Verification-Aided Deep Ensemble Selection

Guy Amir, Tom Zelazny, Guy Katz and Michael Schapira
The Hebrew University of Jerusalem, Jerusalem, Israel
{guyam, tomz, guykatz, schapiram}@cs.huji.ac.il

Abstract—Deep neural networks (DNNs) have become the technology of choice for realizing a variety of complex tasks. However, as highlighted by many recent studies, even an imperceptible perturbation to a correctly classified input can lead to misclassification by a DNN. This renders DNNs vulnerable to strategic input manipulations by attackers, and also oversensitive to environmental noise. To mitigate this phenomenon, practitioners apply joint classification by an ensemble of DNNs. By aggregating the classification outputs of different individual DNNs for the same input, ensemble-based classification reduces the risk of misclassifications due to the specific realization of the stochastic training process of any single DNN. However, the effectiveness of a DNN ensemble is highly dependent on its members not simultaneously erring on many different inputs. In this case study, we harness recent advances in DNN verification to devise a methodology for identifying ensemble compositions that are less prone to simultaneous errors, even when the input is adversarially perturbed — resulting in more robustly-accurate ensemble-based classification. Our proposed framework uses a DNN verifier as a backend, and includes heuristics that help reduce the high complexity of directly verifying ensembles. More broadly, our work puts forth a novel universal objective for formal verification that can potentially improve the robustness of real-world, deep-learning-based systems across a variety of application domains.

I. INTRODUCTION

In recent years, deep learning [33] has emerged as the state-of-the-art solution for a myriad of tasks. Through the automated training of deep neural networks (DNNs), engineers can create systems capable of correctly handling previously unencountered inputs. DNNs excel at tasks ranging from image recognition and natural language processing to game playing and protein folding [2], [21], [38], [48], [74], [75], and are expected to play a key role in various complex systems [15], [44].

Despite their immense success, DNNs suffer from severe vulnerabilities and weaknesses. A prominent example is the sensitivity of DNNs to adversarial inputs [34], [49], [80], i.e., slight perturbations of correctly-classified inputs that result in misclassifications. The susceptibility of DNNs to input perturbations involves two risks that limit the applicability of deep learning to mission-critical tasks: (1) falling victim to strategic input manipulations by attackers, and (2) failing to generalize well in the presence of environmental noise. In light of the above, recent work has focused on enhancing the robustness of DNN-based classification to adversarial inputs while preserving accuracy [13], [29], [62], [82], [97]. Informally, a classifier is robustly accurate (aka astute [86]) with respect to a given distribution over inputs, if it continues to correctly classify inputs drawn from this distribution, with high probability, even when these inputs are arbitrarily perturbed (up to some maximally allowed perturbation).

We focus here on a classic technique for improving classification quality [9], [52]: combining the outputs of an ensemble [28], [37], [81] of DNN-based classifiers on an input to derive a joint classification decision for that input. By incorporating the outputs of independently-trained DNNs, ensembles mitigate the risk of misclassification of a single DNN due to a specific realization of its stochastic training process and the specifics of its training data traversal. For a DNN ensemble to provide a meaningful improvement over utilizing a single DNN, its members should not frequently misclassify the same input. Consider, for instance, an extreme example, where an ensemble with $k = 10$ members is used, but for some part of the input space, the 10 DNNs effectively behave identically, making mistakes on the exact same inputs. In this scenario, the ensemble as a whole is no more robust on this input subspace than each of its individual members. Our objective is to demonstrate how recent advances in DNN verification [40], [45] can be harnessed to provide system designers and engineers with the means to avoid such scenarios, by constructing adequately diverse ensembles.

Significant progress has recently been made on formal verification techniques for DNNs [1], [8], [11], [12], [26], [56], [67], [76], [90]. The basic DNN verification query is to determine, given a DNN $N$, a precondition $P$, and a postcondition $Q$, whether there exists an input $x$ such that $P(x)$ and $Q(N(x))$ both hold. Recent verification work has focused on identifying adversarial inputs to DNN-based classification, or formally proving that no such inputs exist [30], [35], [58]. We demonstrate the applicability of DNN verification to solving a new kind of queries, pertaining to DNN ensembles, which could significantly boost the robustness of these ensembles (as opposed to just measuring the robustness of individual DNNs). We note that despite great strides in recent years [47], [58], [76], even state-of-the-art DNN verification tools face severe scalability limitations. This renders solving verification queries pertaining to ensembles extremely challenging, since the complexity of this task grows exponentially with the number of ensemble members (see Section III).

In this case-study paper, we propose and evaluate an efficient and scalable approach for verifying that different ensemble members do not tend to err simultaneously. Specifically, our scheme considers small subsets of ensemble members,1

1While our technique is applicable to subsets of any size, we focused on pairs in our evaluation, as we later elaborate.
and dispatches verification queries to seek perturbations of inputs for which all members in the subset err simultaneously. By identifying such inputs, we can assign a mutual error score to each subset. Using these mutual error scores, we compute, for each individual ensemble member, a uniqueness score that signifies how often it err simultaneously with other ensemble members. This score can be used to detect the “weakest” ensemble members, i.e., those most prone to erring in parallel to others, and replace them with fresh DNNs — thus enhancing the diversity among the ensemble members, and improving the overall robust accuracy of the ensemble.

