I AM GOING MAD: MAXIMUM DISCREPANCY COMPETITION FOR COMPARING CLASSIFIERS ADAPTIVELY

Anonymous authors
Paper under double-blind review

ABSTRACT

The learning of hierarchical representations for image classification has experienced an impressive series of successes due in part to the availability of large-scale labeled data for training. On the other hand, the trained classifiers have traditionally been evaluated on a handful of test images, which are deemed to be extremely sparsely distributed in the space of all natural images. It is thus questionable whether recent performance improvements on the excessively re-used test sets generalize to real-world natural images with much richer content variations. In addition, studies on adversarial learning show that it is effortless to construct adversarial examples that fool nearly all image classifiers, adding more complications to relative performance comparison of existing models. This work presents an efficient framework for comparing image classifiers, which we name the *MAximum Discrepancy* (MAD) competition. Rather than comparing image classifiers on fixed test sets, we adaptively sample a test set from an arbitrarily large corpus of unlabeled images so as to maximize the discrepancies between the classifiers, measured by the distance over WordNet hierarchy. Human labeling on the resulting small and model-dependent image sets reveals the relative performance of the competing classifiers and provides useful insights on potential ways to improve them. We report the MAD competition results of eleven ImageNet classifiers while noting that the framework is readily extensible and cost-effective to add future classifiers into the competition.

1 INTRODUCTION

Large-scale human-labeled image datasets such as ImageNet (Deng et al., 2009) have greatly contributed to the rapid progress of research in image classification. In recent years, considerable effort has been put to designing novel network architectures (He et al., 2016; Hu et al., 2018) and advanced optimization algorithms (Kingma & Ba, 2015) to improve the training of image classifiers based on deep neural networks (DNNs), while little attention has been paid to comprehensive and fair evaluation/comparison of their model performance. Conventional model evaluation methodology for image classification generally follows a three-step approach (Burnham & Anderson, 2003). First, pre-select a number of images from the space of all possible natural images (*i.e.*, natural image manifold) to form the test set. Second, collect the human label for each image in the test set to identify its ground-truth category. Third, rank the competing classifiers according to their goodness of fit (*e.g.*, accuracy) on the test set; the one with the best goodness of fit is declared the winner.

A significant problem with this methodology is the apparent contradiction between the enormous size and high dimensionality of natural image manifold and the limited scale of affordable testing (*i.e.*, human labeling, or verifying predicted labels, which is expensive and time consuming). As a result, a typical "large-scale" test set for image classification allows for only tens of thousands of natural images to be examined, which are deemed to be extremely sparsely distributed in natural image manifold. Model comparison based on a limited number of samples assume that they are *sufficiently representative* of the whole population, an assumption that has been proven to be doubtful in image classification. Specifically, Recht et al. (2019) found that a minute natural distribution shift leads to a large drop in accuracy for a broad range of image classifiers on both CIFAR-10 (Krizhevsky, 2009) and ImageNet (Deng et al., 2009), suggesting that the current test sets may far less suffice to represent hard natural images encountered in the real world. Another problem with the conventional model comparison methodology is that the test sets are pre-selected and therefore fixed. This leaves

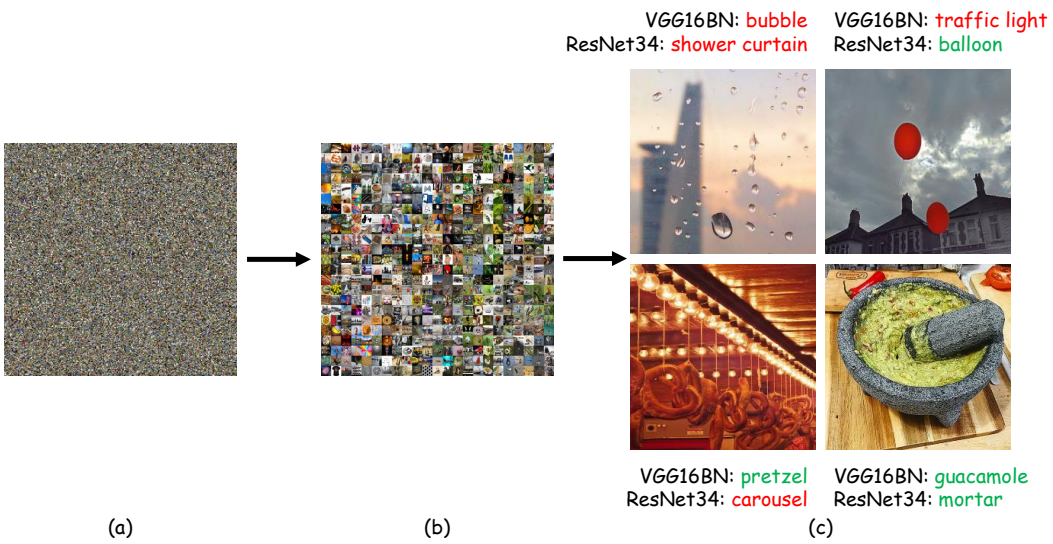

Figure 1: Overview of the MAD competition pipeline. **(a)**: A large unlabeled image set of web scale. **(b)**: The subset of natural images selected from (a) on which two classifiers (VGG16BN and ResNet34 in this case) make different predictions. Note that collecting the class label for each image in this subset may still be prohibitive because of its gigantic size. **(c)**: Representative examples sampled from top-k images on which VGG16BN's and ResNet34's predictions differ the most, quantified by Eq. (3). Although the two classifiers have nearly identical accuracies on ImageNet validation set, the proposed MAD competition methodology successfully distinguishes them by finding their respective counterexamples. This sheds light on potential ways to improve the two classifiers or combine them into a better one. The model predictions are shown along with the images, where green and red indicate correct and incorrect predictions, respectively.

open the door of adapting classifiers to the test images, deliberately or unintentionally, via extensive hyperparameter tuning, raising the risk of overfitting. As a result, it is never guaranteed that image classifiers with highly competitive performance on such small and fixed test sets can generalize to real-world natural images with much richer content variations. In addition, recent studies on adversarial learning (Goodfellow et al., 2015; Madry et al., 2018) indicate that adversarial images produced to mislead a specific image classifier have strong transferability to fool other classifiers even if their design philosophies differ substantially, which further complicates the relative performance comparison of existing classifiers.

