ON THE COARSE ROBUSTNESS OF CLASSIFIERS

Ismail R. Alkhouri[1] *Stanley Bak*[2] *Alvaro Velasquez*[3,4] *George K. Atia*[1,5]

[1] Department of Electrical and Computer Engineering, University of Central Florida, Orlando FL, USA
[2] Department of Computer Science, Stony Brook University, Stony Brook NY, USA
[3] Information Innovation Office, Defense Advanced Research Projects Agency, Arlington VA, USA
[4] Department of Computer Science, University of Colorado, Boulder CO, USA
[5] Department of Computer Science, University of Central Florida, Orlando FL, USA

ABSTRACT

Standard measures of robustness, derived from the least amount of adversarial perturbation, often fail to gauge the ability of a classifier to recognize the coarse genres. It is desirable to have a classifier with high coarse robustness with respect to a grouping that is consistent with the class semantics, so that semantically-plausible coarse categories remain invariant to imperceptible perturbations.. In this work, we formalize a new notion of coarse robustness that is defined with respect to a specified grouping of the class labels. We formulate an optimization problem to obtain the optimal grouping, and develop an algorithm that is shown to perform on par with brute force search. Moreover, we propose a training mechanism that incorporates the coarse label information in addition to the finer ones. We empirically and theoretically show that this mechanism improves the proposed coarse notion of robustness while only requiring a relatively small additional parameters and training time.

***Index Terms*—** Coarse robustness, Adversarial attacks, Best label groupings, Course training

1. INTRODUCTION

A common approach to gauge the robustness of a certain classifier to adversarial attacks revolves around determining the least amount of perturbation capable of inducing a misclassification – e.g., see the CLEVER metric [1] and the work in [2, 3, 4, 5, 6, 7, 8] for some of the state-of-the-art attack and robustness evaluation methods.

Given a trained classifier, most previous works have focused on metrics based on the standard notion of robustness that make use of state-of-the-art adversarial attack methods such as [2, 9, 3, 4, 5]. However, since this approach generally ignores the semantic relations among the classes, it falls short of capturing a measure of *severity* of the induced misclassifications, and in turn of the safety and brittleness of the classifier design.For example, misclassifying 'Can' as 'Cup' is less drastic than classifying 'Truck' as 'Cat' (all distinct categories in CIFAR-10 [10]).

In this paper we introduce a new notion of *coarse robustness* which captures the susceptibility of a classifier to perturbations inducing misclassifications of the coarse labels with respect to (w.r.t.) a specified grouping of the class labels. A clear advantage is that the derived measures could be used to gauge the ability of a classifier to recognize the coarse genres w.r.t. the grouping.

In addition to the proposed measures, a training approach that integrates the coarse information is introduced. When compared to conventional training, the introduced coarse training approach is proven to obtain stronger classifiers as leveraged by the introduced measures that evaluate the coarse robustness w.r.t. consistent groupings of the class labels.

Contributions. First, this work introduces a new notion of global coarse robustness of a classifier which captures the hardness of confusing the coarse predictions induced by a certain grouping of the class labels. Second, to find suboptimal groupings, we propose measures of goodness of class groupings and formulate a mathematical program to optimize such measures. An algorithmic procedure is developed based on proposed measures, shown to perform on par with combinatorial brute force search, while requiring a considerably smaller number of iterations. The utility of the introduced measures and the performance of the search algorithm are demonstrated using image classification on known benchmark datasets. Third is the coarse training approach for obtaining improved coarse robustness models. In comparison to Natural Training (NT), we prove that our approach obtains stronger models in terms of the introduced coarse robustness measure. While requiring the same training time and small relative additional parameters, experiments show that our approach produces improved models when compared to models trained using NT.

This work was supported in part by NSF CAREER Award CCF-1552497 and NSF Award CCF-2106339, DOE Award DE-EE0009152, AFRL Contract Number FA8750-20-3-1004, and AFOSR Award 20RICOR012.

