ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases

Pierre Stock$^{(\boxtimes)}$ and Moustapha Cisse

Facebook AI Research, Paris, France
pstock@fb.com, moustaphacisse@google.com

Abstract. ConvNets and ImageNet have driven the recent success of deep learning for image classification. However, the marked slowdown in performance improvement combined with the lack of robustness of neural networks to adversarial examples and their tendency to exhibit undesirable biases question the reliability of these methods. This work investigates these questions from the perspective of the end-user by using human subject studies and explanations. The contribution of this study is threefold. We first experimentally demonstrate that the accuracy and robustness of ConvNets measured on Imagenet are vastly underestimated. Next, we show that explanations can mitigate the impact of misclassified adversarial examples from the perspective of the end-user. We finally introduce a novel tool for uncovering the undesirable biases learned by a model. These contributions also show that explanations are a valuable tool both for improving our understanding of ConvNets' predictions and for designing more reliable models.

Keywords: Bias detection · Interpretability · Adversarial examples

1 Introduction

Convolutional neural networks [1,2] and Imagenet [3] (the dataset and the challenge) have been instrumental to the recent breakthroughs in computer vision. Imagenet has provided ConvNets with the data they needed to demonstrate their superiority compared to the previously used handcrafted features such as Fisher Vectors [4]. In turn, this success has triggered a renewed interest in convolutional approaches. Consequently, novel architectures such as ResNets [5] and DenseNets [6] have been introduced to improve the state of the art performance on Imagenet. The impact of this virtuous circle has permeated all aspects of computer vision and deep learning at large. Indeed, the use of feature extractors pre-trained on Imagenet is now ubiquitous. For example, the state of the art image segmentation [7,8] or pose estimation models [9,10] heavily rely on pre-trained Imagenet features. Besides, convolutional architectures initially
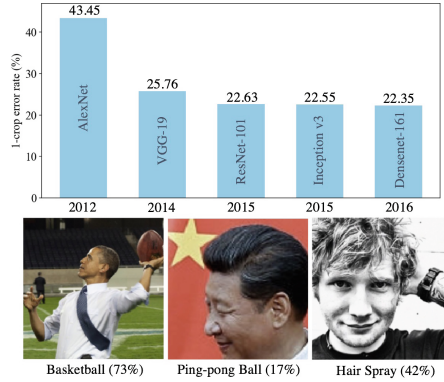
M. Cisse—Now Google AI Ghana Lead.

Fig. 1. Top: performance evolution of various CNN architectures on Imagenet. Bottom: some images sampled from the Internet and misclassified by a ResNet-101.

developed for image classification such as Residual Networks are now routinely used for machine translation [11] and speech recognition [12].

Since 2012, the top-1 error of state of the art (SOTA) models on Imagenet has been reduced from 43.45% to 22.35%. Recently, the evolution of the best performance seems to plateau (see Fig. 1) despite the efforts in designing novel architectures [5,6], introducing new data augmentation schemes [13] and optimization algorithms [14]. Concomitantly, several studies have demonstrated the lack of robustness of deep neural networks to adversarial examples [15–17] and raised questions about their tendency to exhibit (undesirable) biases [18]. Adversarial examples [15] are synthetic images designed to be indistinguishable from natural ones by a human, yet they are capable of fooling the best image classification systems. Undesirable biases are patterns or behaviors learned from the data, they are often highly influential in the decision of the model but are not aligned with the values of the society in which the model operates. Examples of such biases include *racial* and *gender biases* [19]. While accuracy has been the leading factor in the broad adoption of deep learning across the industries, its *sustained improvement* together with other desirable properties such as *robustness to adversarial examples* and *immunity to biases* will be critical in maintaining the trust in the technology. It is therefore essential to improve our understanding of these questions from the perspective of the end-user in the context of Imagenet classification.

In this work, we take a step in this direction by assessing the predictions of SOTA models on the validation set of Imagenet. We show that *human studies* and *explanations* can be valuable tools to perform this task. Human studies yield a new judgment of the quality of the model's predictions from the perspective of the end-user, in addition to the traditional ground truth used to evaluate the model. We also use both *feature-based* and *example-based* explanations. On the one hand, explaining the prediction of a black box classifier by a subset of the features of its input can yield valuable insights into the workings of the

models and underline the essential features in its decision. On the other hand, example-based explanations provide an increased *interpretability* of the model by highlighting instances representative of the distribution of a given category as captured by the model. The particular form of example-based explanation we use is called *model criticism* [20]. It combines both *prototypes* and *criticisms* and is proven to better capture the complex distributions of natural images. Therefore, it facilitates human understanding. Our main findings are summarized below:

