LACONIC IMAGE CLASSIFICATION:
HUMAN VS. MACHINE PERFORMANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose *laconic classification* as a novel way to understand and compare the performance of diverse image classifiers. The goal in this setting is to minimise the amount of information (aka. *entropy*) required in individual test images to maintain correct classification. Given a classifier and a test image, we compute an approximate minimal-entropy positive image for which the classifier provides a correct classification, becoming incorrect upon any further reduction. The notion of entropy offers a unifying metric that allows to combine and compare the effects of various types of reductions (e.g., crop, colour reduction, resolution reduction) on classification performance, in turn generalising similar methods explored in previous works. Proposing two complementary frameworks for computing the minimal-entropy positive images of both human and machine classifiers, in experiments over the ILSVRC test-set, we find that machine classifiers are more sensitive entropy-wise to reduced resolution (versus cropping or reduced colour for machines, as well as reduced resolution for humans), supporting recent results suggesting a texture bias in the ILSVRC-trained models used. We also find, in the evaluated setting, that humans classify the minimal-entropy positive images of machine models with higher precision than machines classify those of humans.

1 INTRODUCTION

Deep neural networks now surpass human-level performance on a variety of specific tasks and metrics relating to visual recognition. In a widely-used yardstick for human-level performance, Russakovsky et al. (2015) estimated that an expert human with prior training can achieve a top-5 classification error rate of 5.1% on a dataset of 1,500 ILSVRC images and 1,000 target classes. Shortly after, He et al. (2015) surpassed human-level performance on the same task achieving 4.9% top-5 error with PReLU-net. Later works would further reduce this error rate, including ResNet (3.6%) (He et al., 2016), Trimps-Soushen (3.0%) (Shao et al., 2016), SeNet (2.3%) (Hu et al., 2019), etc., with contemporary state-of-the-art models more than halving estimated human error for this specific task.

Though such results represent landmark advances, by focusing on classification errors alone, they do not reveal the full story of relative machine performance for image classification. Works on adversarial examples (Dalvi et al., 2004; Szegedy et al., 2014; Nguyen et al., 2015), for instance, establish that human and machine perception diverges greatly for specifically constructed images. Other works have presented bespoke experiments comparing human and machine performance beyond classification errors, presenting evidence for a lack of robustness in the presence of noisy (Russakovsky et al., 2015; Dodge & Karam, 2017; Hosseini et al., 2018; Dodge & Karam, 2019) or incomplete information (Ullman et al., 2016; Wick et al., 2016; Linsley et al., 2017; Ho-Phuoc, 2018; Srivastava et al., 2019), a lack of generalisation (Geirhos et al., 2018), a bias towards texture (Geirhos et al., 2019), etc., in evaluated models. By applying specific transformations on test images prior to classification, these latter works provide insights into the specific differences in the types of information that humans and machines rely on to perform adequately at this task.

These latter recent works suggest the need for an information-theoretic framework that generalises such issues: a framework within which the performance of classifiers – be they human, machine or other – can be compared and understood, allowing to quantify, in a more fine-grained manner, the type of information in the input on which a given classifier depends. While previous works address

individual or multiple types of information reduction on input images in isolation, a more general framework should allow to combine and compare different types of reduction on test inputs.

In this paper, we propose such a framework, based on the principle of computing and analysing *minimal entropy positive inputs*: inputs with minimal information with respect to yielding correct classification results. The notion of entropy intuitively generalises and allows to compare the relative effects of different reductions on inputs – and their combinations – on classifier performance; such reductions may include, for example, cropping, downsampling and quantisation. The goal in this framework thus shifts from precise classification to *laconic classification*: providing a classifier that minimises the entropy of input(s) required for correct classification. Being based on a continuous notion of entropy rather than a discrete notion of correct/incorrect, this latter goal thus presents a novel challenge beyond minimising classification error—a goal for which state-of-the-art approaches are close to achieving perfect results on datasets like ILSVRC. In fact, existing datasets – such as ILSRVC – can be straightforwardly leveraged for evaluating classifiers under this new goal. Models performing more laconic classification (we currently conjecture) should likewise perform more robustly in practical settings involving incomplete or noisy information capture.

Though the framework we propose can be applied to any classifier for any classification task, herein we first instantiate the framework on the aforementioned problem of image classification. We consider three general operations for reducing the entropy of test images: crop, resolution reduction (downsampling) and colour reduction (quantisation). We then propose two methods for finding the minimal entropy positive images under these reductions for two different types of classifier.

Given a pre-trained machine classifier – where classification can be separated from learning – an input test image, and a set of reduction operations, we use a known search algorithm (Powell, 1964) and apply the given reduction functions to the input image to find (under certain assumptions) the lowest entropy image that the model classifies correctly such that applying any further reduction of entropy leads to incorrect classification. We apply the aforementioned framework to find the minimal entropy images from a sample ILSVRC test-set for state-of-the-art deep neural-network (DNN) models (GoogLeNet, SqueezeNet, ResNet50 and SeNetResNet50), with respect to the three aforementioned reduction operations and their combination. Our experiments show that minimal entropy images are considerably smaller than original images for DNNs; for example, we find that with only 2–6% of the information content of the original test-images (on average) the best performing machine model can still produce a correct classification of the reduced image.

We then consider also *human classifiers* in order to compare their ability to perform laconic classification with DNN models; applying our framework for humans was not trivial since learning cannot be separated from classification (we cannot start with the full input test-image and reduce it since the human will remember the image) and an automatic optimization algorithm was not possible. To cope with this we designed an experiment *reversing* the optimization goal: starting from a void image, the human evaluator may add information incrementally until they believe that a confident classification of the displayed image is possible. We apply this framework with more than 500 human subjects through an online interface. Our results show that the minimal-entropy positive images for humans are considerably smaller in the case of colour and resolution reductions (53–62% the size of the corresponding size for the best performing machine model).

Finally we cross-classified the minimal-entropy positive images among DNNs and humans: given classifier A and B, we presented A's minimal-entropy positive images to B and vice versa, computing classification precision. We show that the precision of human classifiers on DNN minimal-entropy positive images was considerably better (0.74 precision in the worst case) than the corresponding results for cross-classification among DNN models (0.02 precision in the worst case for the best model). Furthermore, we find that the DNN models on human images give precisions of 0.03–0.43, depending on the reductions used for the images and the model used for classification.

Our results provide insights into how different classifiers – both machine and human – rely differently on different types of information, providing further evidence to support, for example, prior claims of the lack of robustness to incomplete information relative to humans in such models (Dodge & Karam, 2017), a bias towards texture in ILSRVC-trained models (Geirhos et al., 2019), and so forth. As a more general conclusion, we show that humans are capable of performing (much) more laconic classification in the evaluated setting than state-of-the-art machine classifiers. We conclude with open challenges regarding the laconic classification of images using DNNs.