To evaluate our scheme, we implemented it as a proof-of-concept tool, and used this tool to conduct extensive experimentation on DNN ensembles for classifying digits and clothing items. Our results demonstrate that by identifying the weakest ensemble members (using verification) and replacing them, the robust accuracy of the ensemble as a whole may be significantly improved. Additional experiments that we conducted also demonstrate that our verification-driven approach affords significant advantages when compared to competing, non-verification-based, methods. Together, these results showcase the potential of our approach. Our code and benchmarks are publicly available online [6].

The rest of the paper is organized as follows. Section II contains background on DNN ensembles and DNN verification. In Section III we present our verification-based methodology for ensemble selection, and then present our case study in Section IV. Next, in Section V we compare our verification-based approach to state-of-the-art, gradient-based, methods. Related work is covered in Section VI, and we conclude and discuss future work in Section VII.

II. BACKGROUND

Deep Neural Networks. A deep neural network (DNN) [33] is a directed graph, comprised of layers of nodes (also known as neurons). In feed-forward DNNs, data flows sequentially from the first (input) layer, through a sequence of intermediate (hidden) layers, and finally into an output layer. The network’s output is evaluated by assigning values to the input layer’s neurons and computing the value assignment for neurons in each of the following layers, in order, until reaching the output layer and returning its neuron values to the user. In classification networks, which are our subject matter here, each output neuron corresponds to an output class; and the output neuron with the highest value represents the class, or label, which the particular input is being classified as.

Fig. 1 depicts a toy DNN. It has an input layer with two neurons, followed by a weighted sum layer, which computes an affine transformation of values from its preceding layer. For example, for input \( v_1 = [1, -5]^T \), the second layer’s computed values are \( v_2 = [-8, 1]^T \). Next is a ReLU layer, which applies the ReLU function \( \text{ReLU}(x) = \max(0, x) \) to each individual neuron, resulting in \( v_3 = [0, 1]^T \). Finally, the network’s output layer again computes an affine transformation, resulting in the output \( v_4 = [6, 3]^T \). Thus, input \( [1, -5]^T \) is classified as the label corresponding to neuron \( v_4^1 \). For additional details, see [33].

Accuracy, Robustness, and Deep Ensembles. The weights of a DNN are determined through its training process. In supervised learning, we are provided a set of pairs \((x_i, l_i)\) drawn according to some (unknown) distribution \( D \), where \( x_i \) is an input point and \( l_i \) is a ground-truth label for that input. The goal is to select weights for the DNN \( N \) that maximize its accuracy, which is defined as: \( Pr_{(x, l) \sim D}(N(x) = l) \) (we slightly abuse notation, and use \( N(x) \) to denote both the network’s output vector, as well as the label it assigns \( x \)).

We restrict our attention to the classification setting, in which labels are discrete. The training of a DNN-based classifier is typically a stochastic process. This process is affected, for example, by the initial assignment of weights to the DNN, the order in which training data is traversed, and more. A prominent method for avoiding misclassifications originating from the stochastic training of a single DNN is employing deep ensembles. A deep ensemble is a set \( E = \{N_1, \ldots, N_k\} \) of \( k \) independently-trained DNNs. The ensemble classifies an input by aggregating the individual classification outputs of its members (see Fig. 2). The collective decision is typically achieved by averaging over all members’ outputs. Ensembles have been shown to often achieve better accuracy than their individual members [9], [52], [57], [92].

A critical condition for the success of ensemble-based classifiers is that the ensemble members’ misclassifications are not strongly correlated [53], [63], [79]. This key property is crucial in order to avoid a scenario where many different members of the ensemble frequently make mistakes on the same input, causing the ensemble as a whole to also err on that input. Heuristics for achieving diversity across ensemble members include, e.g., training the members simultaneously with diversity-aware loss [43], [52], randomly initializing different weights for the ensemble members [50], and other methods [63], [73].

Since the discovery of adversarial inputs, practitioners have become interested in DNNs that are not only accurate but also robustly accurate. We say that a network \( N \) is \( \epsilon \)-robust around the point \( x \) if every input point that is at most \( \epsilon \) away from \( x \) receives the same classification as \( x \): \( \|x' - x\| \leq \epsilon \Rightarrow N(x) = N(x') \), where \( N(x) \) is the label assigned to \( x \); and the definition of accuracy is generalized to \( \epsilon \)-robust
accuracy as follows: $\Pr_{(x,l) \sim D}(\|x' - x\| \leq \epsilon \Rightarrow N(x') = l)$. While improvements in accuracy afforded by ensembles are straightforward to measure, this is typically not the case for robust accuracy, as we discuss in Section III.

**DNN Verification.** Given a DNN $N$, a verification query on $N$ specifies a precondition $P$ on $N$’s input vector $x$, and a postcondition $Q$ on $N$’s output vector $N(x)$. A DNN verifier needs to determine whether there exists a concrete input $x_0$ that satisfies $P(x_0) \land Q(N(x_0))$ (the SAT case), or not (the UNSAT case). Typically, $P$ and $Q$ are expressed in the logic of linear real arithmetic. For instance, the $\epsilon$-robustness of a DNN around a point $x$ can be phrased as a DNN verification problem that is known to be NP-complete [46].