– The accuracy of convolutional networks evaluated on Imagenet is *vastly underestimated*. We find that when the mistakes of the model are assessed by human subjects and considered correct when at least four out of five humans agree with the model's prediction, the top-1 error of a ResNet-101 trained on Imagenet and evaluated on the standard validation set decreases from 22.69% to 9.47%. Similarly, the top-5 error decreases from 6.44% to 1.94%. This observation holds across models. It explains the marked slowndown in accuracy improvement and suggests that Imagenet is almost solved.
– The robustness of ConvNets to adversarial examples is also underestimated. In addition, we show that providing explanations helps to mitigate the misclassification of adversarial examples from the perspective of the end-user.
– *Model Criticism* is a valuable tool for detecting the biases in the predictions of the models. Further, adversarial examples can be effectively used for model criticism.

   Similar observations to our first point existed in prior work [21]. However, the scale, the conclusions and the implications of our study are different. Indeed we consider that if top-5 error is the measure of interest, Imagenet is (almost) solved. In the next section, we summarize the related work before presenting our experiments and results in details.

2   Related Work

Adversarial Examples. Deep neural networks can achieve high accuracy on previously unseen examples while being vulnerable to small adversarial perturbations of their inputs [15]. Such perturbed inputs, called adversarial examples, have recently aroused keen interest in the community [15,17,22,23]. Several studies have subsequently analyzed the phenomenon [24–26] and various approaches have been proposed to improve the robustness of neural networks [13,27–30]. More closely related to our work are the different proposals aiming at generating better adversarial examples [22,31]. Given an input (train or test) example (x, y), an adversarial example is a perturbed version of the original pattern $\tilde{x} = x + \delta_x$ where δ_x is small enough for $\tilde{x}$ to be undistinguishable from x by a human, but causes the network to predict an incorrect target. Given the network g_θ (where θ is the set of parameters) and a p-norm, the adversarial example is formally defined as:

$$\tilde{x} = \operatorname*{argmax}_{\tilde{x}:\|\tilde{x}-x\|_p \leq \epsilon} \ell\big(g_\theta(\tilde{x}), y\big) \tag{1}$$

where ϵ represents the strength of the adversary. Assuming the loss function $\ell(\cdot)$ is differentiable, [25] propose to take the first order taylor expansion of $x \mapsto \ell(g_\theta(x), y)$ to compute δ_x by solving the following simpler problem:

$$\tilde{x} = \underset{\tilde{x}:\|\tilde{x}-x\|_p \leq \epsilon}{\operatorname{argmax}} \left(\nabla_x \ell(g_\theta(x), y)\right)^T (\tilde{x} - x). \tag{2}$$

When $p = \infty$, then $\tilde{x} = x + \epsilon \cdot \operatorname{sign}(\nabla_x \ell(g_\theta(x), y))$ which corresponds to the *fast gradient sign method* [22]. If instead $p = 2$, we obtain $\tilde{x} = x + \epsilon \cdot \nabla_x \ell(g_\theta(x), y)$ where $\nabla_x \ell(g_\theta(x), y)$ is often normalized. Optionally, one can perform more iterations of these steps using a smaller step-size. This strategy has several variants [31,32]. In the rest of the paper, we refer to this method by *iterative fast gradient method* (IFGM) and will use it both to measure the robustness of a given model and to perform model criticism.

Model Criticism. Example-based explanations are a well-known tool in the realm of cognitive science for facilitating human understanding [33]. They have extensively been used in case-based reasoning (CBR) to improve the interpretability of the models [34–36]. In most cases, it consists in helping a human to understand a complex distribution (or the statistics captured by a model) by presenting her with a set of prototypical examples. However, when the distribution or the model for which one is seeking explanation is complex (as is often the case with real-world data), prototypes may not be enough. Recently, Kim et al. [20] have proposed to use, in addition to the prototypes, data points sampled from regions of the input space not well captured by the model or the prototypes. Such examples, called *criticism*, are known to improve the human's mental model of the distribution.