2 RELATED WORKS

A number of previous works have studied the robustness of image classification for DNNs and humans. Ullman et al. (2016) introduced the notion of a "minimal image" as the smallest region of an image that is still recognisable by a human. They show that a small change in these minimal images can have a drastic effect on recognition. Moreover, Ullman et al. (2016) show that DNNs are unable to accurately recognise human minimal images. To compute the minimal images, they started from the complete image and then iteratively cropped it showing the new cropped image to a different human subject every time. The process stopped when the accuracy of recognizing every crop of the current patch dropped from a threshold. Given the difficulty of computing minimal regions recognisable by humans, their experiments were limited to 10 images (Ullman et al., 2016).

Srivastava et al. (2019) extended the previous work by showing that the sharp drop in accuracy for minimal images can also be observed in DNNs. They defined the notion of "fragile recognition image" as a region of an image for which a small change in size produces a considerable change in accuracy of recognition by DNNs. Srivastava et al. (2019) showed that fragile recognition images are abundant and can occur for several different sizes. The work by Ullman et al. (2016) and Srivastava et al. (2019) can be modelled inside our minimal-entropy framework, which considers not only regions of images (i.e., crop), but also reductions in resolution, colour, and combinations thereof.

Another line of related works study the robustness of DNN image classification with respect to distortions to the input image (Hendrycks & Dietterich, 2019; Geirhos et al., 2018; 2019; Dodge & Karam, 2017; 2019). Dodge & Karam (2017; 2019) also study how humans react to these distortions showing that the performance of DNNs is much lower than human performance on distorted images. Hendrycks & Dietterich (2019) constructed a benchmark for image classifier robustness based on several different image distortions. Although all these distortions are designed to make the image recognition task more difficult, not all of them reduce the information in the image. For example, two transformations considered by Geirhos et al. (2018) are image rotation and colour inversion that do not alter the image file size. Other transformations considered by Geirhos et al. (2018) – such as Eidolon and additive noise – may even increase the image file size. In contrast we only consider transformations that actually reduce the information content of the input image file.

3 APPROXIMATING MINIMAL-ENTROPY POSITIVE IMAGES

We now discuss the framework we use for evaluating the laconic classification of images. We first discuss the three forms of entropy reduction applied to images in this work. We then discuss the computation of approximate minimal-entropy positive images (MEPIs) for DNNs and for humans.

3.1 ENTROPY REDUCTION

Given an $m \times n$ matrix $\mathbf{A}$, we consider three forms of entropy reduction based on computing:

1. A (nearest-element) $m \times n$ quantised matrix $\mathbf{A}'$ based on mapping the elements of $\mathbf{A}$ to their nearest value in a fixed set of quantisation values V; more specifically, given the quantisation value κ, we define $V = \{1, \ldots, \kappa\}$, normalise the values of $\mathbf{A}$ according to $\frac{\kappa}{\max(\mathbf{A})}\mathbf{A}$, and map each element a_{ij} of $\mathbf{A}$ to $a'_{ij} := \arg\min_{v \in V}(v - a_{ij})^2$ in $\mathbf{A}'$.

2. A $r \times s$ downsampled matrix $\mathbf{A}'$ ($r \leq m$, $s \leq n$) computed based on Guassian pooling; more specifically, given parameter σ ($0 < \sigma \leq 1$, $\lfloor \sigma m \rfloor = r$, $\lfloor \sigma n \rfloor = s$), we apply a downscaling of $\mathbf{A}$, configuring the pooling to generate the scaling factor σ.

3. A (contiguous) $p \times q$ submatrix $\mathbf{A}'$ ($p \leq m$, $q \leq n$) by deleting rows and columns of $\mathbf{A}$; more specifically, given parameters $\alpha, \beta, \gamma, \delta$ ($\alpha < k$, $\beta = \alpha + p$, $\gamma < n$ and $\delta = \gamma + q$), we delete rows $1, \ldots, \alpha - 1$ and $\beta, \ldots, m$ and columns $1, \ldots, \gamma - 1$ and $\delta, \ldots, n$ from $\mathbf{A}$, resulting in $\mathbf{A}' = \mathbf{A}[\alpha, \alpha + 1, ..., \beta - 1; \gamma, \gamma + 1, \ldots, \delta - 1]$.

4. A combination of the above.

In the case of images, these reductions correspond, respectively, to colour reduction (parameter: κ), resolution reduction (parameter: σ), crop (parameters: $\alpha, \beta, \gamma, \delta$), and their combination (parameters: $\alpha, \beta, \gamma, \delta, \kappa, \sigma$). We thus have six parameters by which to reduce entropy. We adapt these

reductions to multi-channel images in the natural way. In the case of crop, we apply the reduction to all channels separately; however, based on initial experiments with DNNs, rather than remove the rows and columns of the image's channels, we rather replace them with a constant neutral value, which allowed further entropy reduction in positive images by avoiding distortions once images are internally rescaled by the network (i.e., images with narrow crops being "stretched out"). In the case of colour reduction, the nearest quantisation values are computed in the multi-channel case based on Euclidean distance; we further normalise the output values to fill the colour space after the quantisation, choosing equidistant points. In the case of downsampling, the reduction is applied to each channel; by definition, the downsampled image maintains almost the same aspect ratio as the input image (and exactly the same ratio in the case of square images).

Given an image, we define an *atomic reduction step* as a minimal change ϵ to a parameter resulting in a distinct (smaller) matrix. In the case of $\alpha, \beta, \gamma, \delta$ and κ, the step size is one ($\epsilon = 1$). In the case of σ, though for convenience we define it as a real value between 0 and 1, we can consider discrete steps where ϵ is the minimum value such that $\lfloor \sigma m \rfloor \neq \lfloor (\sigma - \epsilon)m \rfloor$ or $\lfloor \sigma n \rfloor \neq \lfloor (\sigma - \epsilon)n \rfloor$.

3.2 ENTROPY MEASURE

We are left to find the minimal-entropy positive images (MEPIs) for various classifiers based on these forms of entropy reduction. Although there exist various proposed measures (Wu et al., 2013; Larkin, 2016), the analysis of entropy for images is not straightforward due to their dimensionality, particularly in the case of multi-channel images where measures of entropy are further complicated by the conditional and joint entropy that must be measured across the channels. We mitigate this issue by adopting the compressed size of an image using a lossless image encoder as an estimation of image entropy (Larkin, 2016). For these purposes we propose to use the Portable Network Graphics (PNG) encoder – proposed as a successor for the Graphics Interchange Format (GIF) – which combines an initial predictive filtering step to improve compressibility, followed by a DEFLATE compression. The size of PNG images has been shown to correlate with more complex measures of entropy (Larkin, 2016), while being widely implemented in a variety of libraries.