### III. IMPROVING ROBUST ACCURACY USING VERIFICATION

#### A. Directly Quantifying Robust Accuracy is Hard

In order to construct a robustly-accurate ensemble $E$ with $k$ members, we train a set of $n > k$ DNNs that provides high robust accuracy. This method of training multiple models and then discarding a subset thereof is known as ensemble pruning, and is a common practice in deep-ensemble training [14], [98]. In our case, a straightforward approach to do so would be to quantify the robust accuracy for all possible $k$-sized DNN-subsets, and then pick the best one. This, however, is computationally expensive, and requires an accurate estimate of the robust accuracy of an ensemble.

A natural approach for estimating the $\epsilon$-robust accuracy of a DNN is to verify, for many points in the test data, that the DNN yields an accurate label not only on each data point itself, but also on each and every input derived from that data point via an $\epsilon$-perturbation [30]. The fraction of tested points for which this is indeed the case can be used to estimate the accuracy of the classifier on the underlying distribution from which the data is generated.

A similar process can be performed for an ensemble $E = \{N_1, \ldots, N_k\}$, by first constructing a single, large DNN $N_E$ that aggregates $E$’s joint classification, and then verifying its robustness on a set of points from the test data (see the extended version of this paper [7]). However, this approach faces a significant scalability barrier: the DNN ensemble, $N_E$, comprised of all $k$ member-DNNs is (roughly) $k$ times larger than any of the $N_i$’s, and since DNN verification becomes exponentially harder as the DNN size increases, $N_E$’s size might render efficient verification infeasible. As we demonstrate later, this is the case even when the constituent networks themselves are fairly small. Our proposed methodology circumvents this difficulty by only solving verification queries pertaining to very small sets of DNNs.

#### B. Mutual Error Scores and Uniqueness Scores

In general, the less likely it is that members of an ensemble err simultaneously with other members, the more accurate the ensemble is. This motivates our definition of mutual error scores below.

**Definition 1 (Agreement Points):** Given an ensemble $E = \{N_1, N_2, \ldots, N_k\}$, we say that an input point $x_0$ is an agreement point for $E$ if there is some label $y_0$ such that $N_i(x_0) = y_0$ for all $i \in [k]$. We let $E(x_0)$ denote the label $y_0$.

As we later discuss, the $\epsilon$-neighborhoods of agreement points are natural locations for detecting hidden tendencies of ensemble members to err together.

**Definition 2 (Mutual Errors):** Let $E$ be an ensemble, and let $x_0$ be an agreement point for $E$. Let $B_{x_0, \epsilon}$ be the $\epsilon$-ball around $x_0$, $B_{x_0, \epsilon} = \{x \mid \|x - x_0\|_\infty \leq \epsilon\}$. We say that $N_1$ and $N_2$ have a mutual error in $B$ if there exists a point $x \in B_{x_0, \epsilon}$ such that $N_1(x) \neq E(x_0)$ and $N_2(x) \neq E(x_0)$.

Intuitively, if $N_1$ and $N_2$ have many mutual errors, incorporating both into an ensemble is a poor choice. This naturally gives rise to the following definition:

**Definition 3 (Mutual Error Scores):** Let $A$ be a finite set of $m$ agreement points in an ensemble $E$’s input space, and let $B_1, B_2, \ldots, B_m$ denote the $\epsilon$-balls surrounding the points in $A$. Let $N_1$, $N_2$ denote two members of $E$. The mutual error score of $N_1$ and $N_2$ with respect to $E$ and $A$ is denoted by $ME_{E,A}(N_1, N_2)$, and defined as:

$$ME_{E,A}(N_1, N_2) = \frac{\{i \mid N_1 \text{ and } N_2 \text{ have a mutual error in } B_i\}}{m}$$

Observe that $ME_{E,A}(N_1, N_2)$ is always in the range $[0,1]$. The closer it is to 1, the more mutual errors $N_1$ and $N_2$ have, making it unwise to place them in the same ensemble.
Definition 4 (Uniqueness Scores): Given an ensemble \( \mathcal{E} = \{N_1, N_2, \ldots, N_n\} \) and a set \( A \) of agreement points for \( \mathcal{E} \), we define, for each ensemble member \( N_i \), the uniqueness score for \( N_i \) with respect to \( \mathcal{E} \) and \( A \), \( \text{US}_{\mathcal{E},A}(N_i) \), as:

\[
\text{US}_{\mathcal{E},A}(N_i) = 1 - \frac{\sum_{j \neq i} \text{ME}_{\mathcal{E},A}(N_i, N_j)}{n - 1}
\]

The uniqueness score (US) of \( N_i \) is the complement of its average mutual error score with the other ensemble members. When this score is close to 0, \( N_i \) tends to err simultaneously with other members of the ensemble on points in \( A \). In contrast, the closer the uniqueness score is to 1, the rarer it is for \( N_i \) to misclassify the same inputs as other members of the ensemble. Hence, ensemble members with low uniqueness scores are, intuitively, good candidates for replacement.

