Why do deep convolutional networks generalize so poorly to small image transformations?

Aharon Azulay
ELSC
Hebrew University of Jerusalem
aharon.azulay@mail.huji.ac.il

Yair Weiss
CSE, ELSC
Hebrew University of Jerusalem
yweiss@cs.huji.ac.il

Abstract

Deep convolutional network architectures are often assumed to guarantee generalization for small image translations and deformations. In this paper we show that modern CNNs (VGG16, ResNet50, and InceptionResNetV2) can drastically change their output when an image is translated in the image plane by a few pixels, and that this failure of generalization also happens with other realistic small image transformations. Furthermore, the deeper the network the more we see these failures to generalize. We show that these failures are related to the fact that the architecture of modern CNNs ignores the classical sampling theorem so that generalization is not guaranteed. We also show that biases in the statistics of commonly used image datasets makes it unlikely that CNNs will learn to be invariant to these transformations. Taken together our results suggest that the performance of CNNs in object recognition falls far short of the generalization capabilities of humans.

1 Introduction

Deep convolutional neural networks (CNNs) have revolutionized computer vision. Perhaps the most dramatic success is in the area of object recognition, where performance is now described as "superhuman" [20]. A key to the success of any machine learning method is the *inductive bias* of the method, and clearly the choice of architecture in a neural network significantly affects the inductive bias. In particular, the choice of convolution and pooling in CNNs is motivated by the desire to endow the networks with invariance to irrelevant cues such as image translations, scalings, and other small deformations [18, 48]. This motivation was made explicit in the 1980s by Fukushima in describing the "neocognitron" architecture, which served as inspiration for modern CNNs [27], "After finishing the process of learning, pattern recognition is performed on the basis of similarity in shape between patterns, and is not affected by deformation, nor by changes in size, nor by shifts in the position of the input patterns." [17]

Despite the excellent performance of CNNs on object recognition, the vulnerability to adversarial attacks suggests that superficial changes can result in highly non-human shifts in prediction [43, 35, 26, 1, 15, 41, 3, 40]. In addition, filtering the image in the Fourier domain (in a way that does not change human prediction) also results in a substantial drop in prediction accuracy [23]. These and other results [34] indicate that CNNs are not invariant to cues that are irrelevant to the object identity.

An argument against adversarial attacks on CNNs is that they often involve highly unnatural transformations to the input images, hence in some sense we would not expect CNNs to be invariant to these transformations (see [13] for a case for human adversarial examples). When considering more natural transformations (e.g. translations, rotations, and scalings), there is preliminary evidence that AlexNet [25] is robust to some of them [48]. On the other hand, there is also preliminary evidence for lack of robustness in the more modern networks [4] along with studies suggesting that with small CNNs and MNIST data, data augmentation is the main feature affecting CNN invariance [24]. An indirect

Preprint. Work in progress.