3.3 MINIMAL-ENTROPY POSITIVE IMAGES

Given an input image, a set of entropy reductions, and a classifier, we define a *naive minimal-entropy positive image* to be the smallest (PNG) image derived from the input image using the given reductions for which the classifier provides a *correct* classification (we consider a classification to be correct if and only if the top predicted class is correct). However, this definition may lead to undesirable results: consider, for example, the case of a classifier making predictions based on the individual (nondescript) pixels of the input image. With varying probability – depending on the number of pixels, images, and classes, the form of classification, etc. – a classifier may for some pixel "guess" the correct class, with that pixel potentially becoming a naive minimal-entropy positive image. To avoid such cases, we add a continuity condition: a *minimal-entropy positive image* (*MEPI*) is then defined to be the smallest image derived from the input image using the given reductions for which the classifier provides a correct classification, and for which there is a continuous path of atomic reduction steps from the input image giving correct classifications.

3.4 APPROXIMATING MEPIS FOR DNNS

In the case of DNNs, for a given image and applicable reductions, we apply a search for the value(s) of the parameter(s) that correspond to the MEPI of the image. Our general method begins with the input image, incrementally applying atomic reduction steps while the prediction of the classifier remains correct, backtracking in the case of an incorrect prediction. While relatively simple to implement for reductions with a single parameter, in the case of crop or combined reductions, we have multiple parameters with respect to which we must search; for example, in the case of crop, for an $m \times n$ image, the number of images to check in the worst-case is potentially $\binom{m}{2}\binom{n}{2}$, i.e., more than 274 billion contiguous sub-images for a 1024×1024 input image (and this considering crop alone without combination with resolution or colour). Approximation is thus sought.

First, our assumption of continuity helps us to prune the search space: if we reach a set of parameter values for which classification is correct but for which any further atomic reduction is incorrect, we

know we can rule out all further reductions from that point. We may still, however, encounter an infeasible search space when considering multiple parameters; to improve performance, we choose to apply a greedy search algorithm: given that the function we wish to minimise is not differentiable but has a fixed number of inputs, we apply Powell's method (Powell, 1964) to find a local minimum.

3.5 Approximating MEPIs for Humans

Humans operate differently to DNNs and thus require specialised methods to approximate their MEPIs. First, humans cannot separate classification from learning; hence we cannot show the full input image and apply reductions as the human will (of course) remember the full input image. Second, humans require (much) more time per classification; hence the search steps must be adjusted to quickly reach the MEPI. Hence we rather compute a bottom-up approximation of MEPIs for humans starting from a void image; at each step the human can then select an atomic step to *increment* the entropy of the image, or can select a class for the image from a list if they believe that they are able to correctly classify the image. If the human guesses correctly, this is considered to be an approximate MEPI (for the image, parameters and human); if not, the image is discarded and a fresh image is considered. To mitigate fatigue, larger atomic steps are defined such that no more than 20 are required to reach the input image on a given dimension; furthermore, in the case of crop, starting from the central pixel, we offer the option to increment in all directions, while in the combined case, we offer the option to increment all parameters. This method makes it feasible to approximate MEPIs for humans, but it is important to note that with the limitations of coarser steps and always starting at the central pixel, the MEPIs approximated using this method are more likely to overestimate the entropy required for correct classification versus the method used for DNNs.

4 Experimental Setting

We base our experiments on images from the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) (Russakovsky et al., 2015). We select four DNN classifiers for evaluation (all trained on the ILSVRC2012 training set): SqueezeNet (Iandola et al., 2016), GoogLeNet (Szegedy et al., 2015), ResNet50 (He et al., 2016) and SeNetResNet50 (Hu et al., 2019); as discussed in the introduction, the latter two classifiers surpass human-level performance in terms of top-5 classification error on a dataset of 1,500 ILSVRC images and 1,000 target classes.

As acknowledged by Russakovsky et al. (2015), the 1,000 detailed classes of ILSVRC (e.g., COUCAL, SEALYHAM TERRIER) require extensive previous training for humans to perform adequately at classification, not only to visually distinguish the objects, but also to be able to retrieve their label from the 1,000 possible labels. We thus opt to greatly simplify the classes, selecting the following twenty: BEAR, BIRD, CAT, DOG, FISH, FLOWER, FOX, FRUIT, FUNGUS, HIPPOPOTAMUS, INSECT, LION, MONKEY, REPTILE, SHARK, SPIDER, TIGER, VEGETABLE, VEHICLE, WOLF. We select these classes as they should be generally recognisable to humans without prior training; furthermore, we select mostly plants and animals to provide a more challenging classification task, with visually similar classes, such as LION/TIGER, FRUIT/VEGETABLE, DOG/WOLF, INSECT/SPIDER, etc., providing non-trivial cases to visually distinguish. The results of DNN models are then mapped hierarchically to these higher-level classes (e.g., COUCAL ↦ BIRD, SEALYHAM TERRIER ↦ DOG). We sample 15 images for each of the 20 classes from the ILSVRC2012 test set for experiments.

5 Experimental Results

With our experimental results, we address two main questions:

1. How does the entropy required by DNN and human classifiers compare?

2. How do the classifiers perform in terms of precision for each others' MEPIs?

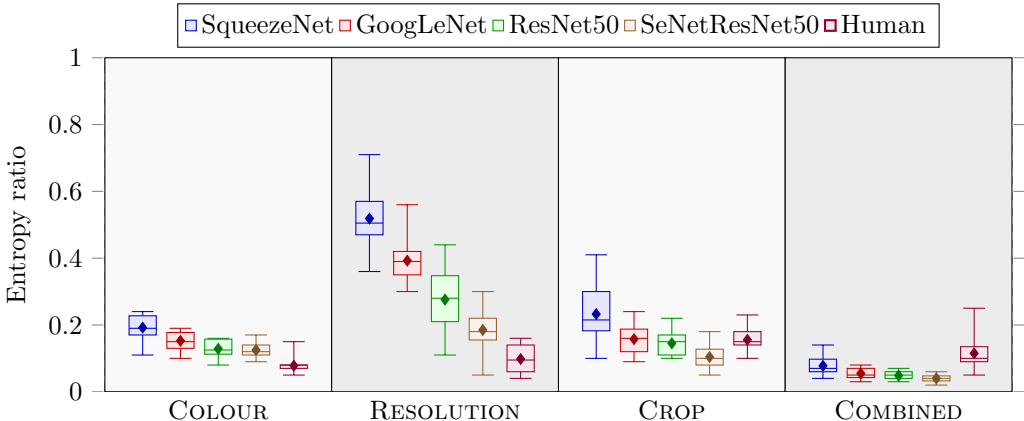We address these two questions in the following subsections.