We point out that our definitions above can naturally be generalized to larger subsets of the ensemble members — thus measuring robust accuracy more precisely, but rendering these measurements more complex to perform in practice.

Computing Mutual Errors. The only computationally complex step in determining the uniqueness scores of individual ensemble members is computing the pairwise mutual errors for the ensemble. To this end, we leverage DNN verification technology. Specifically, given two ensemble members \( N_1 \) and \( N_2 \), an agreement point \( a \) for the ensemble with label \( l \), and \( \epsilon > 0 \), an appropriate DNN verification query can be formulated as follows. First, we construct from \( N_1 \) and \( N_2 \) a single, larger DNN \( N \), which captures \( N_1 \) and \( N_2 \) simultaneously processing a shared input vector, side-by-side. This network \( N \) is then passed to a DNN verifier, with the precondition that the input be restricted to \( B \), an \( \epsilon \)-ball around \( a \), and the postcondition that (1) among \( N \)'s output neurons that correspond to the outputs of \( N_1 \), the neuron representing \( l \) be not maximal, and (2) among \( N \)'s output neurons that correspond to the outputs of \( N_2 \), the neuron representing \( l \) be not maximal. Such queries are supported by most available DNN verification engines. We note that this encoding (depicted in Figure 3), where two networks and their output constraints are combined into a single query, is crucial for finding inputs on which both DNNs err simultaneously. For additional details, see the extended version of this paper [7].

C. Ensemble Selection using Uniqueness Scores

An Iterative Scheme. Building on our verification-based method for computing mutual error scores, we propose an iterative scheme for constructing an ensemble. Our scheme consists of the following steps:

1) independently train a set \( \mathcal{N} \) of \( n \) DNNs, and identify a set \( A \) of \( m \) agreement points that are correctly classified by all \( n \) DNNs.\(^2\) This is done by sequentially checking points from the validation dataset;

2) arbitrarily choose an initial candidate ensemble \( \mathcal{E} \) of size \( k < n \);

3) compute (using a verification engine backend) all mutual error scores for the DNN members comprising \( \mathcal{E} \), with respect to \( A \);

4) compute the uniqueness score for each ensemble member, and identify a DNN member \( N_i \) with a low score;

5) identify a fresh DNN \( N_f \), not currently in \( \mathcal{E} \), that has a higher uniqueness score than \( N_i \), if one exists, and replace \( N_i \) with \( N_f \). Specifically, identify a DNN \( N_f \in \mathcal{N} \setminus \mathcal{E} \) such that the uniqueness score of \( N_f \) with respect to the ensemble \( \mathcal{E} \setminus \{N_i\} \cup \{N_f\} \) and the point set \( A \), namely \( \text{US}_{\mathcal{E}\setminus\{N_i\}\cup\{N_f\},A}(N_f) \), is maximal. If this score is greater than \( \text{US}_{\mathcal{E},A}(N_i) \), replace \( N_i \) with \( N_f \), i.e. set \( \mathcal{E} := \mathcal{E} \setminus \{N_i\} \cup \{N_f\} \); and

6) repeat Steps (3) through (5), until no \( N_f \) is found or until the user-provided timeout or maximal iteration count are exceeded.

Intuitively, after starting with an arbitrary ensemble, we run multiple iterations, each time trying to improve the ensemble. Specifically, we identify the “weakest” member of the current ensemble, and replace it with a fresh DNN that obtains a higher uniqueness score relevant to the remaining members — thus ensuring that each change that we make improves the overall robust accuracy on the fixed set of agreement points.

The greedy search procedure is repeated for the new candidate ensemble, and so on. The process terminates after a predefined number of iterations is reached, when the process converges (no further improvement is achievable on the fixed set of agreement points), or when a predefined timeout value is exceeded.

On the Importance of Agreement Points. Our iterative scheme for constructing an ensemble starts with an arbitrary selection of \( k \) candidate members, and then computes the uniqueness score for each member. As mentioned, the uniqueness scores are computed with respect to a fixed set of agreement points, pre-selected from the validation data (which is labeled data, not used for training the DNNs).

We point out that agreement points are data points on which there is overwhelming consensus among ensemble members, despite the specific realization of the training process of each member. As such, agreement points correspond to data points that are “easy” to label correctly. Consequently, data points in close proximity of an agreement point are rarely classified differently than the agreement point by an individual ensemble member, let alone by multiple members simultaneously. As our objective is to expose implicit tendencies of ensemble members to err together, the close neighborhood of agreement points is a natural area for seeking joint deviations from the consensual label (which are expected to be extremely rare). In our evaluation, we computed uniqueness scores based solely on correctly-classified agreement points and ignored any incorrectly-classified agreement points.\(^3\)

As we later demonstrate, a small set of correctly-classified agreement points from the validation set can be used, in

\(^2\)In our experiments, we arbitrarily chose \( k = 5 \), \( n = 10 \) and \( m = 200 \).

\(^3\)For example, in our MNIST experiments 99.7% of the agreement points were correctly classified by all individual DNNs, and by the ensemble as a whole.
practice, to identify ensemble members that tend to err simultaneously on other data points. We note that this is also the case even when the chosen agreement points are all identically labeled.