Kim et al. [20] introduced MMD-critic, an approach inspired by bayesian model criticism [37] to select the *prototypes* and the *critics* among a given set of examples. MMD-critic uses the maximum mean discrepancy [38] and large-scale submodular optimization [39]. Given a set of examples $\mathcal{D} = \{(x, y)\}_{i=1}^n$, let $S \subset \{1, \ldots, n\}$ such that $\mathcal{D}_S = \{x_i, i \in S\}$. Given a RKHS with kernel function $k(\cdot, \cdot)$, the prototypes of MMD-critic are selected by minimizing the maximum mean discrepancy between $\mathcal{D}$ and $\mathcal{D}_S$. This is formally written as:

$$\max_{S \in 2^{[n]}, |S| \leq m} J_b(S) = \frac{2}{n|S|} \sum_{i \in [n], j \in S} k(x_i, x_j) - \frac{1}{|S|^2} \sum_{i,j \in S} k(x_i, x_j) \tag{3}$$

Given a set of prototypes, the criticisms are similarly selected using MMD to maximize the deviation from the prototypes. The objective function in this case is regularized to promote diversity among the criticisms. A greedy algorithm can be used to select both prototypes and criticisms since the corresponding optimization problems are provably submodular and monotone under certain conditions [39]. In our experimental study, we will use MMD-critic as a baseline for example-based explanations.

Feature-Based Explanation. Machine learning and especially Deep Neural Networks (DNNs) lie at the core of more and more technological advances across various fields. However, those models are still widely considered as *black boxes*, leading end users to mistrust the predictions or even the underlying models. In order to promote the adoption of such algorithms and to foster their positive technological impact, recent studies have been focusing on understanding a model from the human perspective [40–43].

In particular, [44] propose to explain the predictions of any classifier g_θ (denoted as g) by approximating it locally with an interpretable model h. The role of h is to provide qualitative understanding between the input x and the classifier's output $g(x)$ for a given class. In the case where the input x is an image, h will act on vector $x' \in \{0,1\}^d$ denoting the presence or absence of the d super-pixels that partition the image x to *explain* the classifier's decision.

Finding the best explanation $\xi(x)$ among the candidates h can be formulated as:

$$\xi(x) = \underset{h \in H}{\operatorname{argmin}} \, \mathcal{L}(g, h, \pi_x) + \Omega(h) \tag{4}$$

where the best explanation minimizes a local weighted loss $\mathcal{L}$ between g and h in the vicinity π_x of x, regularized by the complexity $\Omega(h)$ of such an explanation. The authors restrict $h \in H$ to be a linear model such that $h(z') = w_h z'$. They further define the vicinity of two samples using the exponential kernel

$$\pi_x(z) = e^{\frac{\|x - z\|^2}{\sigma^2}} \tag{5}$$

and define the local weighted loss $\mathcal{L}$ as:

$$\mathcal{L}(g, h, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(g(z) - h(z') \right)^2 \tag{6}$$

where $\mathcal{Z}$ if the dataset of n perturbed samples obtained from x by randomly activating or deactivating some super-pixels in x. Note that z' denotes the one-hot encoding of the super-pixels whereas z is the actual image formed by those super-pixels. Finally, the interpretability of the representation is controlled by

$$\Omega(h) = \begin{cases} 0 & \text{if} \quad \|w_h\|_0 < K \\ +\infty & \text{otherwise} \end{cases} \tag{7}$$

The authors solve this optimization problem by first selecting K features with Lasso and then learning the weights w_g *via* least squares. In the case of DNNs, an explanation generated by this algorithm called LIME allows the user to just highlight the super-pixels with positive weights towards a specific class (see Fig. 5). In what follows we set $\sigma = 0.25$, $n = 1000$ and keep K constant. In the following we refer to an image prompted solely with its top super-pixels as an image *with attention*.

3   Experiments

3.1   Human Subject Study of Classification Errors

We conduct a study of the misclassifications of various pre-trained architectures (*e.g.* ResNet-18, ResNet-101, DenseNet-121, DenseNet-161) with human subjects on Amazon Mechanical Turk (AMT). To this end, we use the center-cropped images of size 224×224 from the standard validation set of Imagenet. The same setting holds in all our experiments. For every architecture, we consider all the examples misclassified by the model. Each of these examples is presented to five (5) different turkers together with the class (C) predicted by the model. Each turker is then asked the following question "*Is class C relevant for this image.*" The possible answers are *yes* and *no*. The former means that the turker agrees with the prediction of the network, while the latter means that the example is misclassified. We take the following measures to ensured high quality answers: (1) we only recruit *master turkers*, *i.e.* turkers who have demonstrated excellence across a wide range of tasks and are awarded the masters qualification by Amazon (2) we mitigate the effect of bad turkers by aggregating 5 independent answers for every question (3) we manually sample more than 500 questions and successfully cross-validated our own guesses with the answers from the turkers, finding that 4 or more positive answers over 5 leads to almost no false positives.