Figure 1: Box-plots of mean entropy ratio for DNN and human MEPIs across classes

5.1 ENTROPY RATIO IN MEPIS

For the $20 \times 15 = 300$ test images, we first compute the top-down MEPIs for the four DNN classifiers using the method described in Section 3.4. For humans, following the bottom-up method described in Section 3.5, we provide a web interface that – starting with a void image – allows a human user in each step to either increment the image by the given parameters, or select the class for the currently displayed image. In order to achieve many responses, the interface was shared on a university forum as well as on social media. The options and instructions were presented in Spanish, corresponding to the native language of the country in which the university is based. Screenshots of the interface are given in Appendix A. In total, 423 user sessions were logged; of the 1,722 responses obtained (average 4.07 per session), 1,340 (77.8%) resulted in a correct validation and thus a valid MEPI.

In Figure 1 we present the *entropy ratio* for four different settings across five different classifiers; entropy ratio is defined here as the ratio of the (PNG-encoded) size of the original input image versus the size of the extracted MEPI. We take the average entropy ratio for the images of each of the twenty classes. Figure 1 then presents the box-plots – displaying the 1st (min), 25th, 50th (median), 75th, and 100th (max) percentiles with mean marked as a diamond – for the mean ratio across the different classes; to illustrate with an example, in the case of SqueezeNet, considering only the COLOUR experiment, the best class had a mean entropy ratio of 0.11 (bottom whisker), the worst class had a mean entropy ratio of 0.24 (top whisker), etc. We present the DNN models in order of their reported performance for top-5 classification error on the ILSVRC dataset, with SqueezeNet having the highest such error and SeNetResNet50 having the lowest.

From this figure we can draw some high-level observations about the DNN classifiers. First, for the DNN classifiers, the entropy ratio is lowest for the COMBINED experiment (as expected), which offers more avenues by which to reduce the entropy while maintaining a correct classification. Second, the results for DNN classifiers follow the same trend as for performance based on classification error; this suggests that there is a correlation between the goals of laconic classification and precise classification. Third, we see that DNNs are most sensitive to reductions in resolution, which supports the hypothesis that DNNs trained on ILSVRC images are biased towards texture (Geirhos et al., 2019) (with RESOLUTION being the parameter that most affects the ability to distinguish texture). We also see a considerable variation across classes, particularly for RESOLUTION; for space reasons, we provide the detailed results of mean entropy ratio by class in Appendix B.

With respect to human classifiers, we note that they are less sensitive to reductions in COLOUR and much less sensitive to reductions in RESOLUTION than all DNNs, and more sensitive to reductions in CROP than some state-of-the-art DNNs. We also see an unusual result, whereby the reduction ratios for COMBINED are not lower than those for (e.g.,) COLOUR; this suggests that the users may have struggled with the interface for the CROP and COMBINED experiments, where multiple options were provided to increment the entropy of the image (versus COLOUR and RESOLUTION, which only permitted increments in one direction). As such, the results for CROP and COMBINED in human classifiers leave an ambiguity: are the relatively poor results of humans due in this case

	SN	GN	RN	SRN	Hum
SN	1.000	0.414	0.344	0.382	0.136
GN	0.903	1.000	0.632	0.640	0.255
RN	0.905	0.880	1.000	0.771	0.432
SRN	0.919	0.888	0.765	1.000	0.424
Hum	0.923	0.911	0.891	0.914	1.000

(a) COLOUR

	SN	GN	RN	SRN	Hum
SN	1.000	0.224	0.158	0.084	0.039
GN	0.760	1.000	0.275	0.121	0.030
RN	0.888	0.806	1.000	0.314	0.152
SRN	0.950	0.933	0.840	1.000	0.285
Hum	0.929	0.907	0.918	0.864	1.000

(b) RESOLUTION

	SN	GN	RN	SRN	Hum
SN	1.000	0.040	0.080	0.019	0.112
GN	0.597	1.000	0.209	0.041	0.221
RN	0.530	0.214	1.000	0.023	0.226
SRN	0.783	0.572	0.637	1.000	0.424
Hum	0.867	0.816	0.852	0.760	1.000

(c) CROP

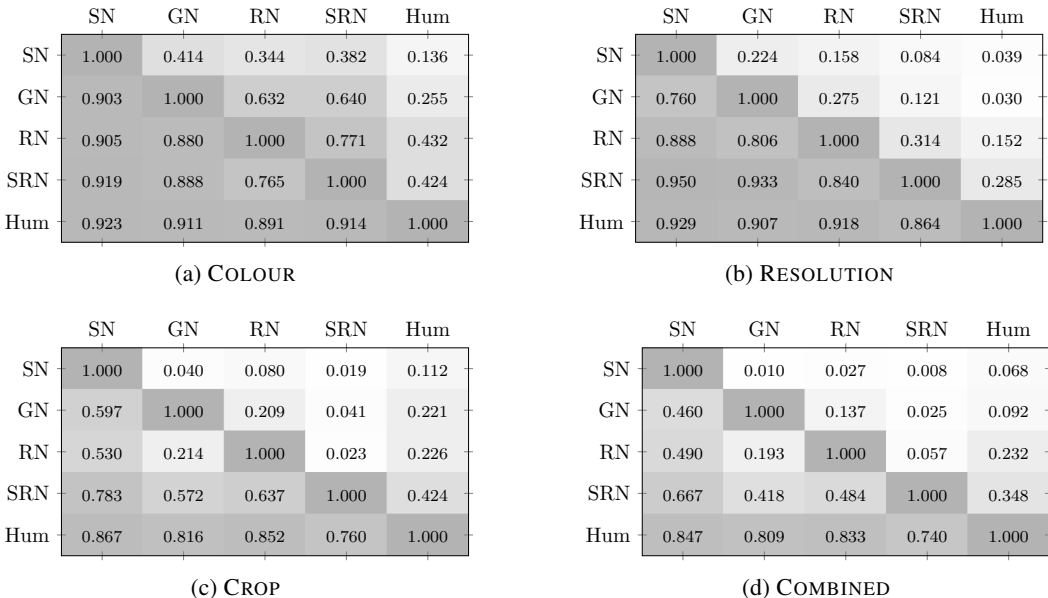|	SN	GN	RN	SRN	Hum
SN	1.000	0.010	0.027	0.008	0.068
GN	0.460	1.000	0.137	0.025	0.092
RN	0.490	0.193	1.000	0.057	0.232
SRN	0.667	0.418	0.484	1.000	0.348
Hum	0.847	0.809	0.833	0.740	1.000

(d) COMBINED

Figure 2: Heatmaps of classification precision on MEPIs for the four types of entropy reduction; rows indicate the classifier, columns indicate the MEPIs classified: SN = SqueezeNet; GN = GoogLeNet; RN = ResNet50; SRN = SeResNet50; Hum = Human

to greater sensitivity to such (multi-parametric) reductions, or because of difficulty using the more complex interface in these cases? We require further experiments to address this ambiguity.