**Monotonicity and Convergence.** Using our approach, an ensemble member is replaced with a fresh DNN only if this replacement leads to strictly fewer joint errors with the remaining members on the fixed set of agreement points. Thus, the total number of joint errors decreases with every replacement; and, as this number is trivially lower-bounded by 0, this (“potential-function” style) argument establishes the process’s monotonicity and convergence.

By iteratively reducing the number of joint errors across all pairs of chosen ensemble members, our iterative process improves the robust accuracy of the resulting ensemble on the fixed set of agreement points. This, however, does not guarantee improved robust accuracy over the entire input domain. Nonetheless, we show in Section IV that such an improvement does typically occur in practice, even on randomly sampled subsets of input points (which are not necessarily agreement points).

**IV. CASE STUDY: MNIST AND FASHION-MNIST**

Below, we present the evaluation of our methodology on two datasets: the MNIST dataset for handwritten digit recognition [51], and the Fashion-MNIST dataset for clothing classification [91]. Our results for both datasets demonstrate that our technique facilitates choosing ensembles that provide high robust accuracy via relatively few, efficient verification queries. The considered datasets are conducive for our purposes since they allow attaining high accuracy using fairly small DNNs, which enables us to directly quantify the robust accuracy of an entire ensemble, by dispatching verification queries that would otherwise be intractable (see Section III-A). This provides the ground truth required for assessing the benefits of our approach. The scalability afforded by our approach is crucial even for handling the relatively modest-sized DNNs considered: on the MNIST data, for instance, mutual-error verification queries for two ensemble members typically took a few seconds, whereas verification queries involving the full ensemble of five networks often timed out (35% of the queries on the MNIST data timed out after 24 hours, versus only roughly 1% of the pairwise mutual-error queries). As constituent DNN sizes and ensemble sizes increase, this gap in scalability is expected to become even more significant.

Our verification queries were dispatched using the Marabou verification engine [47] (although other engines could also be used). Additional details regarding the encoding of the verification queries, as well as detailed experimental results, appear in the extended version of this paper [7]. We have publicly released our code, as well as all benchmarks and experimental data, within an artifact accompanying this paper [6].

**MNIST.** For this part of our evaluation, we trained 10 independent DNNs \(\{N_1, \ldots, N_{10}\}\) over the MNIST dataset [51], which includes 28×28 grayscale images of 10 handwritten digits (from “0” to “9”). Each of these networks had the same architecture: an input layer of 784 neurons, followed by a fully-connected layer with 30 neurons, a ReLU layer, another fully-connected layer with 10 neurons, and a final softmax layer with 10 output neurons, corresponding to the 10 possible digit labels. All networks achieved high accuracy rates of 96.29% – 96.57% (see Table I).

After training, we arbitrarily constructed two distinct ensembles with five DNN members each: \(E_1 = \{N_1, \ldots, N_5\}\) and \(E_2 = \{N_6, \ldots, N_{10}\}\), with an accuracy of 97.8% and 97.3%, respectively. Notice that the ensembles achieve a higher accuracy over the test set than their individual members.

We then applied our method in an attempt to improve the robust accuracy of \(E_1\). We began by searching the validation set, and identifying 200 agreement points (the set \(A\)), all correctly labeled as “0” by all 10 networks. Using the 200 agreement points and 6 different perturbation sizes \(\epsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06\}\), we constructed 1200 \(\epsilon\)-balls around the selected agreement points; and then, for every ball \(B\) and for every pair \(N_1, N_2 \in E_1\), we encoded a verification query to check whether \(N_1\) and \(N_2\) have a mutual error in \(B\) (see example in Fig. 3). This resulted in \(\binom{200}{2} \cdot 6 = 12000\) verification queries, which we dispatched using the Marabou DNN verifier [47] (each query ran with a 2-hour timeout limit). Finally, we used the results to compute the uniqueness score for each network in \(E_1\); these results, which appear briefly in Table I (for \(\epsilon = 0.02\) and appear in full in [7], clearly show that two of the members, \(N_2\) and \(N_5\), are each relatively prone to erring simultaneously with the remaining four members of \(E_1\).

Next, we began searching among the remaining networks, \(N_6, \ldots, N_{10}\), for good replacements for \(N_2\) and \(N_5\). Specifically, we searched for networks that obtained higher US scores than \(N_2\) and \(N_5\). To achieve this, we began modifying \(E_1\), each time removing either \(N_2\) or \(N_5\), replacing them with one of the remaining networks, and computing the uniqueness scores for the new members (with respect to the four remaining original networks). We observed that for both \(N_2\) and \(N_5\), network \(N_9\) was a good replacement, obtaining very high US values. For additional details, see the extended version of our paper [7].