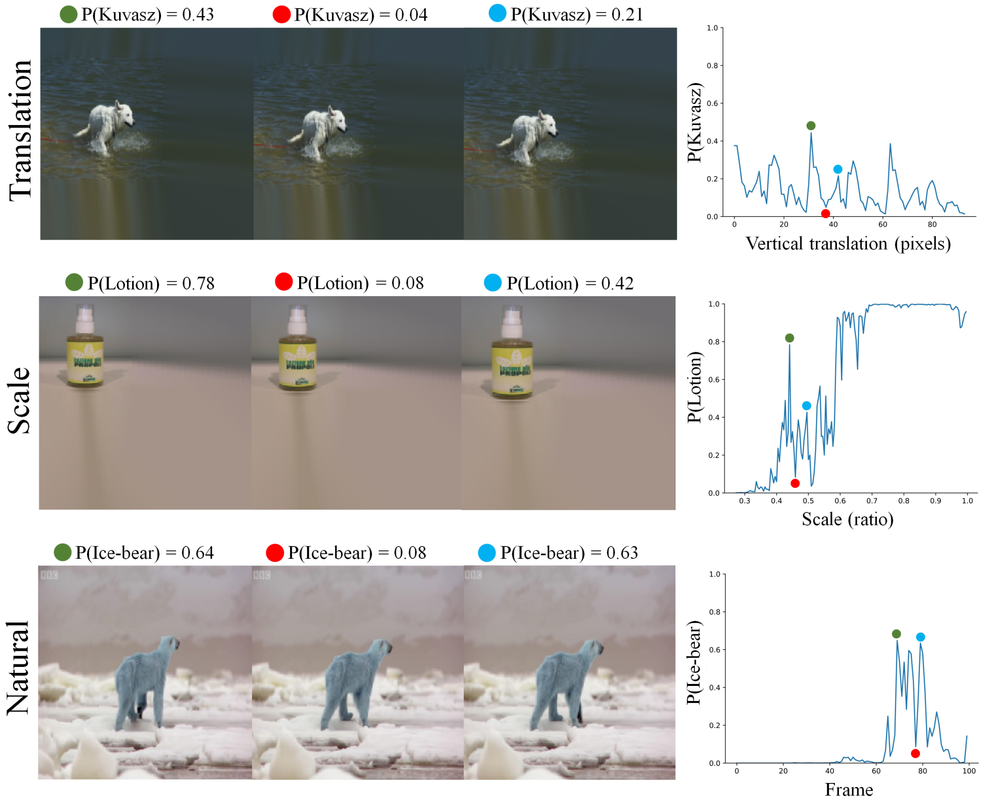

Figure 1: Examples of jagged predictions of modern deep convolutional neural networks. Top: A negligible vertical shift of the object (Kuvasz) results in an abrupt decrease in the network's predicted score of the correct class. Middle: A tiny increase in the size of the object (Lotion) produces a dramatic decrease in the network's predicted score of the correct class. Bottom: A very small change in the bear's posture results in an abrupt decrease in the network's predicted score of the correct class. Colored dots represent images chosen from interesting x-axis locations of the graphs on the right. These dots illustrate sensitivity of modern neural networks to small, insignificant (to a human), and realistic variations in the image (see `https://youtu.be/M4ys8c2NtsE`).

method to probe the invariances measures the linearity of the learned representations under natural transformations to the input image [28, 22, 16, 9]. In this paper, we directly ask "how invariant are modern CNNs to natural image transformations?". Specifically, we systematically examine the invariances of three modern deep CNNs: VGG-16 [39], ResNet-50 [21], and InceptionResNet-V2 [42]. We find that modern deep CNNs are not invariant to translations, scalings and other realistic image transformations, and this lack of invariance is related to the subsampling operation and the biases contained in image datasets.

2 Failures of modern CNNs

Figure 1 contains examples of abrupt failures following tiny realistic transformations for the InceptionResNet-V2 CNN. Shifting or scaling the object by just one pixel could result in a sharp change in prediction. In the top row, we embed the original image in a larger image and shift it in the image plane (while filling in the rest of the image with a simple inpainting procedure). In the middle row, we repeat this protocol with rescaling. In the bottom row, we show frames from a BBC film in which the ice bear moves almost imperceptibly between frames and the network's output changes dramatically (see `https://youtu.be/M4ys8c2NtsE`).

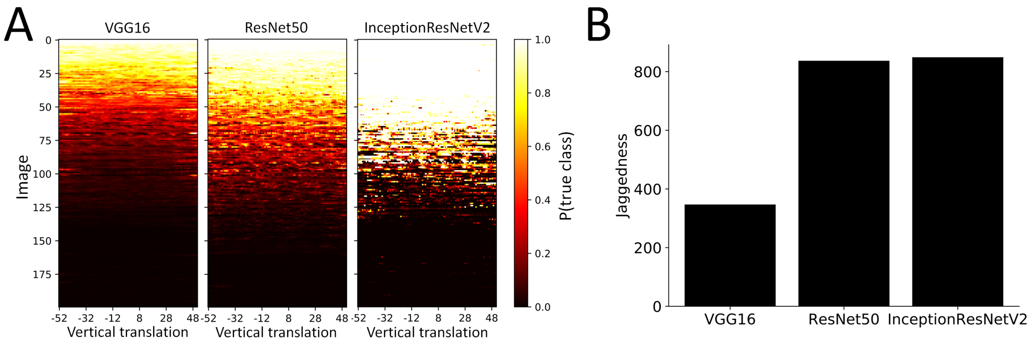

Figure 2: Modern deep convolutional neural networks are sensitive to small image translations. A) Comparison of three networks of various depths in the task of vertical image translation depicted in figure 1. Images (rows) are randomly chosen from the ImageNet dataset [10], and are sorted by the network's prediction sum in a descending order. B) The deeper the network, the more jagged its predictions.