Figure 2 shows a breakdown of the misclassified images by a ResNet-101 (resp. a ResNet-18) according to the number of positive answers they receive from the turkers. Our first results show that for a ResNet-101 (resp. a ResNet-18), for 39.76% (resp. 27.14%) of the misclassified examples, all the turkers (5/5) agree with the prediction of the model. If we further consider the prediction of the model for a given image to be correct if at least four turkers out of five (4/5) agree with it, the rectified Top-1 error of the models are drastically reduced.
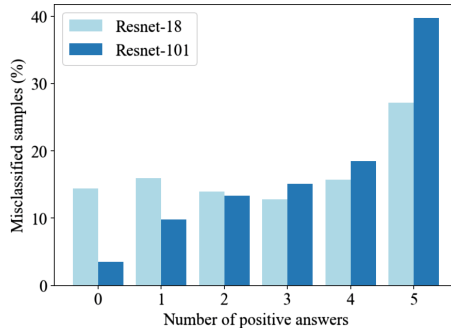

Fig. 2. Positive answers for misclassified samples. An image is prompted to 5 different subjects and a **positive** answer means the subject **agrees with the predicted class**

Table 1 shows the original Top-1 errors together with the rectified versions. For the ResNet-101 and the DenseNet-161, the rectified Top-1 error is respec-

Table 1. Standard and rectified Top-1 errors for various models. Rectified Top-1 errors are significantly lower.

Model	Top-1 error	Rectified Top-1 error
Resnet-18	31.39	17.93
Resnet-101	22.69	9.47
Densenet-121	25.53	14.37
Densenet-161	22.85	10.87

tively 9.47% and 10.87%. Similarly, the top-5 error of the ResNet-101 on a single crop is 6.44%. When we present the misclassified images to the turkers together with the top 5 predictions of the model, the rectified top-5 error drops to 1.94%. When instead of submitting to the turker the misclassified images (*i.e* top-1-misclassified images), we present them the pictures for which the ground truth is not in the top-5 predictions (*i.e.* top-5-misclassified images), the rectified top-5 error drops further to 1.37%. This shows that while the top-5 score is often used to mitigate the fact that many classes are present in the same image, it does not reflect the whole multi-label nature of those images. Moreover, the observation on top-5 is in line with the conclusions regarding top-1. If top-5 is the important measure, this experiment suggests that Imagenet is (almost) solved as far as accuracy is concerned, therefore explaining the marked slowdown in performance improvement observed recently on Imagenet.

The difference between the ground truth labeling of the images and the predictions of the models validated by the turkers can be traced back to the collection protocol of Imagenet. Indeed to create the dataset, Deng et al. [3] first queried several image search engines (using the synsets extracted from WordNet) to obtain good candidate images. Turkers subsequently cleaned the collected images by validating that each one contains objects of a given synset. This labeling procedure ignores the *intrinsic multilabel* nature of the images, neither does it take into account important factors such as *composionality*. Indeed, it is natural that the label `wing` is also relevant for a picture displaying a plane and labelled as `airliner`. Figure 3 shows examples of misclassifications by a RestNet-101 and a DenseNet-161. In most cases, the predicted class is (also) present in the image.

3.2   Study of Robustness to Adversarial Examples

We conducted a human subject study to investigate the appreciation of adversarial perturbation by end-users. Similarly to the human subject study of misclassifications, we used the center-cropped images of the Imagenet validation set. Next, we consider a subset of 20 classes and generate an adversarial example from each legitimate image using the IFGSM attack on a pre-trained ResNet-101. We used a step size of $\epsilon = 0.001$ and a maximum number of iterations of $M = 10$. This attack deteriorated the accuracy of the network on the vali-

Fig. 3. Some test samples misclassified by a ResNet-101 (first row) and a Densenet-161 (second row). The predicted class is indicated in red, the ground truth in black and in parenthesis. All those examples gathered more than four (4 or 5) positive answers over 5 on AMT. Note that no adversarial noise has been added to the images. (Color figure online)

dation set from 77.31% down to 7.96%. Note that by definition, we only need to generate adversarial samples for the correctly predicted test images. We also only consider non-targeted attacks since targeted attacks often require larger distorsion from the adversary and are more challenging if the target categories are far from the ones predicted in case of non-targetted attacks.