5.2 CLASSIFICATION PRECISION FOR MEPIS OF DIFFERENT CLASSIFIERS

We now look at the precision of (top-1) classification across the MEPIs of models, again considering COLOUR, RESOLUTION, CROP and COMBINED. To gather results for human classification, we created a second, simpler, online interface that presents the MEPI of a particular model under a particular reduction and asks the human evaluator to select the class for that MEPI from the list of twenty possible classes. The interface was shared again on a university forum and through social media. We provide a screenshot of this interface in Appendix A. We first ran a control group with 25 trusted users, which logged an aggregate precision of 0.875 with a standard deviation of 0.061; in the open/online evaluation, we then filter user sessions more than two standard deviations from the control mean precision, giving a lower threshold of 0.753. The public evaluation then logged 531 valid user sessions according to the threshold, resulting in 11,588 valid classifications of machine MEPIs (equating to 26.2 classifications on average per session).

The results of the precision for cross-classification of MEPIs are summarised in Figure 2. Cells are shaded darker to indicate better performance. Darker rows indicate better classifier performance while darker columns indicate having MEPIs that are easier for other models to classify. First we observe that all classifiers correctly predict (as expected by definition) all of their own MEPIs.

With DNNs ordered by expected performance, we again see the general trend that fewer reported classification errors in ILSVRC again correlate with better precision in the classification of MEPIs, with, for example, SeNetResNet50 (SRN) having darker rows and lighter columns than other DNNs. We also see good cross-classifier performance for COLOUR and RESOLUTION: given that this are one dimensional reductions, the space of possible images is greatly reduced. On the other hand, DNN models struggle in cross-classification of the MEPIs under CROP and COMBINED: this is perhaps due to the larger search space for these reductions, but also indicates that these MEPIs for DNNs are characteristic of that particular DNN. In general, the MEPIs of SRN are notably more difficult for other models to classify (as could have been expected from Figure 1).

Regarding human classification, we see that in general humans perform best of all classifiers in this experiment, being adaptable enough to classify the MEPIs under all reductions with relatively high

precision. The most difficult MEPIs for human are in the CROP and COMBINED configurations for SeResNet50 (SRN), where a relatively high precision of 74–76% correct classifications is still seen. With respect to the previously discussed ambiguity in the results of the previous sub-section, we can thus see that although CROP and COMBINED are more difficult cases for humans, issues with the more complex interface for these cases may explain the relatively poor results of humans in Figure 1 for CROP and COMBINED. Furthermore, we can see that DNNs often struggle to classify the human MEPIs, where the best precision reached was 42% correct classifications for SRN in he case of CROP; Appendix C provides a sample of three human MEPIs not correctly classified by any DNN model, illustrating some of the most difficult such cases for DNN models.

6 DISCUSSION

Based on our proposed frameworks for computing minimal-entropy positive images (MEPIs) for both machine and human classifiers, our first results were based on the relative entropy of such images for the different classifiers. From these results we can conclude that state-of-the-art machine models are relatively sensitive (entropy-wise) to reductions in resolution when compared with other forms of reduction based on cropping or colour; such observations appear to independently support previous results indicating a bias in texture for images trained on the ILSVRC dataset (Geirhos et al., 2019). On the other hand, humans are relatively sensitive to the cropping of images, but are much less sensitive than machine classifiers to reductions in resolution; this tends to suggest that humans rely more on form and context rather than textures of small regions of the image. In our second set of results, we performed cross-classification of the MEPIs for different models; we saw that humans greatly outperform machine models for classifying the MEPIs of other models, suggesting a generally more robust aptitude in humans for the laconic classification of images.

Our work has a number of limitations that could be addressed for future work. While the method and interface used for computing human MEPIs in a bottom-up fashion work well for simpler (i.e., one-dimensional) forms of image enhancement, the unexpected result in Figure 1 showing an increase in the size of MEPIs in the combined case suggest that the users had some difficulty using the interface for multi-dimensional settings; given that we filter incorrect images from consideration, we speculate that the users often "overshoot" the MEPI in particular dimensions, finding it difficult to select the particular dimension that is most likely to help them correctly classify the image. Refinements of this idea would be interesting to explore in future. Furthermore, in order to facilitate non-expert humans to participate, we selected high-level classes, where it would be of interest to try to repeat the results for lower-level classes (though again presenting challenges for the interface).

On the other hand, we have performed experiments for pre-trained, off-the-shelf DNN models. An interesting line of research would be to rather train models specifically for the task of laconic classification. Along these lines, one could consider computing MEPIs in the training set and feeding them (potentially recursively) back into the model; unlike the internal transformations applied in specific DNN models, such a framework for laconic learning could treat a particular model as a black-box. It would further be interesting to explore how models trained in such a manner perform in more traditional classification metrics – error rates, precision, etc. – on the original input images, as well as whether or not they might help to address the observed lack of robustness of state-of-the-art DNNs in the presence of noisy (Russakovsky et al., 2015; Dodge & Karam, 2017; Hosseini et al., 2018; Dodge & Karam, 2019) or incomplete information (Ullman et al., 2016; Wick et al., 2016; Linsley et al., 2017; Ho-Phuoc, 2018; Srivastava et al., 2019), or their lack of generalisation (Geirhos et al., 2018), or their bias towards texture (Geirhos et al., 2019). Further exploration is needed.

More generally, we view laconic classification as a potentially fruitful avenue to explore: one that is relatively straightforward to implement, can re-use available datasets, connects a number of recent works, provides insights into the performance and robustness of diverse classifiers, and one that may perhaps reveal new challenges and lead to new developments in image classification.

NOTE: *We plan to publish images and other materials online after double-blind review. Please also see appendices for further details.*

REFERENCES

Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 99–108. ACM, 2004.

Samuel Dodge and Lina Karam. Human and DNN Classification Performance on Images With Quality Distortions: A Comparative Study. *ACM Transactions on Applied Perception (TAP)*, 16 (2):7:1–7:17, 2019.