In order to reliably measure the progress in image classification and to fairly test the generalizability of existing classifiers in a natural setting, we believe a much larger test set in the order of millions or even billions must be used. Apparently, the main challenge here is how to exploit such a large-scale test set under the constraint of the very limited budget for human labeling, knowing that collecting ground-truth labels for all images is extremely difficult, if not impossible.

In this work, we propose an efficient and practical methodology, namely the *MAximum Discrepancy* (MAD) competition, to meet this challenge. Instead of trying to *prove* an image classifier to be correct using a small and fixed test set, MAD starts with a large-scale unlabeled image set and attempts to *falsify* a classifier by finding a set of images, whose predictions are in strong disagreements with the rest competing classifiers (See Figure 1). A classifier that is harder to be falsified is considered better. The initial image set for MAD to explore can be made arbitrarily large provided that the cost of computational prediction for all competing classifiers is cheap. To quantify the discrepancy between two classifiers on one image, we propose a weighted distance over WordNet hierarchy (Miller, 1998), which is more semantically aligned with human cognition compared with traditional binary judgment (agree vs. disagree). The set of model-dependent images selected by MAD are the most informative in discriminating the competing classifiers. Subjective experiments on the MAD test set reveal the relative strengths and weaknesses among the classifiers and identify the training techniques and architecture choices that improve the generalizability to natural image manifold. Moreover, careful inspection of the selected images may suggest potential ways to improve a classifier or to combine aspects of multiple classifiers.

We apply the MAD competition to compare eleven ImageNet classifiers and find that MAD verifies the relative improvements achieved by recent DNN-based methods, with a minimal subjective testing budget. MAD is readily extensible, allowing future classifiers to be added into competition with little additional cost. Moreover, the application scope of MAD is far beyond image classification. It can be extended to many other research fields that expect discrete-valued outputs, and is especially useful when the sample space is enormous and the ground-truth measurement is expensive.

2 THE MAD COMPETITION METHODOLOGY

The general problem of model comparison in image classification may be formulated as follows. We work with the natural image manifold $\mathcal{X}$, upon which we define a class label $f(x) \in \mathcal{Y}$ for all natural images $x \in \mathcal{X}$, where $\mathcal{Y} = \{1, 2, \ldots, c\}$ indicates which of c categories the object in x belongs to. We assume a subjective assessment environment, in which a human subject can identify the category membership for any natural image x among all possible categories. A group of image classifiers $\mathcal{F} = \{f_i\}_{i=1}^m$ are also assumed, each of which takes a natural image x as input and makes a prediction of $f(x)$, collectively denoted by $\{f_i(x)\}_{i=1}^m$. The goal is to compare the relative performance of m classifiers under very limited resource for subjective testing.

A conventional model comparison method for image classification first samples a natural image set $\mathcal{D} = \{x_k\}_{k=1}^n \subset \mathcal{X}$. For each image $x_k \in \mathcal{D}$, we ask human annotators to provide the ground-truth label $f(x_k) \in \mathcal{Y}$. Since human labeling is an expensive and time consuming process and DNN-based classifiers are hungry for labeled data in the training stage (Krizhevsky et al., 2012), n is typically small in the order of tens of thousands (Russakovsky et al., 2015). The predictions of the classifiers are compared with the human labels by computing the empirical classification accuracy

$$\mathrm{Acc}(f_i; \mathcal{D}) = \frac{1}{n} \sum_{x \in \mathcal{D}} \mathbb{I}[f_i(x) = f(x)], \text{ for } i = 1, \ldots, m. \tag{1}$$

The classifier f_i with a higher classification accuracy is said to outperform f_j with a lower accuracy.

As an alternative, the proposed MAD competition methodology aims to *falsify* a classifier in the most efficient way with the help of other competing classifiers. A classifier that is more likely to be falsified is considered worse.

2.1 THE MAD COMPETITION PROCEDURE

The MAD competition methodology starts by sampling an image set $\mathcal{D} = \{x_k\}_{k=1}^n$ from the natural image manifold $\mathcal{X}$. Since the number of images selected by MAD for subjective testing is independent of the size of $\mathcal{D}$, we may make n arbitrarily large such that $\mathcal{D}$ provides dense coverage of (*i.e.*, sufficiently represents) $\mathcal{X}$. MAD relies on a distance measure to quantify the degree of discrepancy between the predictions of any two classifiers. The most straightforward measure is the 0-1 loss:

$$d_{01}(f_i(x), f_j(x)) = \mathbb{I}[f_i(x) \neq f_j(x)]. \tag{2}$$

Unfortunately, it ignores the semantic relations between class labels, which may be crucial in distinguishing two classifiers, especially when they share similar design philosophies (*e.g.*, using DNNs as backbones) and is trained on the same image set (*e.g.*, ImageNet). For example, misclassifying a "chihuahua" as a dog of other species is clearly more acceptable compared to misclassifying it as a "watermelon". We propose to leverage the semantic hierarchies in WordNet (Miller, 1998) to measure the distance between two (predicted) class labels. Specifically, we model WordNet as a weighted undirected graph $G(V, E)$[1]. Each edge $e = (u, v) \in E$ connects a parent node u of a more general level (*e.g.*, canine) to its child node v of a more specific level (*e.g.*, dog), for $u, v \in V$. A nonnegative weight $w(e)$ is assigned to each edge $e = (u, v)$ to encode the semantic similarity between u and v. A larger $w(e)$ indicates that u and v are semantically more dissimilar. We measure the distance between two labels as the sum of the weights assigned to the edges along the shortest path $\mathcal{P}$ connecting them

$$d_w(f_i(x), f_j(x)) = \sum_{e \in \mathcal{P}} w(e). \tag{3}$$

[1] Although WordNet is tree-structured, a child node in WordNet may have multiple parent nodes.

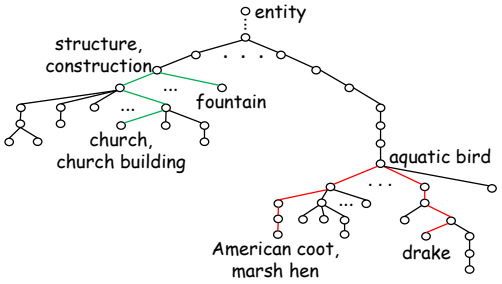

Figure 2: Comparison of weighted and unweighted distances. In the sub-tree of WordNet, we highlight the shortest paths from "fountain" to "church" and from "drake" to "American coot" in green and red, respectively. The semantic distance between the two aquatic birds is much shorter than that between the two constructions of completely different functionalities (verified by our internal subjective testing). The proposed weighted distance is well aligned with human cognition by assigning a much smaller distance to the red path (0.0037) compared to the green one (0.0859).

Eq. (3) reduces to the standard graph hop distance between two vertices by setting $w(e) = 1$. Here we design $w(e)$ to be inversely proportional to the tree depth level l of the parent node (*e.g.*, $w(e) \propto 2^{-l}$). In other words, we prefer the shortest paths to traverse the root node (or nodes with smaller l) as a way of encouraging $f_i(x)$ and $f_j(x)$ to differ in a more general level (*e.g.*, vehicle rather than watercraft). Figure 2 shows the semantic advantages of our choice of weights compared to the equal weight. With the distance measure at hand, the optimal image in terms of discriminating f_i and f_j can be obtained by maximizing the discrepancies between the two classifiers on $\mathcal{D}$

$$x^\star = \arg\max_{x \in \mathcal{D}} d_w(f_i(x), f_j(x)). \tag{4}$$

The queried image label $f(x^\star)$ leads to three possible outcomes (see Figure 1):

- **Case I**. Both classifiers make correct predictions. Although theoretically impossible based on the general problem formulation, it is not uncommon in practice that a natural image may contain multiple distinct objects (*e.g.*, guacamole and mortar). In this case, f_i and f_j successfully recognize different objects in $x^\star$, indicating that both classifiers tend to perform at a high level and cannot be differentiated. By restricting $\mathcal{D}$ to contain only natural images with a single salient object, we may reduce the possibility of this outcome.
- **Case II**. f_i (or f_j) makes correct prediction, while f_j (or f_i) makes incorrect prediction. In this case, MAD automatically identifies a strong failure case to falsify one classifier, not the other; a clear winner is obtained. The selected image $x^\star$ provides the strongest evidence in ranking the relative performance between the two classifiers.
- **Case III**. Both classifiers make incorrect predictions in a multiclass image classification problem (*i.e.*, $c \geq 3$). Although both classifiers make mistakes, they differ substantially during inference, which in turn provides a strong indication of their respective weaknesses and suggests potential ways to combine them into a single better classifier [2]. However, in this case, $x^\star$ contributes little to performance comparison between the two classifiers.

To obtain a reliable performance comparison between f_i and f_j in practice, we choose top-k images in $\mathcal{D}$ with k largest distances computed by Eq. (3) to form the test subset $\mathcal{S}_{\{i,j\}}$. MAD runs this game among all $\binom{m}{2}$ distinct pairs of classifiers, resulting in the final MAD test set $\mathcal{S} = \bigcup \mathcal{S}_{\{i,j\}}$. The number of natural images in $\mathcal{S}$ is at most $m(m-1)k/2$, which is independent of the size n of $\mathcal{D}$. As a result, applying MAD to a larger image set has no impact on the cost of human labeling. In scenarios where the cost of computational prediction can be ignored, MAD encourages to expand $\mathcal{D}$ to cover as many "free" natural images as possible.

We now describe our subjective assessment environment for collecting human labels. Given an image $x \in \mathcal{S}$, which is associated with two classifiers f_i and f_j, we pick two binary questions for human annotators: "Does x contain an $f_i(x)$?" and "Does x contain an $f_j(x)$?". When both answers are no (corresponding to Case III), we cease to source the ground-truth label of x because $f(x)$ is uninformative in distinguishing f_i and f_j. Nevertheless, it is difficult for humans to select one among c classes, especially when c is large and the ontology of classes is complex.

[2] This is in stark contrast to natural adversarial examples in ImageNet-A (Hendrycks et al., 2019), where different image classifiers tend to make consistent mistakes. For example, VGG16BN and ResNet34 make the same incorrect predictions on the $3,149$ out of $7,500$ images in ImageNet-A.