Network	Top-1	Top-5	Parameters	Depth
VGG16 [39]	0.715	0.901	138,357,544	16
ResNet50 [21]	0.759	0.929	25,636,712	50
InceptionResNetV2 [42]	0.804	0.953	55,873,736	134

Table 1: The networks used (taken from (`https://keras.io/applications/`))

In order to measure how typical these failures are, we randomly chose 200 images from the ImageNet validation set and measured the output of three modern CNNs as we embedded these images in a larger image and systematically varied the vertical translation. As was the case in figure 1, we used a simple inpainting procedure to fill in the rest of the image. The networks we chose and their parameters are shown in table 1.

Results are shown in figure 2. Each row corresponds to an image under different translations and the color denotes the network's estimate of the probability of the correct class. Thus a row that is all light corresponds to a correct classification that is invariant to translation, while a row that is all dark corresponds to an incorrect classification that is invariant to translation. Surprisingly, many rows show abrupt transitions from light to dark, indicating that the classification changes abruptly as the object is translated. We quantify the lack of invariance by a measure we call "jaggedness": the number of times the network's predictions had the correct class in its top-5 and after just one pixel shift it moved outside of the top-5 (and also the opposite transition from non-top-5 to top5). Using this measure, we find that for approximately 28% of the images, the output is "jagged", i.e the network changes its prediction by a shift of a single pixel. Also, as shown in the bottom of figure 2, jaggedness is greater for the modern, deeper, networks compared to the less modern VGG16 network. While the deeper networks have better test accuracy, they are also less invariant.

A natural criticism of these results is that they are somehow related to the image resizing and inpainting procedures that we used. To test this possibility, we repeated the experiment with a different protocol where we chose different crops of the original ImageNet image while making sure that the object bounding box remained within the crop. This protocol does not require any inpainting while still translating the object location within the new image. Results are shown in figure 3. We still have a large fraction of images for which the prediction is not invariant to translation, and "jaggedness" is higher for more modern networks.

3 Ignoring the Sampling Thereom

The failure of CNNs to generalize to image translations is particularly puzzling. Intuitively, it would seem that if all layers in a network are convolutional then the representation should simply translate when an image is translated. If the final features for classification are obtained by a global pooling