2. THE COARSE ROBUSTNESS MEASURE

In this section, we first define the classification model and the conventional way of training a neural network based classifier. Define the classifier $h : \mathbb{R}^N \to [M]$, which maps an observation $x \in \mathbb{R}^N$ to one of M possible (fine) labels, where $[M] := \{1, 2, \ldots, M\}$. The predicted label is obtained as the index of the maximizing discriminant probabilistic functional in vector $f : \mathbb{R}^N \to \Delta^M$ where Δ^M is the probability simplex, with entries $f_m, m \in [M]$, as $h(x) = \text{argmax}_{m\in[M]}\ f_m(x)$.

Definition 1. *Given classifier h, parameterized by θ, Natural Training (NT) on dataset $\mathcal{D}$ with entries in the form $(x, 1_y)$, where $y \in [M]$ is the true label of x and 1_y is one hot encoding vector representation of y, can be defined as the task of minimizing the loss in*

$$\min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{(x,1_y)\in\mathcal{D}} \mathcal{L}(f(x;\theta), 1_y)\,. \tag{1}$$

Let $T : [M] \to [M_c]$ be a grouping function that maps the classifier's output to a coarse label $i \in [M_c]$, where $M_c < M$ is the number of coarse classes. The function T induces the coarse class sets $S_i := \{m \in [M] : T(m) = i\}, i \in [M_c]$.

A standard measure of robustness of classifier h w.r.t. a feature vector x is the least amount of perturbation (relative to some norm) $\eta \in \mathbb{R}^N$, required to produce a false prediction. It is defined as $\eta^s_m(x) := \text{argmin}_\eta\{\|\eta\|_p : h(x+\eta) \neq h(x)\}$, where m is the true label.

Definition 2. *We define a Global Standard Robustness (GSR) measure in vector $g_p \in \mathbb{R}^M$ whose entries represent the average of the l_p distances between examples $x \in \mathbb{R}^N$ and their misclassified perturbed versions $\hat{x} = x + \eta$, i.e.,*

$$g_p(m) := \frac{1}{|\mathcal{D}_m|} \sum_{x\in\mathcal{D}_m} \|\eta^s_m(x)\|_p\,, \tag{2}$$

where $\mathcal{D}_m := \{x \in \mathbb{R}^N : h(x) = m\}$.

Given that $h(x) = m$ with the predicted coarser set $S_{T(m)}$, one could define the minimum perturbation required to induce misclassification of the coarse label (coarse misclassification) as $\eta_m(x; T) := \text{argmin}_\eta\{\|\eta\|_p : h(x + \eta) \notin S_{T(m)}\}$.

Definition 3. *Let vector $c_p(T) \in \mathbb{R}^M$, whose entries*

$$c_p(m; T) := \frac{1}{|\mathcal{D}_m|} \sum_{x\in\mathcal{D}_m} \|\eta_m(x; T)\|_p\,, \tag{3}$$

reflect the average minimum perturbations required to cause coarse misclassification from predicted label m, used to define our global measure of coarse robustness, dubbed GCR.

3. FINDING BEST MAPPINGS

In this section, we use targeted perturbations from label m to target n, $\eta_{mn}(x) := \text{argmin}_\eta\{\|\eta\|_p : h(x) = m, h(x + \eta) = n\}$, to find best mappings w.r.t. the notion of coarse robustness. As such, we first define the matrix $G_p \in \mathbb{R}^{M\times M}_+$ with zero diagonal, whose entries represent the average of the l_p distances between examples $x \in \mathbb{R}^N$ and their misclassified perturbed versions $\hat{x} = x + \eta$, i.e., $G_p(m, n) = \frac{1}{|\mathcal{D}_m|} \sum_{x\in\mathcal{D}_m} \|\eta_{mn}(x)\|_p\,,\ m \neq n$.