Algorithm 1: The MAD competition

Input: An unlabeled image set $\mathcal{D}$, a group of image classifiers $\mathcal{F} = \{f_i\}_{i=1}^m$ to be ranked, a distance measure d_w defined over WordNet hierarchy

Output: A global ranking vector $r \in \mathbb{R}^m$

$\mathcal{S} \leftarrow \emptyset$, $B \leftarrow I$
for $i \leftarrow 1$ **to** m **do**
  Compute classifier predictions $\{f_i(x), x \in \mathcal{D}\}$
end
for $i \leftarrow 1$ **to** m **do**
  for $j \leftarrow i+1$ **to** m **do**
    Compute the distances using Eq. (3) $\{d_w(f_i(x), f_j(x)), x \in \mathcal{D}\}$
    Select top-k images with k largest distances to form $\mathcal{S}_{\{i,j\}}$
    $\mathcal{S} \leftarrow \mathcal{S} \bigcup \mathcal{S}_{\{i,j\}}$
  end
end
Source human labels for $\mathcal{S}$
Compute the pairwise accuracy matrix A with $a_{ij} = \mathrm{Acc}(f_i; \mathcal{S}_{\{i,j\}})$ using Eq. (1)
Compute the pairwise dominance matrix B with $b_{ij} = a_{ij}/a_{ji}$
Compute the global ranking vector r using Eq. (5)

Algorithm 2: Adding a new classifier into the MAD competition

Input: An unlabeled image set $\mathcal{D}$, the pairwise dominance matrix $B \in \mathbb{R}^{m \times m}$ for $\mathcal{F} = \{f_i\}_{i=1}^m$, a new classifier f_{m+1} to be ranked, d_w

Output: A global ranking vector $r \in \mathbb{R}^{m+1}$

$\mathcal{S} \leftarrow \emptyset$, $B' \leftarrow \begin{bmatrix} B & 0 \\ 0^T & 1 \end{bmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}$
Compute the predictions of the new image classifier $\{f_{m+1}(x), x \in \mathcal{D}\}$
for $i \leftarrow 1$ **to** m **do**
  Compute the distances using Eq. (3) $\{d_w(f_i(x), f_{m+1}(x)), x \in \mathcal{D}\}$
  Select top-k images with k largest distances to form $\mathcal{S}_{\{i,m+1\}}$
  $\mathcal{S} = \mathcal{S} \bigcup \mathcal{S}_{\{i,m+1\}}$
end
Source human labels for $\mathcal{S}$
for $i \leftarrow 1$ **to** m **do**
  $a_{i,m+1} = \mathrm{Acc}(f_i; \mathcal{S}_{\{i,m+1\}})$
  $a_{m+1,i} = \mathrm{Acc}(f_{m+1}; \mathcal{S}_{\{i,m+1\}})$
end
Update the pairwise dominance matrix B' with $b'_{i,m+1} = 1/b'_{m+1,i} = a_{i,m+1}/a_{m+1,i}$
Compute the global ranking vector $r \in \mathbb{R}^{m+1}$ using Eq. (5)

After subjective testing, we first compare the classifiers in pairs and aggregate the pairwise statistics into a global ranking. Specifically, we compute the empirical classification accuracies of f_i and f_j on $\mathcal{S}_{\{i,j\}}$ using Eq. (1), denoted by a_{ij} and a_{ji}, respectively. When k is small, Laplace smoothing is employed to smooth the estimation. Note that $a_{ij} + a_{ji}$ may be greater than one because of Case I. The pairwise accuracy statistics of all classifiers form a matrix A, from which we compute another matrix B with $b_{ij} = a_{ij}/a_{ji}$ indicating the pairwise dominance of f_i over f_j. We aggregate pairwise comparison results into a global ranking $r \in \mathbb{R}^m$ using Perron rank (Saaty & Vargas, 1984):

$$r = \lim_{t \to \infty} \frac{1}{t} \sum_{\alpha=1}^{t} \frac{B^\alpha 1}{1^T B^\alpha 1}, \tag{5}$$

where 1 is an m-dimensional vector of all ones. The limit of Eq. (5) is the normalized principal eigenvector of B corresponding to the largest eigenvalue, where $r_i > 0$ for $i = 1, \dots m$ and $\sum_i r_i = 1$. The larger r_i is, the better f_i performs in the MAD competition. Other ranking aggregation methods such as HodgeRank (Jiang et al., 2011) may also be applied. We summarize the workflow of the MAD competition in Algorithm 1.

Finally, it is straightforward and cost-effective to add the $m+1$-th classifier into the current MAD competition. No change is necessary for the sampled $\mathcal{S}$ and the associated subjective testing. The only additional work is to select a total of mk new images from $\mathcal{D}$ for human labeling. We then enlarge B by one along its row and column, and insert the pairwise comparison statistics between f_{m+1} and the previous m classifiers. An updated global ranking vector $r \in \mathbb{R}^{m+1}$ can be computed using Eq. (5). We summarize the procedure of adding a new classifier in Algorithm 2.

3 Application to ImageNet Classifiers

In this section, we apply the proposed MAD competition methodology to comparing ImageNet classifiers. We focus on ImageNet (Deng et al., 2009) not only because it is one of the first large-scale

datasets in image classification, but also because the improvements on ImageNet seem to plateau, which sets up an ideal platform for MAD to distinguish the newly proposed image classifiers finer.