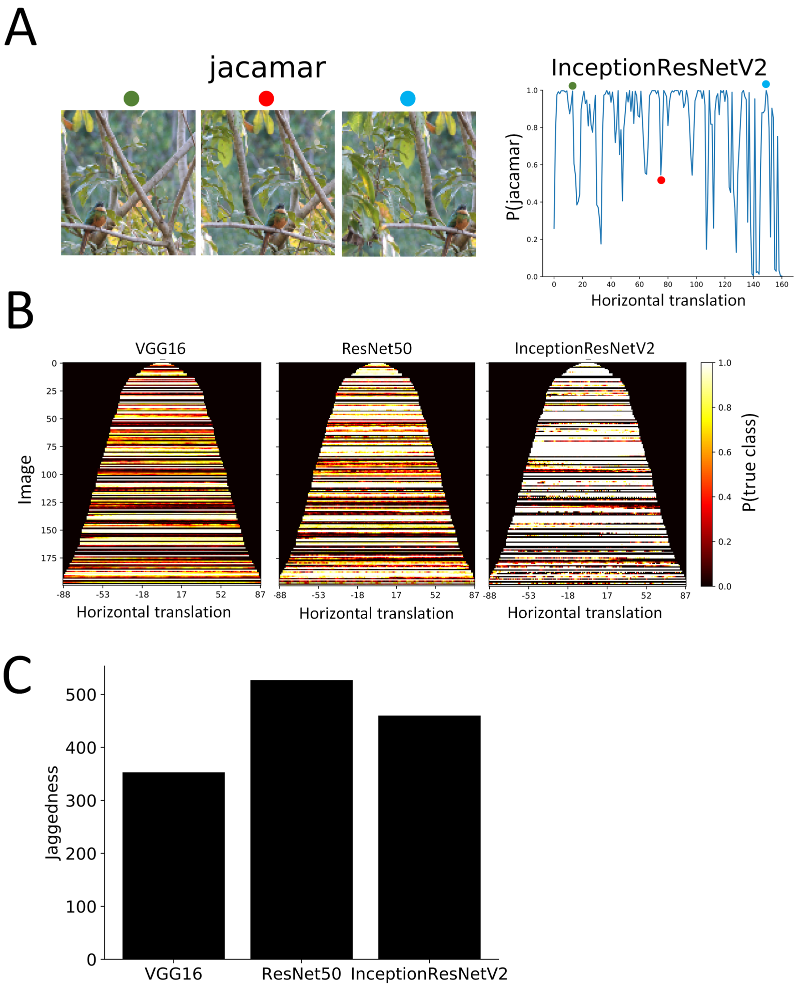

Figure 3: Modern deep convolutional neural networks are sensitive to small image translations (Without image downscaling). A) Example of InceptionResNetV2 sensitivity to very small horizontal translations. B) Comparison of three networks of various depths (16, 50, and 134 for VGG16, ResNet50, and InceptionResNetV2 respectively) in the task of horizontal image translation. Rows are sorted by their the size of the bounding box of the object (large objects have fewer possible translations that keep the entire object in the frame). B) Modern networks have more jagged predictions. Jaggedness is calculated by counting the number of times the network's predictions had the correct class in its top-5 and after just one pixel shift it moved outside of the top-5 (and also the opposite transition from non-top-5 to top5).

operation on the representation (as is done for example in ResNet50 and InceptionResNetV2) then these features should be invariant to translation. Where does this intuition fail?

This intuition ignores the subsampling operation which is prevalent in modern CNNs, also known as "stride". This failure of translation invariance in systems with subsampling was explicitly discussed in Simoncelli et al. [38] who wrote "We cannot literally expect translation invariance in a system based on convolution and subsampling: translation of the input signal cannot produce simple translations of the transform coefficients, unless the translation is a multiple of each of the subsampling factors in the system". Since deep networks often contain many subsampling operations, the subsampling factor of the deep layers may be very large so that "literal" translation invariance only holds for very special translations. In InceptionResnetV2, for example, the subsampling factor is 45, so we expect exact translation invariance to hold only for $\frac{1}{45^2}$ of possible translations.