Samuel F. Dodge and Lina J. Karam. A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions. In *International Conference on Computer Communication and Networks, (ICCCN)*, pp. 1–7. IEEE, 2017.

Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7549–7561, 2018.

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 (ICLR)*. OpenReview.net, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034. IEEE Computer Society, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Tien Ho-Phuoc. CIFAR10 to compare visual recognition performance between deep neural networks and humans. arXiV:1811.07270, 2018.

Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. Assessing Shape Bias Property of Convolutional Neural Networks. arXiV:1803.07739, 2018.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. arXiV:1709.01507v4, 2019.

Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. *CoRR*, abs/1602.07360, 2016. URL `http://arxiv.org/abs/1602.07360`.

Kieran G. Larkin. Reflections on Shannon Information: In search of a natural information-entropy for images. *CoRR*, abs/1609.01117, 2016.

Drew Linsley, Sven Eberhardt, T. Sharma, P. Gupta, and Thomas Serre. What are the Visual Features Underlying Human Versus Machine Vision? In *IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 2706–2714. IEEE Computer Society, 2017.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436. IEEE Computer Society, 2015.

Michael J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, 7(2):155–162, 1964.

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

Jie Shao, Xiaoteng Zhang, Zhengyan Ding, Yixin Zhao, Yanjun Chen, Jianying Zhou, Wenfei Wang, Lin Mei, and Chuanping Hu. Good practices for deep feature fusion. ECCV 2016 Talk, 2016. `http://image-net.org/challenges/talks/2016/Trimps-Soushen@ILSVRC2016.pdf`.

Sanjana Srivastava, Guy Ben-Yosef, and Xavier Boix. Minimal Images in Deep Neural Networks: Fragile Object Recognition in Natural Images. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences (PNAS)*, 113(10):2744–2749, 2016.

Farahnaz Ahmed Wick, Michael L. Wick, and Marc Pomplun. Filling in the details: Perceiving from low fidelity images. arXiV:1604.04125, 2016.

Yue Wu, Yicong Zhou, George Saveriades, Sos S. Agaian, Joseph P. Noonan, and Premkumar Natarajan. Local Shannon entropy measure with statistical tests for image randomness. *Inf. Sci.*, 222:323–342, 2013.

A   SCREENSHOTS OF INTERFACES

We present screenshots of the online interfaces used to collect human MEPIs and classifications. As discussed in the paper, the options and instructions were provided in Spanish as the target audience were students of a university whose native language is Spanish. Figures 3–6 show screenshots of the interfaces used to collect human MEPIs for COLOUR, RESOLUTION, CROP and COMBINED. On the other hand, Figure 7 shows a screenshot of the interface to classify machine MEPIs.
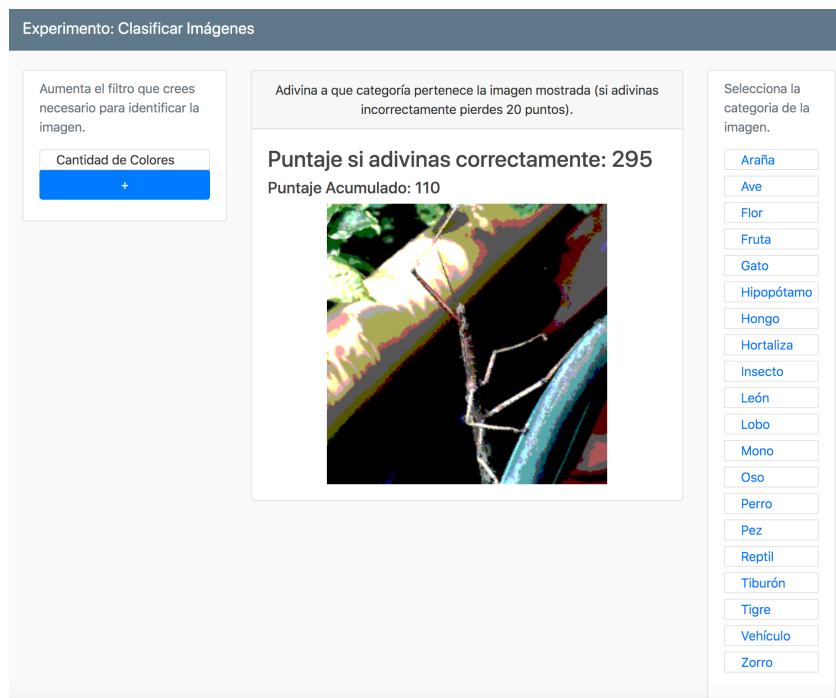

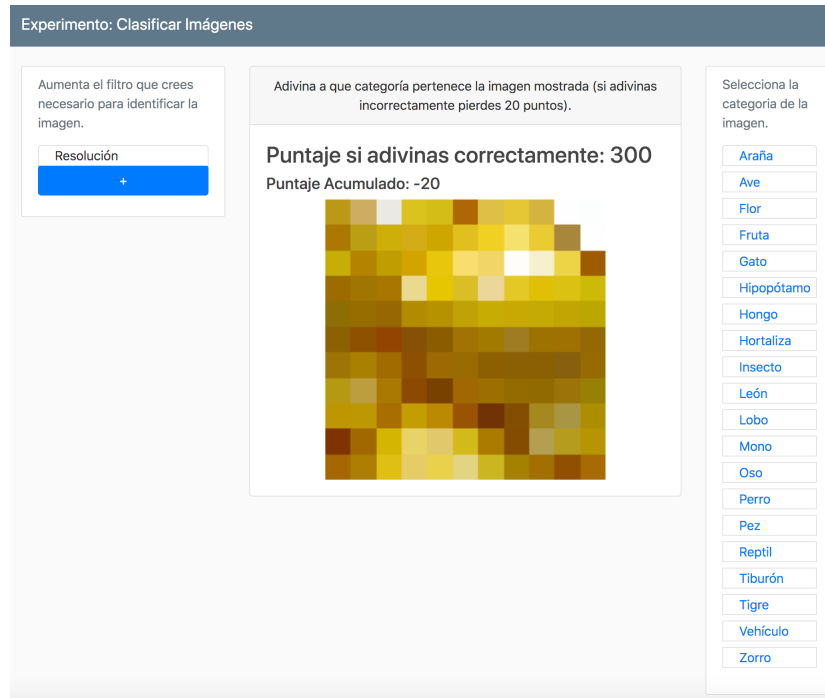Figure 3: Screenshot of interface for humans to specify COLOUR MEPIs