We consider two settings in this experiment. In the first configuration, we present the turkers with the whole adversarial image together with the prediction of the network. We then ask each turker if the predicted label is relevant to the given picture. Again, the possible answers are *yes* and *no*. Each image is shown to five (5) different turkers. In the second configuration, we show each turker the interpretation image generated by LIME instead of the whole adversarial image using the top 8 most important features. The rest of the experimental setup is identical to the previous configuration where we showed the whole images to the turkers. If a turker participated in the second experiment with the interpretation images after participating in the first experiment where the whole images are displayed, his answers could be biased. To avoid this issue, we perform the two studies with three days intervals. Similarly to our previous experiments, we report the rectified Top-1 error on adversarial examples by considering the prediction of the model as correct if at least 4/5 turkers agree with it.

Table 2 shows the standard and the rectified Top-1 errors for the adversarial examples. Two observations can be made. First, the robustness of the models as measured by the Top-1 error on adversarial samples generated from the validation set of Imagenet is also underestimated. Indeed, when the whole images are displayed to them, the turkers agree with 22.01% of the predictions of the networks on adversarial examples. This suggests that often, the predicted label is easily identifiable in the picture by a human. Figure 5 shows an example (labeled
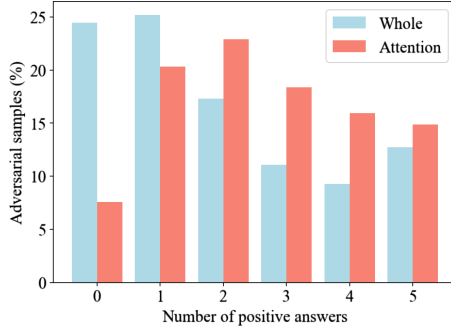
Fig. 4. Positive answers for adversarial samples. The images are either displayed as a whole (entire image) or with attention (explanation). Every image is prompted to 5 different subjects. A **positive** answer means the subject **agrees with the predicted adversarial class**.

Table 2. Standard and rectified Top-1 errors on adversarial examples for the whole images and the explanations.

Setting	Top-1 error	Rectified Top-1 error
Whole image	92.04	76.78
Explanation	92.04	70.68

`jeep`) with the explanations of the predictions for the legitimate and adversarial versions respectively. One can see that the adversarial perturbation exploits the ambiguity of the image by shifting the attention of the model towards regions supporting the adversarial prediction (*e.g.* the red cross).

The second substantial observation is that the percentage of agreement between the predictions of the model and the turkers increases from 22.01% to 30.80% when the explanation is shown instead of the whole image (see also Fig. 4). We further inspected the adversarial images on which the turkers agree with the model's prediction when the explanation is shown but mostly disagree when the whole image is shown. Figure 6 displays examples of such images. In most cases, even though the predicted label does not seem correct when looking at the whole image, the explanation has one of the two following effects. It either reveals the relevant object supporting the prediction (e.g., `Tripod` image) or creates an ambiguous context that renders the predicted label plausible (e.g., `Torch` or `Honeycomb` images). In all cases, providing explanations to the user mitigates the impact of misclassifications due to adversarial perturbations.

3.3 Adversarial Examples for Model Criticism

Example-based explanation methods such as model criticism aim at summarizing the statistics learned by a model by using a carefully selected subset of

Fig. 5. Left: an adversarial image of the true class `Jeep` predicted as `Ambulance` by the network. Center: the explanation of the clean image for its prediction (`Jeep`). Right: the explanation of the adversarial image for its prediction (`Ambulance`). (Color figure online)
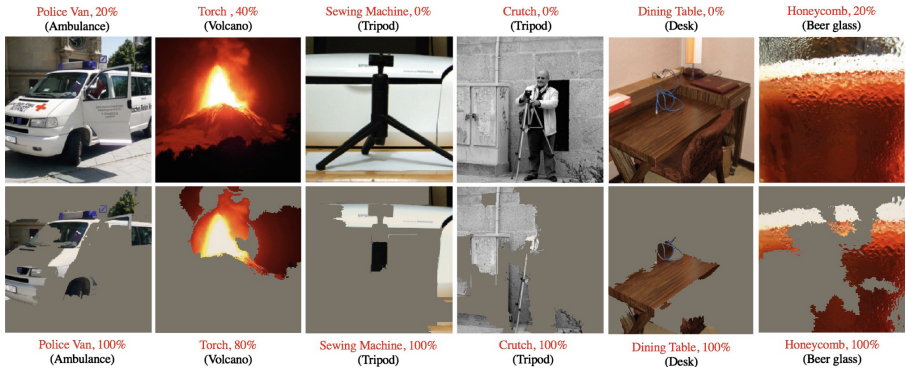