Finally, to evaluate the effect of our changes to \(E_1\), we constructed the two new ensembles, \(E_1^{2\rightarrow9} = \{N_1, N_9, N_3, N_4, N_5\}\) and \(E_1^{9\rightarrow5} = \{N_1, N_2, N_3, N_4, N_9\}\). Computing the new ensembles’ robust accuracy over the entire

\[ ^4 \text{Although the DNNs all have the same size and architecture, common ensemble training processes randomly initialize their weights, and also randomly pick samples from the same training set (see [50]). This is the cause for diversity among ensemble members, which our algorithm later detects.} \]

\[ ^5 \text{In our experiments, we empirically selected 200 agreement points in order to balance between precision (a higher number of points) and verification speed (a smaller number of points). This selection is based on a user’s available computing power.} \]

\[ ^6 \text{The “0” label is the label with the highest accuracy among the trained ensemble members, and thus “0”-labeled agreement points represent areas in the input space with extremely high consensus.} \]

\[ ^7 \text{\(\epsilon\) values which are too small, or too large, render the queries trivial. Thus, we found it to be useful to use a varied selection of \(\epsilon\) values.} \]
TABLE I: Accuracy and uniqueness scores for the MNIST networks. Uniqueness scores are measured with respect to the ensemble (either \(E_1\) or \(E_2\)).

<table>
<thead>
<tr>
<th></th>
<th>(N_1)</th>
<th>(N_2)</th>
<th>(N_3)</th>
<th>(N_4)</th>
<th>(N_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>96.42%</td>
<td>96.55%</td>
<td>96.40%</td>
<td>96.46%</td>
<td>96.29%</td>
</tr>
<tr>
<td><strong>US</strong></td>
<td>90.75%</td>
<td><strong>88.38%</strong></td>
<td>90.63%</td>
<td>92.13%</td>
<td><strong>88.63%</strong></td>
</tr>
</tbody>
</table>

\[ E^{10\rightarrow^4} = \{N_6, N_7, N_8, N_9, N_4\}, \]

and \(E^{20\rightarrow^4} = \{N_6, N_7, N_8, N_9, N_4\}\), and compared their robust accuracy to that of \(E_2\) on 200 random points from the test set. The results, depicted in Fig. 4, indicate that the ensemble’s robust accuracy decreased significantly, as expected.

In both aforementioned experiments, we also computed the accuracy (as opposed to robust accuracy) of the new ensembles, by evaluating them over the test set. All new ensembles had an accuracy that was on par with that of the original ensembles — specifically, within a range of \(\pm 0.2\%\) from the original ensembles’ accuracy.

**Fashion-MNIST.** For the second part of our evaluation, we trained 10 independent DNNs \(\{N_{11}, \ldots, N_{20}\}\) over the Fashion-MNIST dataset [91], which includes \(28 \times 28\) grayscale images of 10 clothing categories (“Coat”, “Dress”, etc.), and is considered more complex than the MNIST dataset. Each DNN had the same architecture as the MNIST-trained DNNs, and achieved an accuracy of 87.05%–87.53% (see Table II). We arbitrarily constructed two distinct ensembles, \(E_3 = \{N_{11}, \ldots, N_{15}\}\) and \(E_4 = \{N_{16}, \ldots, N_{20}\}\), with an accuracy of 88.22% and 88.48%, respectively.

Next, we again computed the US values of each of the networks. The results, which appear in full in [7], indicate a high variance among the uniqueness scores of the members of \(E_4\), as compared to the relatively similar scores of \(E_3\’s\) members. We thus chose to focus on \(E_4\). Based on the computed US values, we identified \(N_{20}\) as its least unique DNN; and, by replacing \(N_{20}\) with each of the five networks not currently in \(E_4\), identified that \(N_{15}\) is a good candidate for replacing \(N_{20}\). Performing our validation step over \(E_4^{20\rightarrow15}\) revealed that its robust accuracy has indeed increased. Running the “reverse” experiment, in which \(E_4\’s\) most unique member is replaced with a worse candidate, led us to consider the ensemble \(E_4^{18\rightarrow13}\), which indeed demonstrated lower robust accuracy than the original ensemble. For additional details, see the extended version of our paper [7].

For the final step of our experiment, we used our approach to iteratively switch two members of an ensemble. Specifically, after creating \(E_4^{20\rightarrow15}\), which had higher robust accuracy than \(E_4\), we re-computed the US scores of its members, and identified again the least unique member — in this case, \(N_{16}\). Per our computation, the best candidate for replacing it was \(N_{12}\). The resulting ensemble, namely \(E_4^{20\rightarrow15,16\rightarrow12}\), indeed demonstrated higher robust accuracy than both its predecessors. Performing another iteration of the “reverse” experiment yielded ensemble \(E_4^{18\rightarrow13,17\rightarrow11}\), with poorer robust accuracy. The results appear in Fig. 5. We note that the only discrepancy, namely the robust accuracy of \(E_4^{20\rightarrow15}\) being lower than that...
TABLE II: Accuracy and uniqueness scores for the Fashion-MNIST networks. Uniqueness scores are measured with respect to the ensemble (either $E_3$ or $E_4$).