Given T, we define matrix $C_p \in \mathbb{R}^{M_c\times M_c}_+$ whose entries,

$$C_p(i, j; T) = \frac{1}{|S_i||S_j|} \sum_{m\in S_i, n\in S_j} G_p(m, n)\,, \tag{4}$$

represent the average of the minimum perturbations to induce a misclassification of the coarse label from set S_i to set S_j. The diagonal of C_p is a measure of the average perturbations required to induce misclassifications within the same coarse set $S_i, i \in [M_c]$. In order to obtain the best grouping function T, for every coarse label $i \in [M_c]$, it requires on average a higher level of perturbation to misclassify the coarse labels (inter-set misclassification) than to misclassify the original prediction within the same coarse set (intra-set misclassification). Further, it is desirable that the mapping holds a semantic-based grouping of the labels. Therefore, for each row $i \in [M_c]$, we require that $\min_{j\in[M_c]\setminus\{i\}} C_p(i, j) = C_p(i, i)$. In order to quantify the quality of the grouping of labels, we seek a measure that captures the hardness of confusing the coarse predictor induced by T relative to classifier h, but also accounts for the variability between labels within each coarse class which is captured in the diagonal of C_p. To this end, we introduce the following.

Definition 4. *The Coarse Mapping Quality matrix (CMQ) is the zero diagonal matrix $\Pi_p(T) \in \mathbb{R}^{M_c\times M_c}$ associated with the grouping induced by the mapping T, whose entries are derived from matrix C_p as*

$$\Pi_p(i, j; T) = C_p(i, j; T) - C_p(i, i; T)\,, \quad i, j \in [M_c]\,. \tag{5}$$

This CMQ identifies features that best separate two or more classes. A larger value in Π_p captures the relative hardness of moving coarse label i to j under grouping function T, measured by the difference of the mean minimum perturbations for altering classifications between the coarse classes and within the class.

To characterize the overall quality of a mapping T based on the CMQ, we use $\alpha_p(T) = \sum_{i\in[M_c]} \sum_{j\in[M_c]} \Pi_p(i, j; T)$, $\beta_p(i; T) = \sum_{j\in[M_c]} \Pi_p(i, j; T)$.

Favorable mappings – in the sense of inducing well-separated coarse classes – will yield larger values of α_p. Therefore, to obtain the overall best mapping, we formulate the optimization problem (BM) in (6).

$$\max_T \{\alpha_p(T) : \Pi_p(i, j; T) > 0, \forall i, j \in [M_c], i \neq j\}\,. \tag{6}$$

Algorithm 1 Finding suboptimal mapping for (BM)

Input: G_p, M_c, Q, T_{ini}
Output: T^*

Initialize $T = T_{\text{ini}}$, P = {.}, iteration = 0
While iteration $\leq Q$
  get $\Pi_p(T)$, $(i_s, i_d) = \text{argmin}_{i,j\in[M_c]}\{\Pi_p(i,j;T) : |S_i| > 1\}$
  get $m_s = \text{argmin}_{m\in S_{i_s}, n\in S_{i_d}} G_p(m,n)$
  re-locate m_s from S_{i_s} to S_{i_d}. **update** T, iteration $\leftarrow$ iteration $+ 1$
  if $T \notin$ P, **update** P $\leftarrow$ P $\cup \{T\}$
  else, **restart** T uniformly at random
get V $= \{T \in$ P : Constrains. of (BM)$\}$
return $T^* = \text{argmax}_{T\in \text{V}}\, \alpha_p(T)$

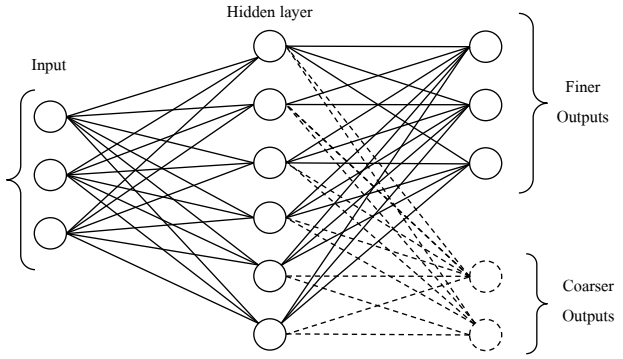

Fig. 1: An example of the additional parameters required for the AHCT method in a neural network classifier.