Fig. 6. Adversarial samples, displayed as a whole (entire image, first row) or with explanation (only k top super-pixels, second row). The adversarial class is indicated in red, the true class in black and between parenthesis. In red is also displayed the percentage of positive answers for the displayed image (over 5 answers). In both cases, a **positive** answer means the subject **agrees with the predicted adversarial class**. (Color figure online)

examples. MMD-critic proposes an effective selection procedure combining prototypes and criticisms for improving interpretability. Though applicable to a pre-trained neural network by using the hidden representation of the examples, it only indirectly exploits the discriminative nature of the classifier. In this work, we argue that adversarial examples can be an accurate alternative to MMD-critic for model criticism. In particular, they offer a natural way of selecting prototypes and criticism based on the number of steps of FGSM necessary to change the decision of a classifier for a given example. Indeed, for a given class and a fixed number of maximum steps M of IFGSM, the examples that are still correctly classified after M steps can be considered as useful prototypes since they are very representative of what the classifier has learned from the data for this class. In contrast, the examples whose decision change after 1 or few steps of FGSM are more likely to be valid criticisms because they do not quite fit the model.

We conducted a human study to evaluate our hypothesis. At each round, we present a turker with six classes represented by six images each, as well a target

Table 3. Results of the Prototype and Criticism study. We report the average success rate of the task (assign a target image to one of the six groups) for the seven conditions. The adversarial selection procedures outperforms MMD-critic as well as the other baselines used in the study.

Condition	Mean (%)	Std (%)
Adv - Prototypes and Criticisms	**57.06**	3.22
MMD - Prototypes and Criticisms	50.64	3.25
Probs - Prototypes and Criticisms	49.36	3.26
Adv - Prototypes only	55.02	3.23
MMD - Prototypes only	53.18	3.24
Probs - Prototypes only	52.44	3.25
Random	49.80	3.25

sample randomly drawn from one of those six classes. We measure how well the subject can assign the target sample to the class it belongs. The assignment task requires a class to be well-explained by its six images. The challenge is, therefore, to select well those candidate images. The adversarial method selects as prototypes the examples not misclassified IFGSM after $M = 10$ steps, and as criticism the examples misclassified after one step. In addition to MMD-critic (using $\lambda = 10^{-5}$), we compare the adversarial approach to a simple baseline using the probabilities for selecting the prototypes (considerable confidence, e.g., >0.9) and the criticisms (little confidence, e.g., <0.1). We also compare with the baseline randomly sampling examples from a given class. For each method (except the random baseline), we experiment with showing six prototypes only vs. showing three prototypes and three criticisms instead.

To properly exploit the results, we discarded answers based on the time spent by the turkers on the question to only consider answers given in the range from 20 s to 3 min. This resulted in 1,600 valid answers for which we report the results in Table 3. Two observations arise from those results. First, the prototypes and criticisms sampled using the adversarial method represent better the class distribution and achieve higher scores for the assignment task both when only prototypes are used (55.02%) and when prototypes are combined with criticisms (57.06%). Second, the use of criticisms additionally to the prototypes always helps to better grasp the class distribution.

More qualitatively, we display in Figs. 7a and b some prototypes and criticisms for the class `Banana` generated using respectively the MMD and the adversarial method for the test samples, that demonstrate the superiority of the adversarial method.

3.4   Uncovering Biases with Model Criticism

To uncover the undesirable biases learned by the model, we use our adversarial example approach to model criticism since it worked better than MMD-critic in
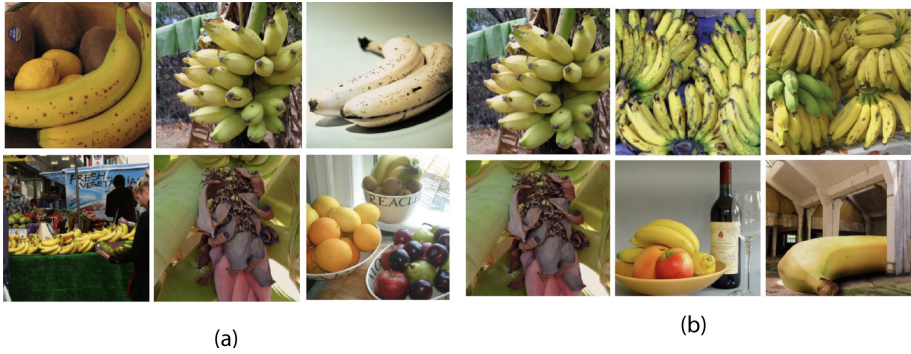
(a)

(b)

Fig. 7. (a) Prototypes (first row) and Criticisms (second row) for the `Banana` class when using the MMD-critic method. (b) Prototypes (first row) and Criticisms (second row), for the `Banana` class when using the adversarial selection.

