quad (27)$$

If we remove g_k from $\{g_1, \dots, g_m\}$, the error $\tilde{\mathcal{R}}^{\ell, \mathcal{G}, w}(h)$ is recomputed as follows.

$$\tilde{\mathcal{R}}^{\ell, \mathcal{G}, w}(h) = \sum_{\substack{i=1 \\ i \neq k}}^m \sum_{\substack{j=1 \\ j \neq k}}^m \Gamma_{ij} / (m-1)^2. \quad (28)$$

Here we consider how the error of the TTA changes when we remove the k -th data augmentation. If we assume that $\mathcal{R}^{\ell, \mathcal{G}, w}(h)$ is greater than or equal to $\tilde{\mathcal{R}}^{\ell, \mathcal{G}, w}(h)$, then

$$(2m-1) \sum_{i=1}^m \sum_{j=1}^m \Gamma_{ij} \leq 2m^2 \sum_{\substack{i=1 \\ i \neq k}}^m \Gamma_{ik} + m^2 \Gamma_{kk}. \quad (29)$$

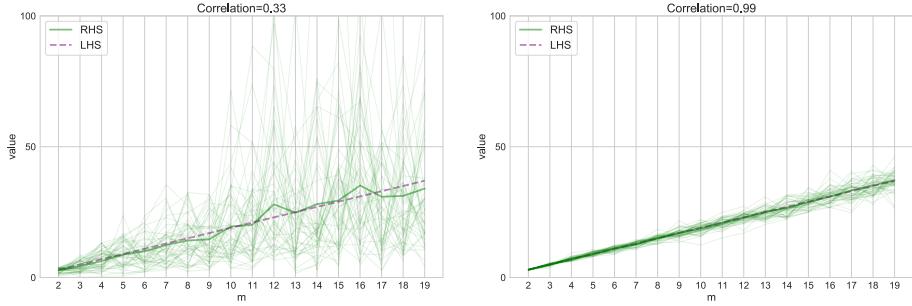


Fig. 1. $(2m - 1) \sum_{i=1}^m \sum_{j=1}^m \Gamma_{ij} = \text{LHS}$ vs $\text{RHS} = 2m^2 \sum_{i \neq k}^m \Gamma_{ik} + m^2 \Gamma_{kk}$ (Eq. (29)). When the correlation is 0.33, the numerical calculation yields $\Pr(\text{RHS} \geq \text{LHS}) \approx 0.38$. On the other hand, when the correlation is 0.99, we yields $\Pr(\text{RHS} \geq \text{LHS}) \approx 0.49$.

From the above equation, we can see that the group of data augmentations with very high correlation is redundant, except for some of them. Fig. 1 shows an example of a numerical experiment to get the probability that Eq. (29) holds. In this numerical experiment, we generated a sequence of random values with the specified correlation to obtain a pseudo (Γ_{ij}) , and calculated the probability that Eq. (29) holds out of 100 trials. From this plot, we can see that (Γ_{ij}) with high correlation is likely to have redundancy. In the following, we introduce ambiguity as another measure of redundancy and show that this measure is highly related to the error of TTA.

3.5 Error decomposition for the TTA

Knowing what elements the error can be broken down into is one important way to understand the behavior of TTA. For this purpose, we introduce the following notion of ambiguity.

Definition 5. (*Ambiguity of the hypothesis set [16]*) For some $\mathbf{x} \in \mathcal{X}$, the ambiguity $\varsigma(h_i|\mathbf{x})$ of the hypothesis set $h = \{h_i\}^m$ is defined as

$$\varsigma(h_i|\mathbf{x}) := \left(h_i(\mathbf{x}) - \sum_{i=1}^m w_i h_i(\mathbf{x}) \right)^2 \quad (\forall i \in \{1, \dots, m\}). \quad (30)$$

Let $\bar{\varsigma}(h|\mathbf{x})$ be the average ambiguity: $\bar{\varsigma}(h|\mathbf{x}) = \sum_{i=1}^m w_i \varsigma(h_i|\mathbf{x})$. From Definition 5, the ambiguity term can be regarded as a measure of the discrepancy between individual hypotheses for input \mathbf{x} . Then, we have

$$\bar{\varsigma}(h|\mathbf{x}) = \sum_{i=1}^m w_i (y - f \circ g_i(\mathbf{x}))^2 - (y - \sum_{i=1}^m w_i f \circ g_i(\mathbf{x}))^2. \quad (31)$$