3.1 EXPERIMENTAL SETUPS

Constructing $\mathcal{D}$ Inspired by (Hendrycks et al., 2019), we focus on the same 200 out of $1,000$ classes to avoid rare and abstract classes, and classes that have changed much since 2012. For each class, we crawl a large number of images from Flickr, resulting in a total of $n = 168,000$ natural images. Although MAD allows us to arbitrarily increase n with essentially no cost, we choose the size of $\mathcal{D}$ to be only approximately three times larger than the ImageNet validation set to provide a relatively easy environment for probing the generalizability of the classifiers. As will be cleared in Section 3.2, the current setting of n is sufficient to discriminate the competing classifiers. To guarantee the content independence between ImageNet and our test set, we collect images that have been uploaded after 2013. It is worth noting that no data cleaning (*e.g.*, inappropriate content and near-duplicate removal) is necessary at this stage since we only need to ensure the selected subset $\mathcal{S}$ for human labeling are eligible.

Competing Algorithms We select eleven representative ImageNet classifiers for benchmarking: VGG16BN (Simonyan & Zisserman, 2014) with batch normalization (Ioffe & Szegedy, 2015), ResNet34, ResNet101 (He et al., 2016), WRN101-2 (Zagoruyko & Komodakis, 2016), ResNeXt101-32×4 (Xie et al., 2017), SE-ResNet-101, SENet154 (Hu et al., 2018), NASNet-A-Large (Zoph et al., 2018), PNASNet-5-Large (Liu et al., 2018), EfficientNet-B7 (Tan & Le, 2019), and WSL-ResNeXt101-32×48 (Mahajan et al., 2018). Since VGG16BN and ResNet34 have nearly identical accuracies (both top-1 and top-5) on ImageNet validation set, it is particularly interesting to see which method generalizes better in natural image manifold. We compare ResNet34 with ResNet101 to see the influence of DNN depth on generalizability. WRN101-2, ResNeXt101-32×4, SE-ResNet-101 are different improved versions over ResNet-101. We also include two state-of-the-art classifiers on ImageNet validation set: WSL-ResNeXt101-32×48 and EfficientNet-B7. The former leverages the power of weakly supervised pre-training on Instagram data, while the latter makes use of compound scaling method. We use publicly available code repositories for all DNN-based models, whose top-1 accuracies on ImageNet validation set are listed in Table 1 for reference.

Constructing $\mathcal{S}$ When constructing $\mathcal{S}$ using the maximum discrepancy principle, we add another constraint based on prediction confidence. Specifically, a candidate image x associated with f_i and f_j is filtered out if it does not satisfy $\min(p_i(x), p_j(x)) \geq T$, where $p_i(x)$ is the confidence score (*i.e.*, probability produced by the last softmax layer) of $f_i(x)$ and T is a predefined threshold set to 0.8. We include the confidence constraint for two main reasons. First, if f_i misclassifies x with low confidence, it is highly likely that x is near the decision boundary and thus contains less information on improving the decision rules of f_i. Second, some images in $\mathcal{D}$ do not necessarily fall into the $1,000$ classes in ImageNet, which are bound to be misclassified (a problem closely related to out-of-distribution detection). If they are misclassified by f_i with high confidence, we consider them as hard counterexamples of f_i. To encourage class diversity in $\mathcal{S}$, we remain a maximum of three images with the same predicted label by f_i. In addition, we exclude images that are non-natural. Figure 3 visually compares representative "manhole cover" images in $\mathcal{S}$ and ImageNet validation set (see more in Figure 6).

Collecting Human Labels As described in 2.1, given an image $x \in \mathcal{S}$, human annotators need to answer two binary questions. In our subjective experiment, we choose $k = 30$ and invite five volunteer graduate students, who are experts in computer vision to label a total of $11 \times 10 \times 30/2 = 1,650$ images. If more than three of them find difficulty in labeling x (associated with f_i and f_j), it is discarded and replaced by $x' \in \mathcal{D}$ with the $k+1$-th largest distance $d_w(f_i(x'), f_j(x'))$. Majority vote is adopted to decide the final label when disagreement occurs. After subjective testing, we find that 53.5% of annotated images belong to Case II, which form the cornerstone of the subsequent data analysis. Besides, 32.9% and 13.6% images pertain to Case I and Case III, respectively.

3.2 EXPERIMENTAL RESULTS

Pairwise Ranking Results Figure 4 shows the pairwise accuracy matrix A in the current MAD competition, where a larger value of an entry (a brighter color) indicates a higher accuracy in the subset selected together by the corresponding row and column models. An interesting phenomenon