<table>
<thead>
<tr>
<th></th>
<th>$N_{11}$</th>
<th>$N_{12}$</th>
<th>$E_3$</th>
<th>$N_{13}$</th>
<th>$N_{14}$</th>
<th>$N_{15}$</th>
<th>$N_{16}$</th>
<th>$N_{17}$</th>
<th>$N_{18}$</th>
<th>$N_{19}$</th>
<th>$N_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>87.14%</td>
<td>87.13%</td>
<td>87.53%</td>
<td>87.34%</td>
<td>87.3%</td>
<td>87.32%</td>
<td>87.35%</td>
<td>87.34%</td>
<td>87.11%</td>
<td>67.38%</td>
<td>72.38%</td>
</tr>
<tr>
<td>US</td>
<td>70.63%</td>
<td>71.5%</td>
<td>69.75%</td>
<td>70.88%</td>
<td>73.25%</td>
<td>67.38%</td>
<td>72.38%</td>
<td>80.13%</td>
<td>71.38%</td>
<td>66.75%</td>
<td>71.38%</td>
</tr>
</tbody>
</table>

Next, we used our three attacks to search for mutual errors of DNN pairs — i.e., adversarial examples that simultaneously affect a pair of DNNs. Specifically, we applied the attacks on both datasets (MNSIT and Fashion-MNIST), and searched for adversarial examples within various $\epsilon$-balls around the same set of agreement points used in our previous experiments. This allowed us to subsequently compute, via gradient attacks, the mutual error scores of DNN pairs, and consequently, the uniqueness scores of each constituent ensemble member. The results of the total number of adversarial inputs found (SAT queries) are summarized in Table III. Each gradient attack typically took a few seconds to run. We also provide further details regarding the uniqueness scores computed by the three gradient-based methods in the extended version of this paper [7], and in our accompanying artifact [6].

The results in Table III include a total of 108000 experiments, on all ensemble pairs. In these experiments, our

\[^{8}\text{The 108000 experiments consist of } 30 \text{ pairs, times 200 agreement points, times 6 perturbation sizes, times 2 datasets.}\]

Fig. 4: The average robust accuracy scores for our original and modified ensembles. The results for $\epsilon = 0.01$ and $\epsilon = 0.06$ are trivial (the ensembles achieve near-perfect or near-zero robustness), and are omitted to reduce clutter.

Fig. 5: The original ensemble $E_4$ (center), ensembles modified to gain robust accuracy (right), and ensembles modified to reduce robust accuracy (left).

Similarly to the MNIST case, the new ensembles in the Fashion-MNIST experiments obtained an accuracy that was on par with that of the original ensembles — specifically, within a range of ±0.17% from the original ensemble’s accuracy.

V. COMPARISON TO GRADIENT-BASED ATTACKS

Current state-of-the-art approaches for assessing a network’s robustness and robust accuracy rely on gradient-based attacks — a popular class of algorithms that, like verification methods, are capable of finding adversarial examples for a given neural network. In this section we compare our verification-based approach to these methods.

Gradient-based attacks generate adversarial examples by optimizing (via various techniques) a loss metric over the network’s output, relative to its input. This allows these methods to effectively search the local surroundings of a fixed input point for local optima, which often constitute adversarial inputs. Gradient-based methods, such as the fast-gradient sign method (FGSM) [39], projected gradient descent (PGD) [60], and others [49], [59], are in widespread use due to their scalability and relative ease of use. However, as we demonstrate here, they are often unsuitable in our setting.

In order to evaluate the effectiveness of gradient-based methods for measuring the robust accuracy of ensembles, we modified the common FGSM [39] and I-FGSM [49] (“Iterative FGSM”) methods, thus extending them into three novel attacks aimed at finding adversarial examples that can fool multiple ensemble members simultaneously. We refer to these attacks as Gradient Attack (G.A.) 1, 2, and 3. For a thorough explanation of these attacks, as well as information about their design and implementation, see the extended version of our paper [7].

Next, we used our three attacks to search for mutual errors of DNN pairs — i.e., adversarial examples that simultaneously affect a pair of DNNs. Specifically, we applied the attacks on both datasets (MNSIT and Fashion-MNIST), and searched for adversarial examples within various $\epsilon$-balls around the same set of agreement points used in our previous experiments. This allowed us to subsequently compute, via gradient attacks, the mutual error scores of DNN pairs, and consequently, the uniqueness scores of each constituent ensemble member. The results of the total number of adversarial inputs found (SAT queries) are summarized in Table III. Each gradient attack typically took a few seconds to run. We also provide further details regarding the uniqueness scores computed by the three gradient-based methods in the extended version of this paper [7], and in our accompanying artifact [6].
TABLE III: The number of SAT queries discovered when searching for an adversarial attack, using the three gradient attack methods (G.A. 1, 2 and 3), and our verification approach.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>G.A. 1</th>
<th>G.A. 2</th>
<th>G.A. 3</th>
<th>verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>1,333</td>
<td>3,886</td>
<td>5,574</td>
<td>16,826</td>
</tr>
<tr>
<td>Fashion-MNIST</td>
<td>17,190</td>
<td>21,245</td>
<td>22,129</td>
<td>33,152</td>
</tr>
<tr>
<td>Total</td>
<td>18,523</td>
<td>25,131</td>
<td>27,703</td>
<td>49,978</td>
</tr>
</tbody>
</table>

verification-based approach returned 49978 SAT results, while the strongest gradient-based method (gradient attack number 3) returned only 27703 SAT results — a 44% decrease in the number of counterexamples found. This discrepancy is on par with previous research [89], which indicates that gradient-based methods may err significantly when used for adversarial robustness analysis. This phenomenon manifests strongly in our setting, which involves many small and medium-sized perturbations that gradient-based approaches struggle with [24].