Since Eq. (31) holds for all $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned} & \sum_{i=1}^m w_i \int_{\mathcal{X} \times \mathcal{Y}} \varsigma(h_i | \mathbf{x}) p(\mathbf{x}, y) d\mathbf{x} dy \\ &= \sum_{i=1}^m w_i \int (y - f \circ g_i(\mathbf{x}))^2 p(\mathbf{x}, y) d\mathbf{x} dy - \int \left(y - \sum_{i=1}^m w_i f \circ g_i(\mathbf{x}) \right)^2 p(\mathbf{x}, y) d\mathbf{x} dy. \end{aligned}$$

Let

$$err(f \circ g_i) = \mathbb{E}[(y - f \circ g_i(\mathbf{x}))^2] = \int (y - f \circ g_i(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} dy, \quad (32)$$

and

$$\varsigma(f \circ g_i) = \mathbb{E}[\varsigma(f \circ g_i | \mathbf{x})] = \int \varsigma(f \circ g_i | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (33)$$

Then, we have

$$\mathcal{R}^{\ell, \mathcal{G}, w}(h) = \sum_{i=1}^m w_i \cdot err(f \circ g_i) - \sum_{i=1}^m w_i \cdot \varsigma(f \circ g_i), \quad (34)$$

where the first term corresponds to the error, and the second term corresponds to the ambiguity. From this equation, it can be seen that TTA yields significant benefits when each $f \circ g_i$ is more accurate and more diverse than the other.

To summarize, we have the following proposition.

Proposition 2. *The error of the TTA can be decomposed as*

$$\mathcal{R}^{\ell, \mathcal{G}, w}(h) = [errors of f \circ g_i] + [ambiguites of f \circ g_i]. \quad (35)$$

3.6 Statistical consistency

Finally, we discuss the statistical consistency for the TTA procedure.

Definition 6. *The ERM is the strictly consistent if for any non-empty subset $\mathcal{H}(c) = \{h \in \mathcal{H} : \mathcal{R}^\ell(h) \geq c\}$ with $c \in (-\infty, +\infty)$ the following convergence holds:*

$$\inf_{h \in \mathcal{H}(c)} \hat{\mathcal{R}}_{\mathcal{S}}^\ell(h) \xrightarrow{P} \inf_{h \in \mathcal{H}(c)} \mathcal{R}^\ell(h) \quad (N \rightarrow \infty). \quad (36)$$

Necessary and sufficient conditions for strict consistency are provided by the following theorem [32, 33].

Theorem 4. *If two real constants $a \in \mathbb{R}$ and $A \in \mathbb{R}$ can be found such that for every $h \in \mathcal{H}$ the inequalities $a \leq \mathcal{R}^\ell(h) \leq A$ hold, then the following two statements are equivalent:*

1. The empirical risk minimization is strictly consistent on the set of functions $\{\ell(y, h(\mathbf{x})) \mid h \in \mathcal{H}\}$.
2. The uniform one-sided convergence of the mean to their expectation takes place over the set of functions $\{\ell(y, h(\mathbf{x})) \mid h \in \mathcal{H}\}$, i.e.,

$$\lim_{N \rightarrow \infty} \Pr \left[\sup_{h \in \mathcal{H}} \left\{ \mathcal{R}^\ell(h) - \hat{\mathcal{R}}_\mathcal{S}^\ell(h) \right\} > \epsilon \right] = 0 \quad (\forall \epsilon > 0). \quad (37)$$

Using these concepts, we can derive the following lemma.

Lemma 1. The empirical risk $\frac{1}{m} \sum_{i=1}^m \hat{\mathcal{R}}_\mathcal{S}^{\ell, \mathcal{G}}(h)$ obtained by ERM with data augmentations $\{g_1, \dots, g_m\}_{i=1}^m$ is the consistent estimator of $\mathbb{E}_{\bar{\mathcal{X}} \times \mathcal{Y}} [\ell(y, f(\mathbf{x}))]$, i.e.,

$$\inf_{h \in \mathcal{H}(c)} \hat{\mathcal{R}}_\mathcal{S}^{\ell, \mathcal{G}}(h) \xrightarrow{P} \mathbb{E}_{\bar{\mathcal{X}} \times \mathcal{Y}} [\ell(y, f(\mathbf{x}))] \quad (N \rightarrow \infty). \quad (38)$$