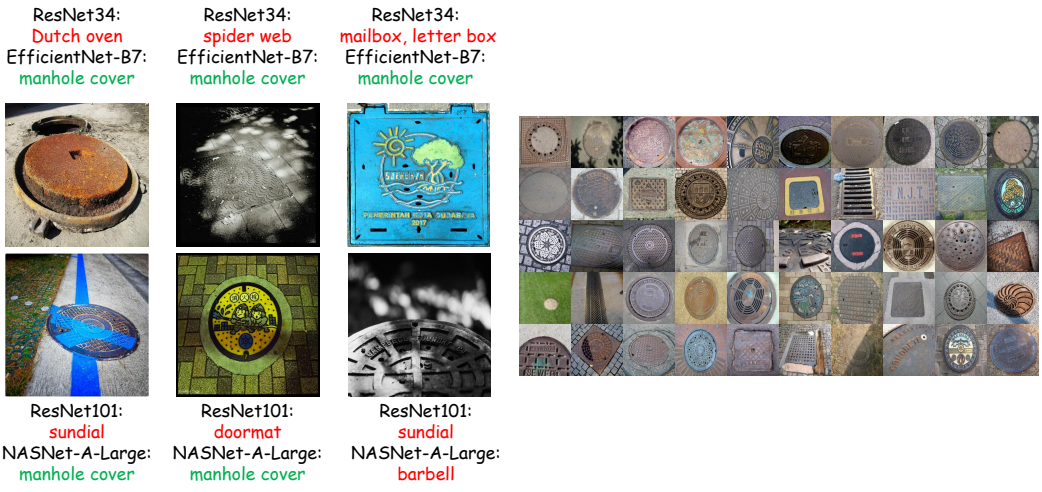

Figure 3: Visual comparison of images selected by MAD and in ImageNet validation set. **(a)**: "manhole cover" images selected by MAD along with the predictions by the associated classifiers. **(b)**: All "manhole cover" images in ImageNet validation set. The MAD-selected images are visually much harder, which contain diverse and non-trivial distortions, *e.g.*, occlusion, shading and unbalanced lightening, complex background, rare colors, and untypical viewing points. In contrast, ImageNet images mainly include a single center-positioned object with relatively clean background, whose shape, color and viewing point are common.

we find is that when two classifiers f_i and f_j perform at a similar level on $\mathcal{S}_{\{i,j\}}$ (*i.e.*, $|\text{Acc}(f_i) - \text{Acc}(f_j)|$ is small), $\max(\text{Acc}(f_i), \text{Acc}(f_j))$ is also small. That is, more images on which they both make incorrect but different predictions (Case III) have been selected compared to images falling into Case I. Taking a closer look at images in $\mathcal{S}_{\{i,j\}}$, we may reveal the respective model biases of f_i and f_j. For example, we find that WSL-ResNeXt101-32×48 tend to focus on foreground objects while EfficientNet-B7 attends more to background objects (See Figure 8). We also observe several common failure modes of the competing classifiers through pairwise comparison, *e.g.*, excessive reliance on relation inference (see Figure 9), bias towards low-level visual features (see Figure 10), and difficulty in recognizing rare instantiations of objects (see Figures 3 and 6).

Global Ranking Results We present the global ranking results by MAD in Table 1, where we find that MAD tracks the steady progress in image classification, as verified by a reasonable Spearman rank-order correlation coefficient (SRCC) of 0.89 between the accuracy rank on ImageNet validation set and the MAD rank on our test set $\mathcal{D}$. Moreover, by looking at the differences between the two rankings, we obtain a number of interesting findings. First, VGG16BN outperforms not only ResNet34 but also ResNet101, suggesting that under similar computation budgets, VGG-like networks may exhibit better generalizability to hard samples than networks with residual connections. Second, both networks equipped with the squeeze-and-extraction mechanism, *i.e.* SE-ResNet-101 and SENet154, move up by two places in the MAD ranking. This indicates that explicitly modeling dependencies between channel-wise feature maps seem quite beneficial to image classification. Third, for the two models that

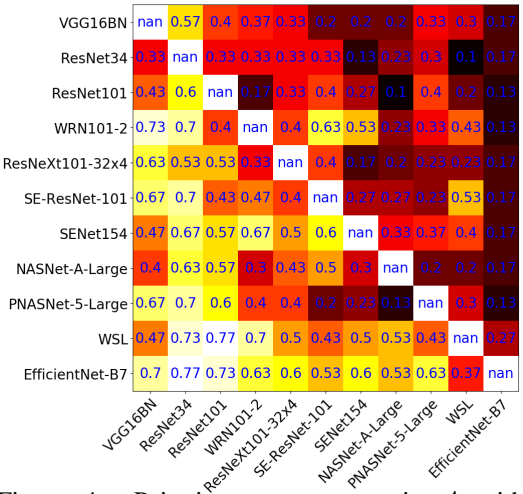

Figure 4: Pairwise accuracy matrix A with brighter colors indicating higher accuracies (blue numbers). WSL-ResNeXt101-32×48 is abbreviated to WSL for neat presentation.

Models	ImageNet top-1 Acc	Acc Rank	MAD Rank	Δ Rank
WSL-ResNeXt101-32×48 (Mahajan et al., 2018)	85.44	1	2	-1
EfficientNet-B7 (Tan & Le, 2019)	84.48	2	1	1
PNASNet-5-Large (Liu et al., 2018)	82.74	3	7	-4
NASNet-A-Large (Zoph et al., 2018)	82.51	4	4	0
SENet154 (Hu et al., 2018)	81.30	5	3	2
WRN101-2 (Zagoruyko & Komodakis, 2016)	78.85	6	6	0
SE-ResNet-101 (Hu et al., 2018)	78.40	7	5	2
ResNeXt101-32×4 (Xie et al., 2017)	78.19	8	8	0
ResNet101 (He et al., 2016)	77.37	9	10	-1
VGG16BN (Simonyan & Zisserman, 2014)	73.36	10	9	1
ResNet34 (He et al., 2016)	73.31	11	11	0

Table 1: Global ranking results. A smaller rank indicates better performance.

exploit neural architecture search, NASNet-A-Large is still ranked high by MAD; interestingly, the rank of PNASNet-5-Large drops a lot. That implies MAD may prefer the global search strategy used in NASNet-A-Large to the progressive cell-wise search strategy adopted in PNASNet-5-Large, although the former is slighter inferior in ImageNet top-1 accuracy. Last but not least, the top-2 performers, WSL-ResNeXt101-32×48 and EfficientNet-B7, are still the best in MAD competition (irrespective of their relative rankings), verifying the effectiveness of large-scale hashtag data pre-training and compound scaling method in the context of image classification.