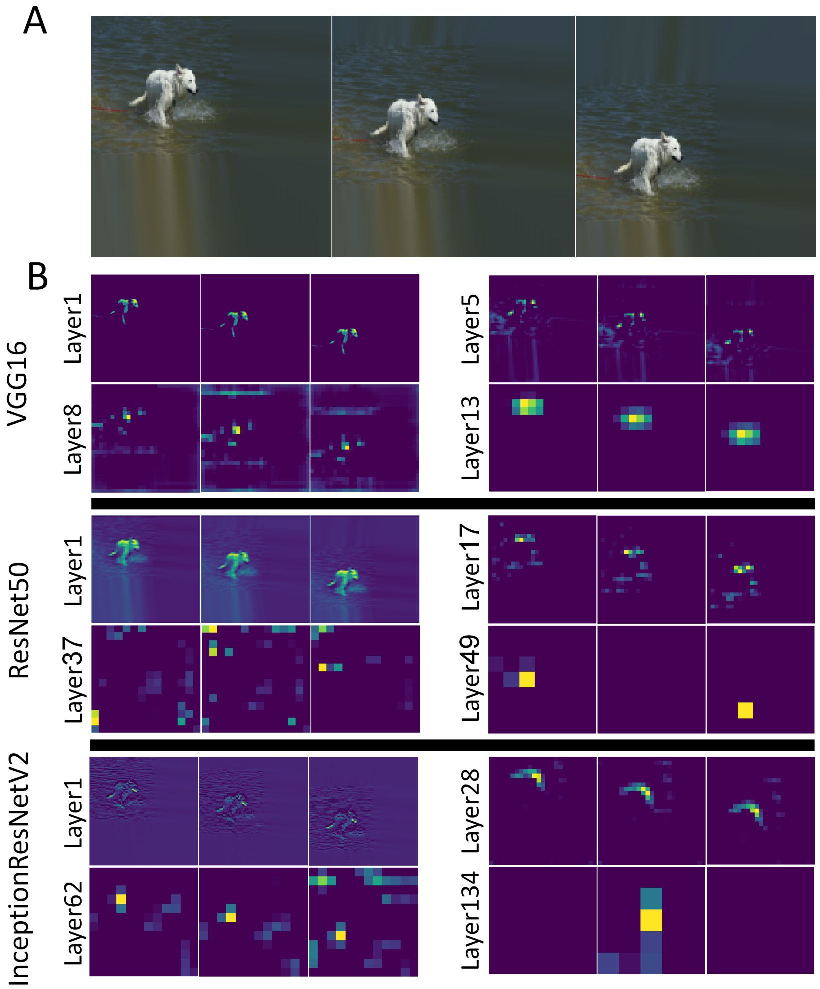

Figure 4: The deeper the network, the less shiftable are the feature maps. A) A vertical shift of a "Kuvasz" dog in the image plane. B) Feature maps from three different network architectures in response to the translated Kuvasz image. Layer depth assignments reflect the number of trainable convolutional layers preceding the selected layer. The last layer is always the last convolutional layer in each network.

Simoncelli et al. also defined a weaker form of translation invariance, which they called "shiftability" and showed that it can hold for systems with subsampling (this is related to weak translation invariance as defined by [28], see also [14, 9] for related ideas applied to neural networks). Here we extend the basic shiftability result to show that when shiftability holds, then global pooling will indeed yield invariant representations.

We define $r(x)$ as the response of a feature detector at location x in the image plane. We say that this response is "convolutional" if translating the image by any translation δ yields a translation of the response by the same δ. This definition includes cases when the feature response is obtained by convolving the input image with a fixed filter, but also includes combinations of linear operations and nonlinear operations that do not include any subsampling.

We start by a trivial observation:

Observation: If $r(x)$ is convolutional then global pooling $r = \sum_x r(x)$ is translation invariant.

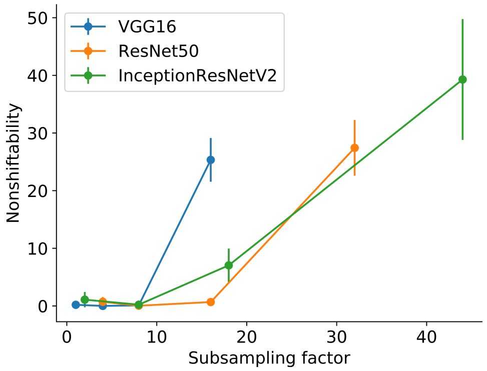

Figure 5: A measure of nonshiftability in different feature maps of VGG, Resnet50 and Inception-ResnetV2 (number of times the global sum of activities in a randomly chosen feature map changes significantly as the pattern is shifted). Architectures that obey the sampling theorem should always be shiftable, but in the more modern networks the deeper layers are not shiftable.

Proof: This follows directly from the definition of a convolutional response. If $r(x)$ is the feature response to one image and $r_2(x)$ is the feature response to the same image translated, then $\sum_x r(x) = \sum_x r_2(x)$ since the two responses are shifts of each other.

Definition: A feature detector $r(x)$ with subsampling factor s is called "shiftable" if for any x the detector output at location x can be linearly interpolated from the responses on the sampling grid:

$$r(x) = \sum_i B_s(x - x_i) r(x_i)$$

where x_i are located on the sampling grid for subsampling factor s and $B_s(x)$ is the basis function for reconstructing $r(x)$ from the samples.

The classic Shannon-Nyquist theorem tells us that $r(x)$ will be shiftable if and only if the sampling frequency is at least twice the highest frequency in $r(x)$.

Claim: If $r(x)$ is shiftable then global pooling on the sampling grid $r = \sum_i r(x_i)$ is translation invariant.

Proof: This follows from the fact that global pooling on the sampling grid is (up to a constant) the same as global pooling for all x.

$$\begin{aligned} \sum_x r(x) &= \sum_x \sum_i r(x_i) B(x - x_i) & (1) \\ &= \sum_i r(x_i) \sum_x B(x - x_i) & (2) \\ &= K \sum_i r(x_i) & (3) \end{aligned}$$

where $K = \sum_x B(x - x_i)$ and K does not depend on x_i.