Proof. Let $\bar{\mathcal{X}}$ be the augmented input space with $\{g_1, \dots, g_m\}_{i=1}^m$. Then, we have

$$\mathbb{E}_{\mathcal{X} \times \mathcal{Y}} [\hat{\mathcal{R}}_\mathcal{S}^{\ell, \mathcal{G}}(h)] = \int_{\mathcal{X} \times \mathcal{Y}} \left\{ \frac{1}{N} \sum_{i=1}^N \frac{1}{m} \sum_{j=1}^m \ell(y, f \circ g_j(\mathbf{x})) \right\} p(\mathbf{x}, y) d\mathbf{x} dy \quad (39)$$

$$= \int_{\mathcal{X} \times \mathcal{Y}} \left\{ \frac{1}{Nm} \sum_{i=1}^N \sum_{j=1}^m \ell(y, f \circ g_j(\mathbf{x})) \right\} p(\mathbf{x}, y) d\mathbf{x} dy \quad (40)$$

$$= \int_{\bar{\mathcal{X}} \times \mathcal{Y}} \left\{ \frac{1}{Nm} \sum_{i=1}^{Nm} \ell(y, f(\mathbf{x})) \right\} p(\mathbf{x}, y) d\mathbf{x} dy \quad (41)$$

$$= \int_{\bar{\mathcal{X}} \times \mathcal{Y}} \ell(y, f(\mathbf{x})) p(\mathbf{x}, y) d\mathbf{x} dy = \mathbb{E}_{\bar{\mathcal{X}} \times \mathcal{Y}} [\ell(y, f(\mathbf{x}))]. \quad (42)$$

From Lemma 1, we can confirm that the ERM with data augmentation is also minimizing the TTA error. This means that the data augmentation strategies used in TTA should also be used during training.

4 Related works

Although there is no existing research that discusses the theoretical analysis of the TTA, there are some papers that experimentally investigate the behavior of the TTA [28]. In those papers, the following results are reported:

- the benefit of TTA depends upon the model’s lack of invariance to the given Test-Time Augmentations;
- as the training sample size increases, the benefit of TTA decreases;
- when TTA was applied to two datasets, ImageNet [15] and Flowers-102 [24], the performance improvement on the Flowers-102 dataset was small.

Because of the simplicity of the concept, several variants of TTA have also been proposed [12, 19, 21]. It is also a critical research direction to consider whether a theoretical analysis of these variants is possible using the same procedure as discussed in this paper.

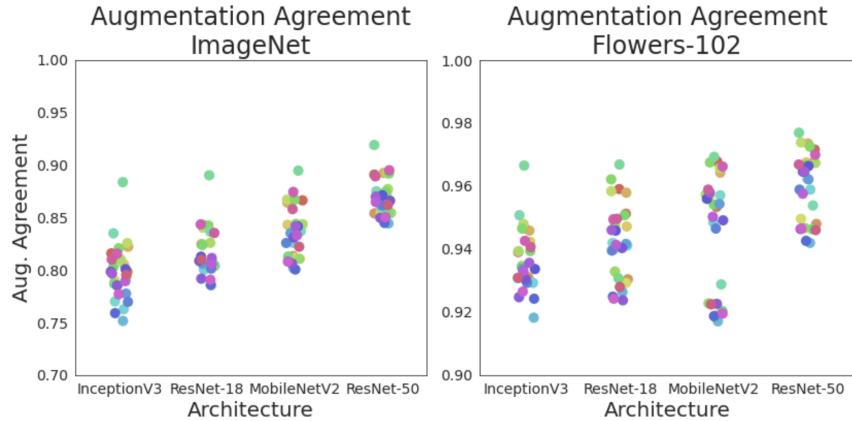


Fig. 2. Architectures that benefit least from standard TTA are also the least sensitive to the augmentations. Note that this figure is created by [28], and see their paper for more details.

5 Conclusion and Discussion

In this paper, we theoretically investigate the behavior of TTA. Our discussion shows that TTA has several theoretically desirable properties. Furthermore, we showed that the error of TTA depends on the ambiguity of the output.

5.1 Future works

In the previous work, some empirical observations are reported [28]. The future of research is to construct a theory consistent with these observations.

- When TTA was applied to two datasets, ImageNet [15] and Flowers-102 [24], the performance improvement on the Flowers-102 dataset was small. This may be because the instances in Flower-102 are more similar to each other than in the case of ImageNet, and thus are less likely to benefit from TTA. Figure 2 shows the relationship between the model architectures and the TTA ambiguity for each dataset [28]. This can be seen as an analogous consideration to our discussion of ambiguity.
- The benefit of TTA varies depending on the model. Complex models have a smaller performance improvement from TTA than simple models. It is expected that the derivation of the generalization bound considering the complexity of the model such as VC-dimension and Rademacher complexity [1, 22] will provide theoretical support for this experiment.
- The effect of TTA is larger in the case of the small amount of data. It is expected to be theorized by deriving inequalities depending on the sample size.