Fig. 8. Pairs of pictures (columns) sampled over the Internet along with their prediction by a ResNet-101. (Color figure online)

our previous experiments. We consider the class `basketball` (for which humans often appear in the images). We select and inspect a reduced subset of prototypes and criticisms from the category `basketball`. The percentage of `basketball` training images on which at least one white person appears is about 55%. Similarly, the percentage of images on which at least one black person appears is 53%. This relative balance contrasts with the statistics captured by the model. Indeed, on the one hand, we found that about 78% of the prototypes contain at least one black person and only 44% for prototypes contain one white person or more. On the other hand, for criticisms, 90% of the images contain at least one white person and only about 20% include one black person or more. This suggests that the model has learned a biased representation of the class `basketball` where images containing black persons are prototypical. To further validate this hypothesis, we sample pairs of similar pictures from the Internet.

Fig. 9. Prototypes (first row) and Criticisms (second row) for the test images of the `Basketball` class, adv method.

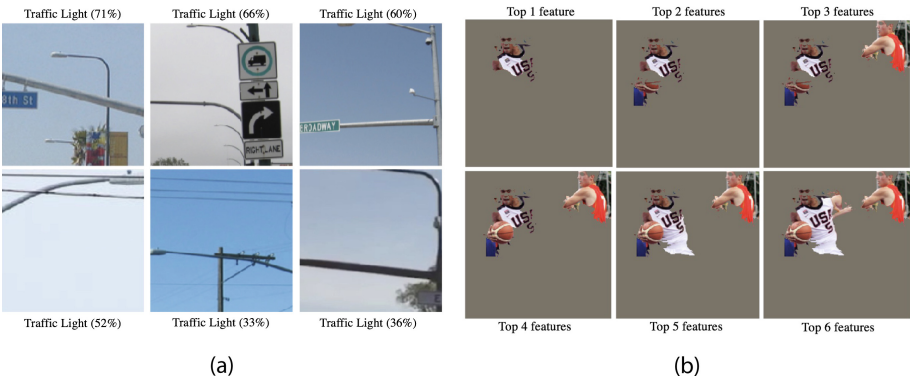(a)   (b)

Fig. 10. (a) Images from the Internet classified as `Traffic Light` for a ResNet-101. (b) Explanation path of a ResNet-101 for a test sample of the `Basketball` class. Note that the first feature to appear includes the writing on the jersey and then the basket ball. (Color figure online)

The pairs are sampled such that the primary apparent difference between the two images is the skin color of the persons. We then fed images to the model and gathered the predictions. Figure 8 shows the results of the experiments. All images containing a black person are classified as basketball while similar photos with persons of different skin color are labeled differently. Figure 10b provides additional supporting evidence for this hypothesis. It shows the progressive feature based explanation path uncovering the super-pixels of an example (correctly classified as `basketball`) by their order of importance. The most importance super-pixels depict the jersey and the skin color of the player. The reasons why the model learns these biases are unclear. One hypothesis is that despite the balanced distribution of races in pictures labeled basketball, black persons are more represented in this class in comparison to the other classes. A similar phenomenon has also been noted in the context of textual data. [19,45]. We defer further investigations on this to future studies (Fig. 9).

Remark. We have focused on racial biases because a human can easily spot them. In fact, we have found similar biases where pictures displaying Asians dressed in red are very often classified as `ping-pong ball`. However, we hypothesize the biases of the model are numerous and diverse. For example, we also have found that the model often predicts the class `traffic light` for images of a blue sky with street lamps as depicted in Fig. 10a. In any case, model criticism has proven effective in uncovering the undesirable hidden biases learned by the model.

4   Conclusion

Through human studies and explanations, we have proven that the performance of SOTA models on Imagenet is underestimated. This leaves little room for improvement and calls for new large-scale benchmarks involving for example multi-label annotations. We have also improved our understanding of adversarial examples from the perspective of the end-user and positioned model criticism as a valuable tool for uncovering undesirable biases. These results open an exciting perspective on designing explanations and automating bias detection in vision models. Our study suggests that more research in these topics will be necessary to sustain the use of machine learning as a general purpose technology and to achieve new breakthroughs in image classification.