The program in (BM) searches for the best placement of M distinct classes into M_c non-empty groups. Therefore, the brute force approach to finding the optimal mapping amounts to a search over $\mathbb{S}(M, M_c)$ (the Stirling number of the second kind [11]) groupings. Thus, the optimal solutions to (BM) is $T^*_{\text{BM}} = \text{argmax}_{T\in L}\{\alpha_p(T) : \Pi_p(i,j;T) > 0, \forall i,j \in [M_c], i \neq j\}$, where L is the set of all possible mappings with $|L| = \mathbb{S}(M, M_c)$. We develop the procedure described in Algorithm 1 where we only consider $Q \ll \mathbb{S}(M, M_c)$ possible groupings depending on the CMQ matrix.

4. IMPROVING THE COARSE ROBUSTNESS

We consider $k : \mathbb{R}^N \to \Delta^{M+M_c}$ with entries $k_i, i \in [M + M_c]$, such that the fine and coarse predictions are given as $h(x) = \text{argmax}_{i\in[M]}\, k_i(x)$, and

$$T(x) = \underset{i\in\{M+1,\ldots,M+M_c\}}{\text{argmax}} \; k_i(x) - M \; . \tag{7}$$

We use θ_a to denote the set of the augmented adjustable parameters along with θ. Moreover, we introduce training dataset $\mathcal{D}_a$ with entries (x, p) where x is from $\mathcal{D}$, and vector $p \in \Delta^{M+M_c}$ is obtained from y as follows. The true finer and coarser labels are presented in vector p by stacking vectors $0.5\ 1_y$ and $0.5\ 1_z$, i.e., $p = 0.5[1_y \quad 1_z]^T$. See Fig. 1 for an example.

Definition 5. *Given classifier h', parameterized by θ_a, we define the Augmented Heads Coarse Training (AHCT) on dataset $\mathcal{D}_a$ as the task of minimizing the loss function in* (8) *in order to obtain high classification accuracy w.r.t. to the fine and coarse information.*

$$\min_{\theta_a} \frac{1}{|\mathcal{D}_a|} \sum_{(x,p)\in\mathcal{D}_a} \mathcal{L}(k(x;\theta_a), p) \; . \tag{8}$$

Theorem 1. *Given mapping T, NT classifier h with $c_p^{NT}(m;T)$, and AHCT classifier h' with $c_p^{AHCT}(m;T)$, then*

$$\mathbb{E}_{m\in[M]} c_p^{NT}(m;T) \leq \mathbb{E}_{m\in[M]} c_p^{AHCT}(m;T) \; . \tag{9}$$

Proof. Given an observation vector x, we use $\eta^{\text{NT}}(x)$ and $\eta^{\text{AHCT}}(x)$ to denote the minimum perturbations needed to induce coarse miss-classification for models trained by the NT and AHCT methods, respectively. This means,

$$\eta^{\text{NT}}(x) = \underset{\eta}{\text{argmin}}\{\|\eta\|_p : \underset{m\in[M]}{\text{argmax}}\, f_m(x+\eta) \notin S_{T(y)}\} \; , \tag{10}$$

$$\eta^{\text{AHCT}}(x) = \underset{\eta}{\text{argmin}}\{\|\eta\|_p : \underset{i\in[M]}{\text{argmax}}\, k_i(x+\eta) \notin S_{T(y)} \bigwedge \underset{i\in\{M+1,\ldots,M+M_c\}}{\text{argmax}} \; k_i(x+\eta) \neq T(y)\} \; . \tag{11}$$

Given the definition of the GCR, if we prove that $\|\eta^{\text{NT}}(x)\|_p \leq \|\eta^{\text{AHCT}}(x)\|_p$, then it follows that (9) is satisfied. The values of $\eta^{\text{NT}}(x)$ and $\eta^{\text{AHCT}}(x)$ are obtained using targeted attacks as follows. All target labels in the set outside the true coarser set, $t \in \bar{S}_{T(y)}$, are tried to get $\eta_t^{\text{NT}}(x)$ and $\eta_t^{\text{AHCT}}(x)$, then the minimum is selected. Given a target t, the targeted perturbations are obtained using the standard unrestricted targeted attack formulation for the NT and AHCT models as given in (12) and (13), respectively.