References

1. Alzubi, J., Nayyar, A., Kumar, A.: Machine learning from theory to algorithms: an overview. In: Journal of physics: conference series. vol. 1142, p. 012012. IOP Publishing (2018)
2. Amiri, M., Brooks, R., Behboodi, B., Rivaz, H.: Two-stage ultrasound image segmentation using u-net and test time augmentation. International journal of computer assisted radiology and surgery **15**(6), 981–988 (2020)
3. Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K.W., Schindelin, J., Cardona, A., Sebastian Seung, H.: Trainable weka segmentation: a machine learning tool for microscopy pixel classification. Bioinformatics **33**(15), 2424–2426 (2017)
4. Dietterich, T.G., et al.: Ensemble learning. The handbook of brain theory and neural networks **2**, 110–125 (2002)
5. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Frontiers of Computer Science **14**(2), 241–258 (2020)
6. Fersini, E., Messina, E., Pozzi, F.A.: Sentiment analysis: Bayesian ensemble learning. Decision support systems **68**, 26–38 (2014)
7. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing **321**, 321–331 (2018)
8. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 289–293. IEEE (2018)
9. Hataya, R., Zdenek, J., Yoshizoe, K., Nakayama, H.: Faster AutoAugment: Learning augmentation strategies using backpropagation. In: Computer Vision – ECCV 2020. pp. 1–16. Springer International Publishing (2020)
10. Hawkins, D.M.: The problem of overfitting. Journal of chemical information and computer sciences **44**(1), 1–12 (2004)
11. Indurkhy, N., Damerau, F.J.: Handbook of natural language processing, vol. 2. CRC Press (2010)
12. Kim, I., Kim, Y., Kim, S.: Learning loss for test-time augmentation. arXiv preprint arXiv:2010.11422 (2020)
13. Kimura, M.: Why mixup improves the model performance (Jun 2020)
14. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering **160**(1), 3–24 (2007)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
16. Krogh, A., Vedelsby, J.: Validation, and active learning. Advances in neural information processing systems **7** 7, 231 (1995)
17. Li, Y., Hu, G., Wang, Y., Hospedales, T., Robertson, N.M., Yang, Y.: Differentiable automatic data augmentation. In: Computer Vision – ECCV 2020. pp. 580–595. Springer International Publishing (2020)
18. Lim, S., Kim, I., Kim, T., Kim, C., Kim, S.: Fast AutoAugment (May 2019)
19. Lyzhov, A., Molchanova, Y., Ashukha, A., Molchanov, D., Vetrov, D.: Greedy policy search: A simple baseline for learnable test-time augmentation. In: Peters, J., Sontag, D. (eds.) Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI). Proceedings of Machine Learning Research, vol. 124,

- pp. 1308–1317. PMLR (03–06 Aug 2020), <http://proceedings.mlr.press/v124/lyzhov20a.html>
20. Mikolajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: 2018 international interdisciplinary PhD workshop (IIPhDW). pp. 117–122. IEEE (2018)
 21. Mocerino, L., Rizzo, R.G., Peluso, V., Calimera, A., Macii, E.: Adaptive test-time augmentation for low-power CPU. CoRR **abs/2105.06183** (2021), <https://arxiv.org/abs/2105.06183>
 22. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2018)
 23. Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., Horvath, P.: Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific reports* **10**(1), 1–7 (2020)
 24. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008)
 25. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: SpecAugment: A simple data augmentation method for automatic speech recognition (Apr 2019)
 26. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621 (2017)
 27. Polikar, R.: Ensemble learning. In: Ensemble machine learning, pp. 1–34. Springer (2012)
 28. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: When and why test-time augmentation works. arXiv preprint arXiv:2011.11156 (2020)
 29. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *Journal of Big Data* **6**(1), 1–48 (2019)
 30. Tian, K., Lin, C., Sun, M., Zhou, L., Yan, J., Ouyang, W.: Improving Auto-Augment via Augmentation-Wise weight sharing (Sep 2020)
 31. Van Dyk, D.A., Meng, X.L.: The art of data augmentation. *Journal of Computational and Graphical Statistics* **10**(1), 1–50 (2001)
 32. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)
 33. Vapnik, V.N.: An overview of statistical learning theory. *IEEE transactions on neural networks* **10**(5), 988–999 (1999)
 34. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (2019)
 35. Wang, G., Li, W., Ourselin, S., Vercauteren, T.: Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In: International MICCAI Brainlesion Workshop. pp. 61–72. Springer (2018)
 36. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
 37. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. *AAAI* **34**(07), 13001–13008 (Apr 2020)