

# ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases

Pierre Stock<sup>[0000–0002–3623–3899]</sup> and Moustapha Cisse

Facebook AI Research

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 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

---

Now Google AI Ghana Lead.





