$$\begin{gathered} \min_{\eta} \|\eta\|_p \quad \text{s.t.} \\ f_t(x+\eta) > f_m(x+\eta), \forall m \in [M] \setminus \{t\} \; . \end{gathered} \tag{12}$$

$$\begin{gathered} \min_{\eta} \|\eta\|_p \quad \text{s.t.} \\ k_t(x+\eta) > k_i(x+\eta), \forall i \in [M] \setminus \{t\} \; , \\ k_{M+T(t)}(x+\eta) > k_{M+i}(x+\eta), \forall i \in [M_c] \setminus \{M + T(t)\} \; . \end{gathered} \tag{13}$$

Given the additional constraints in (13), the feasible set of (13) is a subset of the feasible region of (12). Hence, $\|\eta_t^{\text{NT}}(x)\|_p \leq \|\eta_t^{\text{AHCT}}(x)\|_p$ which yields to $\|\eta^{\text{NT}}(x)\|_p \leq \|\eta^{\text{AHCT}}(x)\|_p$, and that concludes the proof. □

5. NUMERICAL RESULTS

We consider the CIFAR10 [10] (FMNIST [12]) dataset by which class numbers 0-9 represent airplane (T-shirt), car

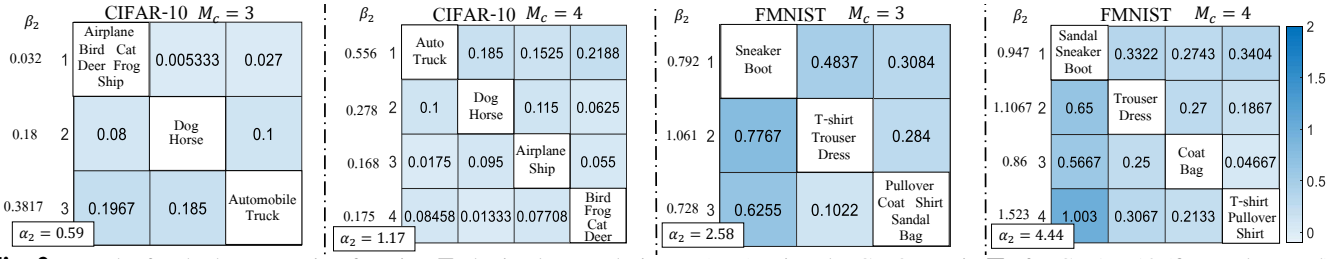

Fig. 2: Results for the best mapping function T obtained as a solution to (BM) using the CMQ matrix Π_2 for CIFAR-10 (first and second) and FMNIST (third and fourth).

Table 1: Performance of Algorithm 1 in obtaining the best overall mapping in comparison with the brute force method.

Dataset	$\mathbb{S}(10,3)$	α_2^*	$\mathbb{E}(\alpha_2)$	$\mathbb{E}(Q)$	$\mathbb{S}(10,4)$	α_2^*	$\mathbb{E}(\alpha_2)$	$\mathbb{E}(Q)$	$\mathbb{S}(10,5)$	α_2^*	$\mathbb{E}(\alpha_2)$	$\mathbb{E}(Q)$
CIFAR-10	9330	0.59	0.59	2139	34105	1.17	1.17	2017	42525	1.645	1.645	2369.8
FMNIST	9330	2.58	2.58	1308.4	34105	4.44	4.44	1828.4	42525	7.11	2477.6	3198.6

(trouser), bird (dress), cat (coat), deer (sandal), dog (shirt), frog (sneaker), horse (bag), ship (boot), and truck. We use standard convolutional neural network classifiers. We use the cross-entropy loss for $\mathcal{L}$ in both NT and AHCT, and the state-of-the-art targeted version of the Projected Gradient Descent attack [13] method with $p = 2$ to generate the perturbations. We use Intel(R) Core(TM) i9-9940 CPU @ 3.30GHz machine.