The reduced precision afforded by gradient-based approaches can, in some cases, lead to sub-optimal ensemble selection choices when compared to our verification-based approaches. Specifically, even if a gradient-based approach produces a uniqueness score ranking that coincides with the one produced using verification, the dramatic decrease in the number of SAT queries leads to much smaller mutual error scores, and consequently — to uniqueness score values that are overly optimistic, and less capable of distinguishing between poor and superior robust accuracy results.

For example, when observing the first two arbitrary ensembles on the MNIST dataset, $E_1$ and $E_2$, the three gradient approaches (G.A. 1, 2 and 3) respectively assign average uniqueness scores of $⟨95.4\%, 97.8\%⟩$, $⟨87.5\%, 94.5\%⟩$ and $⟨83.1\%, 92.5\%⟩$ to the two ensembles (when averaging the US over all ensemble members and all perturbations). This indicates that the robust accuracy of the two ensembles is fairly similar (see appendices in [7]). In contrast, when using the more sensitive, verification-based approach, we find a substantially higher number of mutual errors (see Table III), and consequently, detect a much larger gap between the uniqueness scores of the two ensembles: 55% and 77%.

Another example that demonstrates the increased sensitivity of our method, when compared to gradient-based approaches, is obtained by observing the average uniqueness score of $E_3$ and $E_4$ on the Fashion-MNIST dataset. The strongest gradient attack that we used assigned almost identical average uniqueness scores to both ensembles (up to a difference of 0.01%), while our approach was sensitive enough to find a 2% difference between the average US of the two ensembles.

Finally, we note that, unlike verification-based approaches, gradient attacks are incomplete, and are consequently unable to return UNSAT. This makes them less suitable for assessing any additional uniqueness metrics based on robust $\epsilon$-balls. We thus argue that, although gradient-based methods are faster and more scalable than verification, our results showcase the benefits of using verification-based approaches for assessing uniqueness scores and for ensemble selection.

VI. RELATED WORK

Due to its pervasiveness, the phenomenon of adversarial inputs has received a significant amount of attention [27], [34], [61], [65], [66], [80], [99]. More specifically, the machine learning community has put a great deal of effort into measuring and improving the robustness of networks [18]–[20], [29], [36], [54], [60], [68], [71], [72], [87], [94]. The formal methods community has also been looking into the problem, by devising scalable DNN verification, optimization and monitoring techniques [1], [5], [8], [10]–[12], [16], [26], [41], [42], [55], [56], [64], [67], [70], [76], [90], [96]. To the best of our knowledge, ours is the first attempt to apply DNN verification to the setting of DNN ensembles. We note that our approach uses a DNN verifier strictly as a black-box backend, and so its scalability will improve as DNN verifiers become more scalable.

Obtaining DNN specifications to be verified is a difficult problem. While some studies have successfully applied verification to properties formulated by domain-specific experts [3], [4], [22], [25], [45], [78], most research has been focusing on universal properties, which pertain to every DNN-based system; specifically, local adversarial robustness [17], [35], [58], [76], fairness properties [83], network simplification [31] and modification [23], [32], [69], [77], [84], [93], and watermark resilience [32].

VII. CONCLUSION AND FUTURE WORK

In this case-study paper, we demonstrate a novel technique for assessing a deep ensemble’s robust accuracy through the use of DNN verification. To mitigate the difficulty inherent to verifying large ensembles, our approach considers pairs of networks, and computes for each ensemble member a score that indicates its tendency to make the same errors as other ensemble members. These scores allow us to iteratively improve the robust accuracy of the ensemble, by replacing weaker networks with stronger ones. Our empiric evaluation indicates the high practical potential of our approach; and, more broadly, we view this work as a part of the ongoing endeavor for demonstrating the real-world usefulness of DNN verification, by identifying additional, universal, DNN specifications.

Moving forward, we plan to tackle the natural open questions raised by our work: specifically, how our methodology for selecting robustly accurate ensembles can be extended beyond the current greedy search heuristic, as well as how ensembles should be selected in the context of other performance objectives, beyond robust accuracy. We also plan on experimenting with multiple stopping conditions for the ensemble member replacement process; as well as explore potential synergies between our verification-based approach and gradient-based approaches for computing mutual error scores. In addition, we note that we are currently extending
our approach to regression learning ensembles and deep reinforcement learning ensembles. Finally, we are in the process of optimizing our approach by using lighter-weight, incomplete verification tools (e.g., [76], [88], [95]), which afford better scalability, and also support parallelization. This will hopefully allow us to handle significantly larger DNNs and more complex datasets.

Acknowledgements. We thank Haoze Wu for his contribution to this project. The first three authors were partially supported by the Israel Science Foundation (grant number 683/18). The first author was partially supported by the Center for Interdisciplinary Data Science Research at The Hebrew University of Jerusalem. The fourth author was partially supported by funding from Huawei.

References