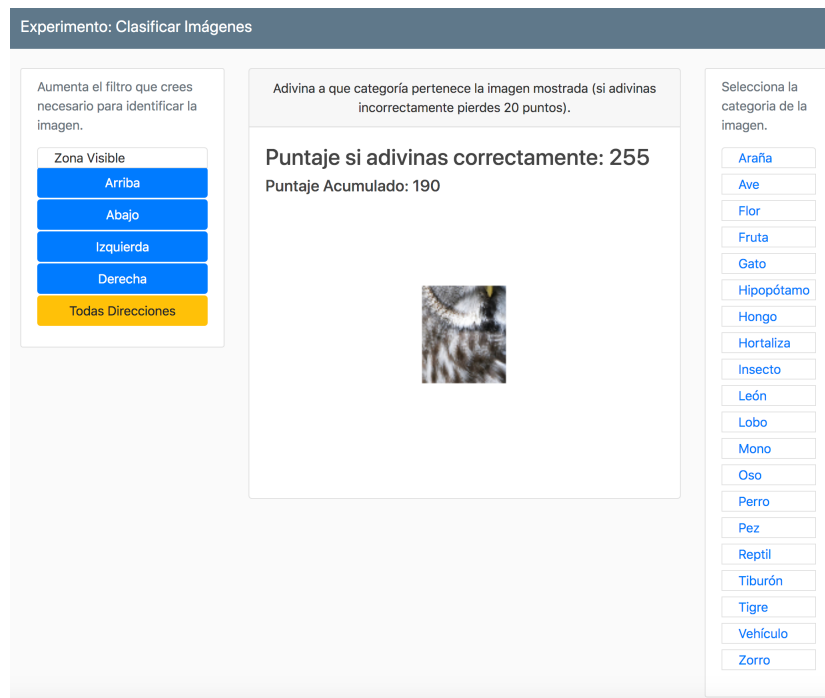Figure 4: Screenshot of interface for humans to specify RESOLUTION MEPIs

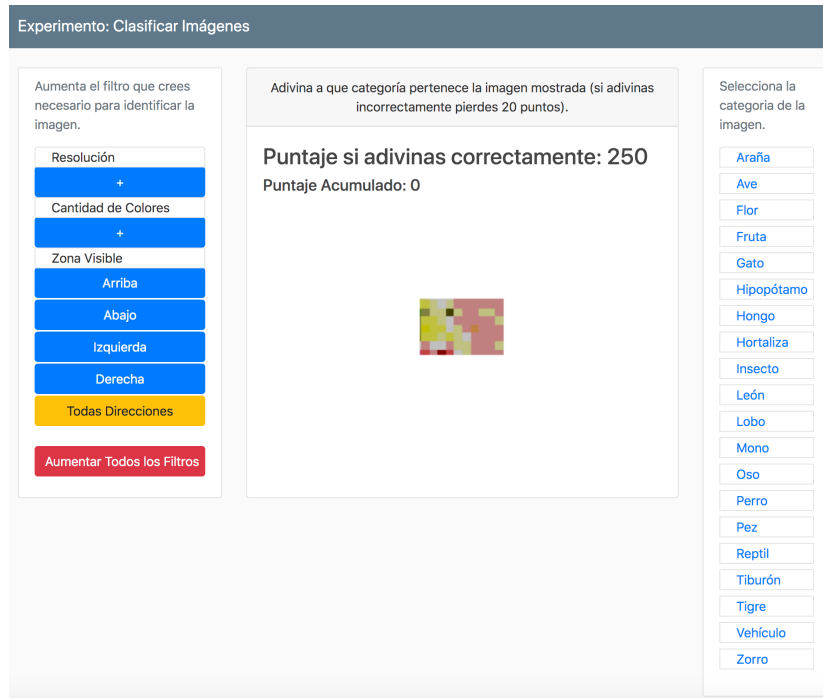Figure 5: Screenshot of interface for humans to specify CROP MEPIs

Figure 6: Screenshot of interface for humans to specify COMBINED MEPIs
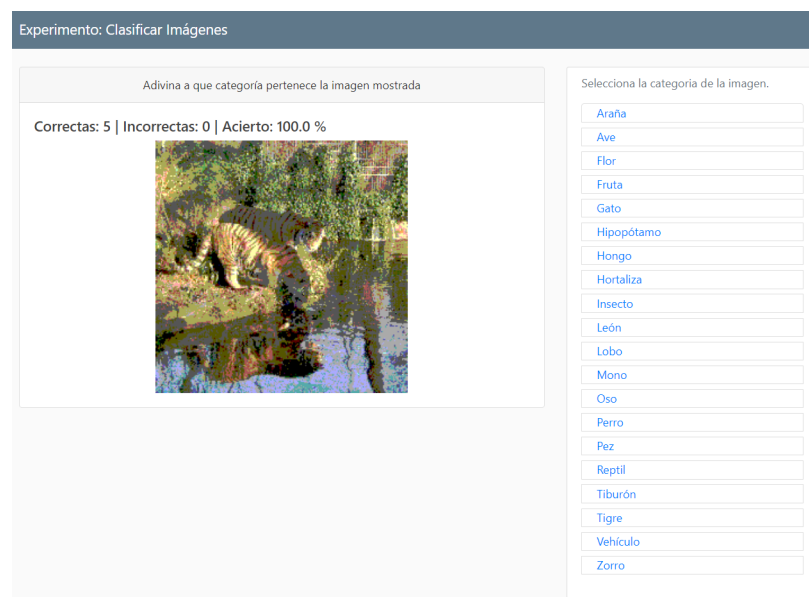

Figure 7: Screenshot of interface for humans to classify machine MEPIs (example using COLOUR)

B   MEAN ENTROPY RATIO PER CLASS

In Table 1 we present the mean entropy ratio of MEPIs per class.