For our first experiment, where the outcomes are given in Fig. 2, we present results for the best mapping function T obtained as a solution to (BM) using the CMQ matrix Π_2. The first (last) two plots show the results for CIFAR-10 (FMNIST) with $M_c = 3$ and $M_c = 4$, respectively. The positive non-diagonal entries show the existence of mappings for which the average minimum perturbations that cause inter-set misclassification are larger than those causing intra-set misclassification. Under the derived mapping, it is most hard to misclassify the coarse class {'Auto','Truck'} as {'Bird', 'Cat', 'Deer', 'Frog'} for CIFAR-10 and {'Sandal','Sneaker','Boot'} as {'T-shirt', 'Pullover', 'Shirt'} for FMNIST. Moreover, irrespective of M_c, we observe that the obtained mappings are consistent with semantic-based groupings. For example, all animals are grouped in two coarse sets for CIFAR-10 with $M_c = 4$, 'Auto' and 'Truck' are grouped in CIFAR-10 for $M_c = 3$ and $M_c = 4$, and all foot-wearable items {'Sandal','Sneaker', and 'Boot'} are grouped in FMNIST for $M_c = 4$. Furthermore, higher values in Π_2 are associated with groupings that are more semantically consistent. For example, in the first CMQ, the grouping in the first row, which places instances of 'Ship' and 'Cat' in one set, returns a low $\beta_2 = 0.032$, while the semantic-based grouping of 'Auto' and 'Truck' returns $\beta_2 = 0.3817$. Interestingly, class separation is not always in one-to-one correspondence with natural semantic-based groupings as for the former example. Therefore, our analysis also sheds some light on a source of vulnerability of classifiers to imperceptible adversarial attacks in that they may not capture semantic similarities of classes. Specifically, the long-observed vulnerability is not very surprising considering the fact that certain labels could be close with regard to the amount of perturbation causing their misclassification while not being close semantically.

In Table 1, the performance of Algorithm 1 in obtaining the best overall mapping in comparison with the brute force method is demonstrated. We report the index of the optimal mapping α_2^* (brute force), the index of the mapping obtained by Algorithm 1 averaged over 5 random initial mappings T_{ini}, and the required number of iterations for both approaches. As shown, Algorithm 1 succeeds in obtaining the exact optimal mapping in all scenarios. The optimal values are obtained by our algorithm are confirmed by the brute force method for which all the $\mathbb{S}(10, M_c)$ possibilities are tried. For all cases, in general, Algorithm 1 requires a considerably smaller number of iterations to converge than brute force search.

For our third experiment in Table 2, we present results comparing the global standard and coarse robustness measures using natural training and the proposed augmented heads coarse training. Table 2 presents CIFAR10 and FMNIST results using semantic based and random mappings. For each dataset, the mapping that returns the highest CMQ is also considered. The first column is used to reference the experimental setting of each case. The last two columns represent the average global standard and coarse robustness measures which will be used as our evaluation of robustness when we present the following observations.

In all the considered scenarios (trained models and groupings), the GCR is higher than GSR. This indicates that inducing coarse miss-classification requires, on average, larger amount of perturbations when compared to those needed to induce any miss-classification. Furthermore, when we compare NT and AHCT models for the amount of average perturbations needed to cause any miss-classification using the GSR results, we observe that it is not necessary the AHCT scores higher as seen in run ID 10 vs. run ID 11 for an example. This reflects that inducing any miss-classification in the AHCT models may became easier as we are not using one-hot encoding for the finer label as in NT, and use a value of 0.5 to represent the true fine label as presented in the AHCT method.

Table 2: Results for enhancing the coarse robustness evaluated using the GCR and GSR measures for models trained using NT and AHCT.