References

1. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE (1998)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 1 (2012)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR09 (2009)
4. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher Kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010 Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). https://doi. org/10.1007/978-3-642-15561-1_11
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR (2016)
7. Girshick, R.: Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). ICCV 2015 (2015)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR abs/1703. 06870 (2017)
9. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deeper-Cut: a deeper, stronger, and faster multi-person pose estimation model. CoRR abs/1605.03170 (2016)

10. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016 Part VII. LNCS, vol. 9911, pp. 717–732. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_44
11. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. CoRR (2017)
12. Wang, Y., Deng, X., Pu, S., Huang, Z.: Residual convolutional CTC networks for automatic speech recognition. CoRR (2017)
13. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. CoRR (2017)
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR (2014)
15. Szegedy, C., et al.: Intriguing properties of neural networks. CoRR (2013)
16. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014)
17. Cissé, M., Adi, Y., Neverova, N., Keshet, J.: Houdini: fooling deep structured visual and speech recognition models with adversarial examples. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA (2017)
18. Ritter, S., Barrett, D.G.T., Santoro, A., Botvinick, M.M.: Cognitive psychology for deep neural networks: a shape bias case study. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research (2017)
19. Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. CoRR (2016)
20. Kim, B., Khanna, R., Koyejo, O.O.: Examples are not enough, learn to criticize! Criticism for interpretability. In: Advances in Neural Information Processing Systems 29 (2016)
21. Karpathy, A.: What i learned from competing against a convnet on imagenet. http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/
22. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)
23. Tabacof, P., Valle, E.: Exploring the space of adversarial images. CoRR (2015)
24. Fawzi, A., Fawzi, O., Frossard, P.: Analysis of classifiers' robustness to adversarial perturbations. CoRR (2015)
25. Shaham, U., Yamada, Y., Negahban, S.: Understanding adversarial training: increasing local stability of neural nets through robust optimization. CoRR (2015)
26. Fawzi, A., Moosavi-Dezfooli, S., Frossard, P.: Robustness of classifiers: from adversarial to random noise. CoRR (2016)
27. Papernot, N., McDaniel, P.D., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP) (2016)
28. Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., Usunier, N.: Parseval networks: improving robustness to adversarial examples. In: Proceedings of the 34th International Conference on Machine Learning (2017)
29. Kurakin, A., Boneh, D., Tramr, F., Goodfellow, I., Papernot, N., McDaniel, P.: Ensemble adversarial training: attacks and defenses (2018)
30. Guo, C., Rana, M., Cissé, M., van der Maaten, L.: Countering adversarial images using input transformations. CoRR (2017)

31. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. CoRR (2015)
32. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. CoRR (2016)
33. Simon, H.A., Newell, A.: Human problem solving: the state of the theory in 1970. Am. Psychol. **26**, 145 (1972)
34. Aamodt, A., Plaza, E.: Case-based reasoning; foundational issues, methodological variations, and system approaches. AI Commun. **7**, 39–59 (1994)
35. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences: what's next? Artif. Intell. Med. **36**, 127–135 (2006)
36. Kim, B., Rudin, C., Shah, J.A.: The Bayesian case model: a generative approach for case-based reasoning and prototype classification. In: Advances in Neural Information Processing Systems 27 (2014)
37. Gelman, A.: Bayesian data analysis using R (2006)
38. Gretton, A., Borgwardt, K.M., Rasch, M., Schölkopf, B., Smola, A.J.: A kernel method for the two-sample-problem. In: Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS 2006 (2006)
39. Badanidiyuru, A., Mirzasoleiman, B., Karbasi, A., Krause, A.: Streaming submodular maximization: massive data summarization on the fly. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014)
40. Lipton, Z.C.: The mythos of model interpretability. CoRR (2016)
41. Samek, W., Wiegand, T., Muller, K.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. CoRR (2017)
42. Montavon, G., Samek, W., Muller, K.: Methods for interpreting and understanding deep neural networks. CoRR (2017)
43. Dong, Y., Su, H., Zhu, J., Bao, F.: Towards interpretable deep neural networks by leveraging adversarial examples. CoRR (2017)
44. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
45. Paperno, D., Marelli, M., Tentori, K., Baroni, M.: Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood. Cogn. Psychol. **74**, 66–83 (2014)