Ablation Study We analyze the key hyper-parameter k in MAD, *i.e.*, the number of images in $\mathcal{S}_{\{i,j\}}$ selected for subjective testing. We calculate the SRCC values between the top-30 ranking (as reference) and other top-k rankings with $k = \{1, 2, \ldots, 29\}$. As shown in Figure 5, the ranking results are fairly stable (SRCC > 0.90) when $k > 15$. This supports our choice of $k = 30$ since the final global ranking already seems to enter a stable plateau.

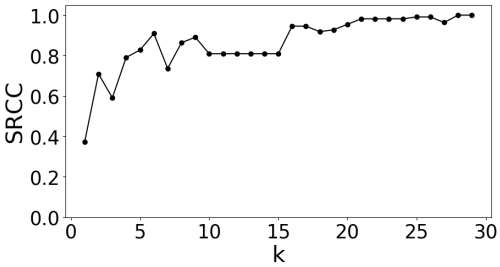

Figure 5: The SRCC values between top-30 and other top-k rankings, $k = \{1, 2, \ldots, 29\}$.

4 Discussion

We have presented a new methodology for comparing image classification models. MAD effectively mitigates the conflict between the prohibitively large natural image manifold that we have to evaluate against and the expensive human labeling effort that we aim to minimize. Much of our endeavor has been dedicated to selecting natural images that are optimal in term of distinguishing or falsifying classifiers. MAD requires explicit specification of image classifiers to be compared, and provides an effective means of exposing the respective flaws of competing classifiers. It also directly contributes to model interpretability and helps us analyze the models' focus and bias when making predictions. We have demonstrated the effectiveness of MAD competition on ImageNet classifiers, and concluded a number of interesting observations, which were not apparently drawn from the (often quite close) accuracy numbers on the small and fixed ImageNet validation set.

MAD is widely applicable to computational models that produce discrete-valued outputs, and is particularly useful when the sample space is large and the ground-truth label being predicted is expensive to measure. Examples include medical and hyperspectral image classification (Filipovych & Davatzikos, 2011; Wang et al., 2014), where signification domain expertise is crucial to obtain correct labels. MAD can also be applied towards spotting rare but fatal failures in high-cost and failure-sensitive applications, *e.g.*, comparing perception systems of autonomous cars (Chen et al., 2015) in unconstrained real-world weathers, lighting conditions, and road scenes. In addition, by restricting the test set to some domain of interest, MAD allows comparison of classifiers in more specific applications, *e.g.*, fine-grained image recognition (Xiao et al., 2015).

We also feel it important to note the current limitations of MAD. First, MAD cannot prove a model's predictions to be correct and therefore it should be viewed as complementary to, rather than a replacement for, the conventional accuracy comparison for image classification. Second, although the

distance in Eq. (3) is sufficient to distinguish multiple classifiers in the current experimental setting, it does not yet fully reflect human cognition of image label semantics. Third, the computation of the confidence used to select images is not perfectly grounded. How to marry the MAD competition with Bayesian probability theory to model uncertainties during image selection is an interesting direction for future research.

Our method arises as a natural combination of concepts drawn from two separate lines of research. The first explores the idea of model falsification as a model comparison. Wang & Simoncelli (2008) introduced the maximum differentiation competition for comparing computational models of *continuous* perceptual quantities, which was further extended by (Ma et al., 2019). Berardino et al. (2017) developed a computational method for comparing hierarchical image representations in terms of their ability to explain perceptual sensitivity in humans. MAD, on the other hand, is tailored to applications with *discrete* model responses and relies on a semantic distance measure to compute model discrepancy. The second endeavour arises from machine learning literature on generating adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2015; Madry et al., 2018) and evaluating image classifiers on new test sets (Geirhos et al., 2019; Recht et al., 2019; Hendrycks & Dietterich, 2019; Hendrycks et al., 2019). The images selected by MAD can be seen as a form of natural adversarial examples as each of them is able to fool at least one classifier (when Case I is eliminated). Unlike adversarial images with inherent transferability to mislead most classifiers, MAD-selected images emphasize on their discriminability of the competing models. Different from recently created test sets, the MAD-selected test set is adapted to the competing classifiers with the goal of minimizing human labeling effort.

REFERENCES

Alexander Berardino, Valero Laparra, Johannes Ballé, and Eero Simoncelli. Eigen-distortions of hierarchical representations. In *Advances in Neural Information Processing Systems*, pp. 3530–3539. 2017.

Kenneth P Burnham and David R Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media, 2003.

Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. DeepDriving: Learning affordance for direct perception in autonomous driving. In *IEEE International Conference on Computer Vision*, pp. 2722–2730, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Roman Filipovych and Christos Davatzikos. Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI). *NeuroImage*, 55(3):1109–1119, 2011.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *European Conference on Computer Vision*, pp. 19–34, 2018.

Kede Ma, Zhengfang Duanmu, Zhou Wang, Qingbo Wu, Wentao Liu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision*, pp. 181–196, 2018.

George A Miller. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? *arXiv preprint arXiv:1902.10811*, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

Thomas L Saaty and Luis G Vargas. Inconsistency and rank preservation. *Journal of Mathematical Psychology*, 28(2):205–214, 1984.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2013.

Mingxing Tan and Quoc V Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

Zhangyang Wang, Nasser M Nasrabadi, and Thomas S Huang. Semisupervised hyperspectral classification using task-driven dictionary learning with Laplacian regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1161–1173, 2014.

Zhou Wang and Eero P Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):1–13, 2008.

Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 842–850, 2015.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018.

5 APPENDIX

Figure 6: Visual comparison of "broccoli" and "soccer ball" images selected by MAD (a) and in ImageNet validation set (b).