Run ID	Dataset	Model	Training	# of Parameters	Coarse Sets	GSR $\mathbb{E}_m(g_2(m))$	GCR $\mathbb{E}_m(c_2(m;T))$
0	CIFAR-10	h	NT	1211786	$\{2,3,4,5,6,7\},\{0,8\},\{1,9\}$	0.149	0.2164
1	CIFAR-10	h'	AHCT	1212170	$\{2,3,4,5,6,7\},\{0,8\},\{1,9\}$	0.192	0.396
2	CIFAR-10	h	NT	1211786	$\{0,1,2,3\},\{4,5,6\},\{7,8,9\}$	0.149	0.171
3	CIFAR-10	h'	AHCT	1212170	$\{0,1,2,3\},\{4,5,6\},\{7,8,9\}$	0.207	0.241
4	CIFAR-10	h	NT	1211786	$\{0,2\},\{1,8,9\},\{3,5\},\{4,6,7\}$	0.149	0.182
5	CIFAR-10	h'	AHCT	1212302	$\{0,2\},\{1,8,9\},\{3,5\},\{4,6,7\}$	0.194	0.244
6	CIFAR-10	h	NT	1211786	$\{0,1,2\},\{3,4\},\{5,6,7\},\{8,9\}$	0.149	0.168
7	CIFAR-10	h'	AHCT	1212302	$\{0,1,2\},\{3,4\},\{5,6,7\},\{8,9\}$	0.202	0.22
8	FMNIST	h	NT	26506	$\{0,1,2\},\{3,4,5\},\{6,7,8,9\}$	0.399	0.424
9	FMNIST	h'	AHCT	26605	$\{0,1,2\},\{3,4,5\},\{6,7,8,9\}$	0.415	0.71
10	FMNIST	h	NT	26506	$\{1,3\},\{5,7,9\},\{0,2,4,6,8\}$	0.399	0.588
11	FMNIST	h'	AHCT	26605	$\{1,3\},\{5,7,9\},\{0,2,4,6,8\}$	0.385	1.369
12	FMNIST	h	NT	26506	$\{0,1,2\},\{3,4\},\{5,6,7\},\{8,9\}$	0.399	0.421
13	FMNIST	h'	AHCT	26638	$\{0,1,2\},\{3,4\},\{5,6,7\},\{8,9\}$	0.39	0.443
14	FMNIST	h	NT	26506	$\{0,6\},\{1,4,8\},\{2,3\},\{5,7,9\}$	0.399	0.467
15	FMNIST	h'	AHCT	26638	$\{0,6\},\{1,4,8\},\{2,3\},\{5,7,9\}$	0.467	0.545

6. CONCLUSION

In this paper, we first introduced global measures of the notion of coarse robustness. An efficient algorithm to identify robust groupings relative to the introduced measures was developed. Furthermore, we presented a method to improve the coarse robustness using a modified structure that incorporates the coarse information. Further, we proved the enhancement theoretically and empirically. An intriguing observation of our experiments on benchmark datasets for image classification is that semantically plausible groupings of the class labels are often consistent with large values of the measures introduced.

7. REFERENCES

[1] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," in *International Conference on Learning Representations*, 2018.

[2] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[3] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami, "The limitations of deep learning in adversarial settings," in *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016, pp. 372–387.

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193.

[5] Ismail R. Alkhouri and George K. Atia, "Adversarial attacks on coarse-to-fine classifiers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 2855–2859.

[6] Emilio Rafael Balda, Arash Behboodi, and Rudolf Mathar, "Perturbation analysis of learning algorithms: Generation of adversarial examples from classification to regression," *IEEE Transactions on Signal Processing*, vol. 67, no. 23, pp. 6078–6091, 2019.

[7] Ismail R Alkhouri, Alvaro Velasquez, and George K Atia, "Adversarial perturbation attacks on nested dichotomies classification systems," in *31st IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021, pp. 1–6.

[8] Gengxing Wang, Xinyuan Chen, and Chang Xu, "Adversarial watermarking to attack deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1962–1966.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[10] Alex Krizhevsky et al., "Learning multiple layers of features from tiny images," 2009, https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf.

[11] Louis Comtet, *Advanced Combinatorics: The art of finite and infinite expansions*, Springer Science & Business Media, 2012.

[12] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[13] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.