Table 1: Mean entropy ratio per class for DNN and human MEPIs

COLOUR									
	bear	**bird**	**cat**	**dog**	**fish**	**flower**	**fox**	**fruit**	**fungus**	**hippo**
SqueezeNet	0.24	0.16	0.22	0.23	0.17	0.11	0.23	0.15	0.21	0.17
GoogLeNet	0.18	0.14	0.17	0.16	0.15	0.1	0.18	0.13	0.19	0.15
Resnet50	0.16	0.12	0.12	0.12	0.12	0.08	0.16	0.11	0.16	0.13
Se-Resnet50	0.17	0.11	0.12	0.14	0.13	0.09	0.14	0.15	0.14	0.11
Humans	0.1	0.07	0.07	0.06	0.08	0.09	0.08	0.07	0.15	0.08
	insect	**lion**	**monkey**	**reptile**	**shark**	**spider**	**tiger**	**vegetable**	**vehicle**	**wolf**
SqueezeNet	0.15	0.22	0.23	0.19	0.18	0.17	0.23	0.19	0.18	0.21
GoogLeNet	0.12	0.19	0.18	0.15	0.13	0.12	0.17	0.15	0.12	0.17
Resnet50	0.09	0.16	0.15	0.1	0.13	0.09	0.16	0.14	0.15	0.12
Se-Resnet50	0.11	0.12	0.15	0.12	0.12	0.1	0.13	0.14	0.09	0.12
Humans	0.07	0.07	0.08	0.08	0.05	0.06	0.08	0.09	0.08	0.05

RESOLUTION									
	bear	**bird**	**cat**	**dog**	**fish**	**flower**	**fox**	**fruit**	**fungus**	**hippo**
SqueezeNet	0.5	0.51	0.64	0.47	0.57	0.36	0.57	0.43	0.39	0.46
GoogLeNet	0.35	0.37	0.51	0.3	0.42	0.35	0.42	0.38	0.3	0.45
Resnet50	0.28	0.34	0.21	0.11	0.28	0.14	0.39	0.29	0.32	0.4
Se-Resnet50	0.18	0.27	0.17	0.05	0.17	0.11	0.22	0.17	0.19	0.25
Humans	0.07	0.05	0.06	0.06	0.14	0.09	0.16	0.06	0.16	0.11
	insect	**lion**	**monkey**	**reptile**	**shark**	**spider**	**tiger**	**vegetable**	**vehicle**	**wolf**
SqueezeNet	0.5	0.53	0.48	0.71	0.5	0.6	0.55	0.47	0.59	0.54
GoogLeNet	0.37	0.33	0.39	0.56	0.42	0.39	0.4	0.33	0.4	0.41
Resnet50	0.33	0.14	0.25	0.23	0.35	0.21	0.36	0.19	0.44	0.26
Se-Resnet50	0.22	0.13	0.17	0.3	0.2	0.26	0.18	0.13	0.18	0.15
Humans	0.09	0.05	0.11	0.14	0.04	0.14	0.1	0.15	0.07	0.11

CROP									
	bear	**bird**	**cat**	**dog**	**fish**	**flower**	**fox**	**fruit**	**fungus**	**hippo**
SqueezeNet	0.2	0.18	0.16	0.12	0.33	0.19	0.1	0.23	0.31	0.25
GoogLeNet	0.12	0.12	0.12	0.09	0.22	0.16	0.1	0.12	0.22	0.16
Resnet50	0.18	0.12	0.11	0.1	0.16	0.17	0.11	0.11	0.2	0.17
Se-Resnet50	0.09	0.08	0.08	0.06	0.14	0.13	0.08	0.05	0.12	0.13
Humans	0.18	0.16	0.13	0.19	0.23	0.14	0.16	0.14	0.21	0.18
	insect	**lion**	**monkey**	**reptile**	**shark**	**spider**	**tiger**	**vegetable**	**vehicle**	**wolf**
SqueezeNet	0.19	0.26	0.19	0.27	0.41	0.35	0.24	0.2	0.31	0.16
GoogLeNet	0.18	0.18	0.14	0.16	0.24	0.19	0.17	0.16	0.2	0.1
Resnet50	0.17	0.11	0.15	0.13	0.22	0.15	0.15	0.11	0.17	0.11
Se-Resnet50	0.11	0.11	0.1	0.13	0.18	0.12	0.09	0.08	0.1	0.1
Humans	0.18	0.14	0.18	0.16	0.14	0.1	0.14	0.14	0.1	0.11

COMBINED									
	bear	**bird**	**cat**	**dog**	**fish**	**flower**	**fox**	**fruit**	**fungus**	**hippo**
SqueezeNet	0.06	0.07	0.06	0.04	0.09	0.04	0.04	0.07	0.12	0.08
GoogLeNet	0.05	0.04	0.05	0.03	0.08	0.05	0.03	0.05	0.08	0.06
Resnet50	0.05	0.04	0.05	0.04	0.06	0.05	0.04	0.03	0.06	0.06
Se-Resnet50	0.04	0.04	0.04	0.02	0.06	0.04	0.04	0.02	0.05	0.05
Humans	0.12	0.07	0.09	0.05	0.15	0.17	0.09	0.1	0.25	0.11
	insect	**lion**	**monkey**	**reptile**	**shark**	**spider**	**tiger**	**vegetable**	**vehicle**	**wolf**
SqueezeNet	0.06	0.09	0.07	0.11	0.14	0.12	0.1	0.06	0.08	0.05
GoogLeNet	0.04	0.06	0.05	0.07	0.08	0.07	0.07	0.05	0.05	0.04
Resnet50	0.04	0.04	0.06	0.05	0.07	0.05	0.05	0.03	0.06	0.04
Se-Resnet50	0.04	0.04	0.04	0.06	0.06	0.04	0.02	0.03	0.03	0.04
Humans	0.11	0.1	0.1	0.2	0.08	0.09	0.12	0.14	0.07	0.09

C HUMAN MEPIs NOT CLASSIFIED CORRECTLY BY ANY DNN MODEL

Here we show three examples of human MEPIs not correctly classified by any of the four evaluated DNN models: Figure 8 is of a fox, Figure 9 is of a fungus, while Figure 10 is of a vehicle. All three were generated under COMBINED. The gray border is used to highlight the cropped region.
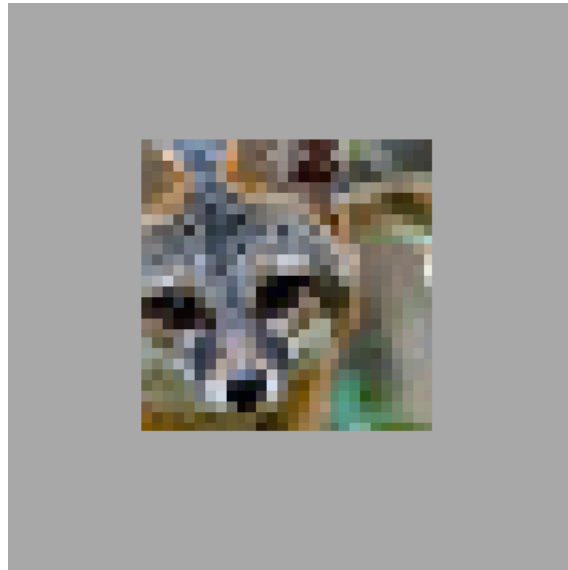

Figure 8: Human MEPI of fox not classified correctly by any DNN model

Figure 9: Human MEPI of fungus not classified correctly by any DNN model

Figure 10: Human MEPI of vehicle not classified correctly by any DNN model