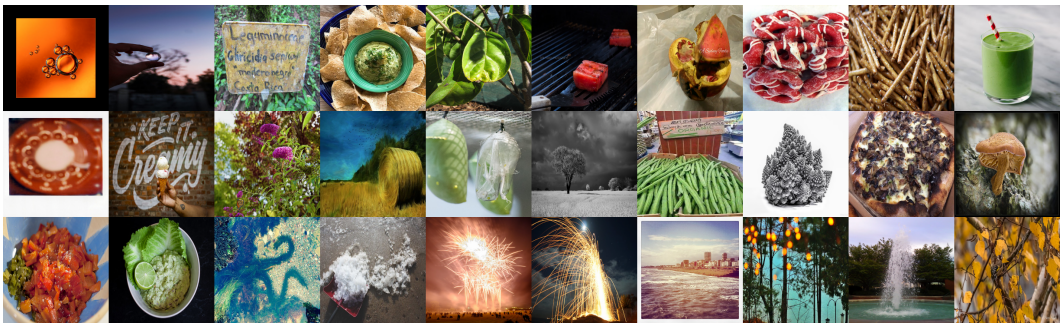

Figure 7: Top-30 images selected by MAD when comparing ResNet101 and EfficientNet-B7.

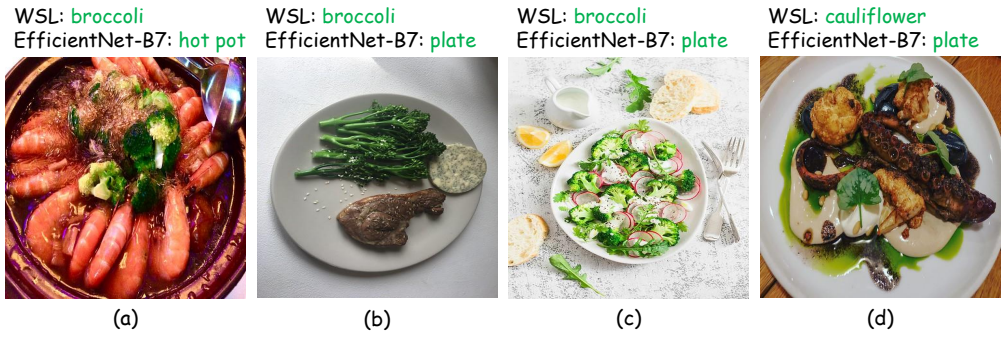

Figure 8: Examples of network bias. WSL-ResNeXt101-32×48 (WSL) tend to focus on foreground objects, while EfficientNet-B7 attends more to background objects.

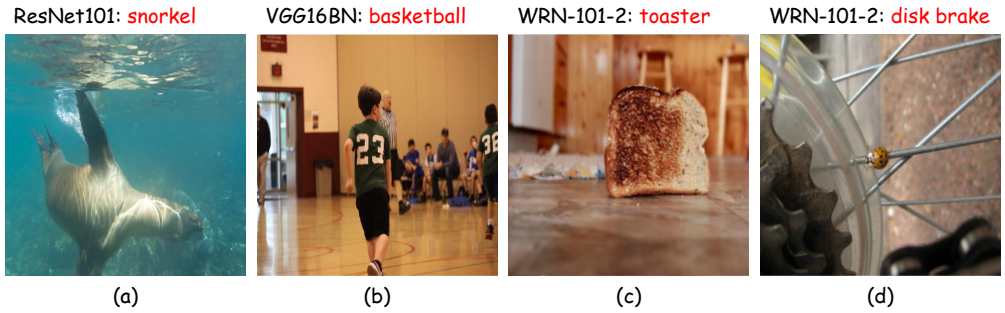

Figure 9: Examples of relation inference. **(a)**: Snorkel is correlated to underwater environment. **(b)**: Basketball is correlated to basketball court. **(c)**: Toaster is correlated to toasted bread. **(d)**: Disk brake is correlated to freewheel and spokes. Similar with how humans recognize objects, it would be reasonable for DNN-based classifiers to make predictions by inferring useful information from object relationships, only when their prediction confidence is low. However, this is not the case in our experiments, which show that classifiers may make high-confidence predictions by leveraging object relations only without really "seeing" the predicted object.

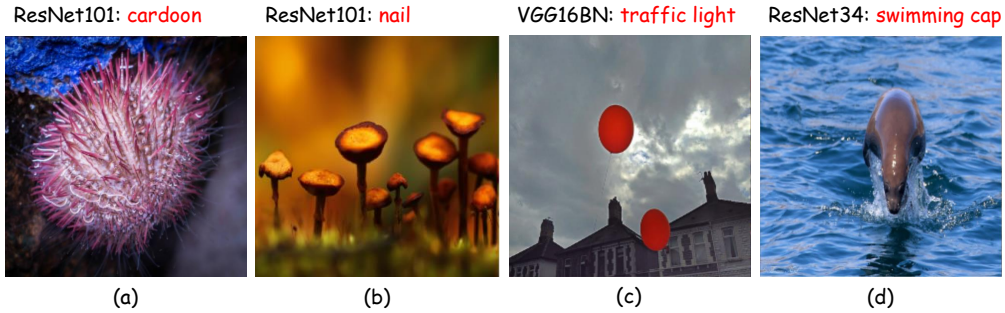

Figure 10: Examples of network bias to low-level visual features, such as color, shape and texture, while overlooking conflicting semantic cues. An ideal classifier is expected to utilize both low-level (appearance) and high-level (semantic) features when making predictions.