Figure 4 examines the extent to which the representations learned by modern CNNs are invariant or shiftable. The top row shows an image that is translated vertically, while the bottom three rows show the representations in different layers for the three CNNs in table 1. For VGG16 the representation appears to shift along with the object, including the final layer where the blurred pattern of response is not a simple translation of the original response, but seems to preserve the global sum for this particular image. For the two more modern networks, the responses are sharper but lose their shiftability in the later layers. In particular, the final layers show approximate invariance to one special translation but no response at all to another translation, suggesting that the many layers of subsampling yield a final response that is not shiftable.

As a more quantitative measure of shitability, figure 5 counts the number of times the global sum of activities in a randomly chosen feature map changes significantly (more than 20% of mean) as the input is shifted (averaged over 10 randomly chosen images). According to the preceding analysis, in architectures that obey the sampling theorem, the global sum should be invariant to input translation so nonshiftability should be zero in all layers. As can be seen, in the more modern networks the deeper layers are increasingly not shiftable.

The basic message of the sampling theorem is that *you should always blur before subsampling*. Translated to the language of neural networks this means that stride (i.e. subsampling) should always be combined with pooling (i.e. blurring) in the preceding layer. While VGG16 has 5 pooling operations in its 16 layers, Resnet50 has only one pooling operation among its 50 intermediate layers and InceptionResnetV2 has only 5 among its intermediate 134 layers. Thus the fact that in the more modern CNNs, the global sum of the feature maps is not preserved with input translation is not surprising.

4 Why don't modern CNNs learn to be invariant from data?

While the preceding discussion suggests that the CNN architecture will not yield translation invariance "for free", there is still the possibility that the CNN will learn a translation invariant prediction from the training examples. This requires that the training set will actually be invariant to the irrelevant transformations. We examined the degree of invariance in the ImageNet training set by manually labeling the training images in five categories: Tibetan terrier, elephant, pineapple, bagel and rowing paddle.

Consistent with previous results on "dataset bias" [37, 33, 2, 44, 45, 32] we find that the ImageNet dataset is not invariant to translations and rescalings. Figure 6 shows the distribution of the distances between the eyes of a "Tibetan terrier" and the positions of the center point between the dog's eyes. Notice that both distributions are far from uniform. Similar results are obtained for the other four categories.

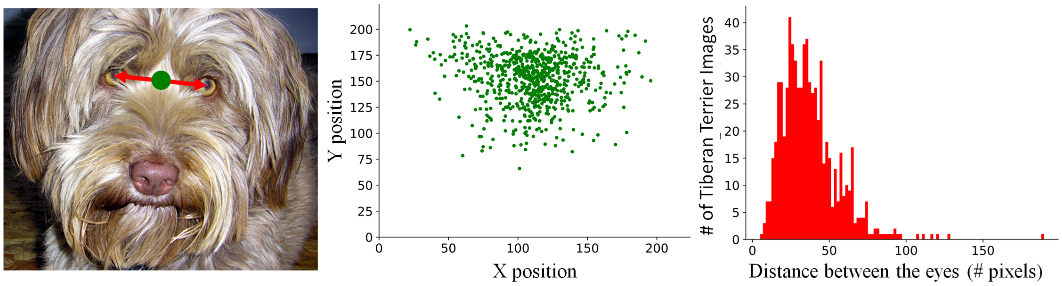

Figure 6: Photographer's biases in the ImageNet's "Tibetan terrier" category. Left: Example of the hand-labeling procedure. Middle: Positions of the middle point between the dog's eyes. Right: Histogram of distances between the dog's eyes. Notice the bias in both the object's position and scale.

To be more quantitative, we used the available bounding-box labels, and extracted the center point of the bounding-box and its height as proxies for the object position and size respectively. We then applied a statistical significance test to ask whether object location and object sizes were uniform for that category. For more than 900 out of the 1000 categories we found that location and size were highly non uniform ($P < 10^{-10}$). In summary, whether we use our own hand labeling to measure invariance or we use the available bounding boxes, the ImageNet dataset has strong biases for object position and size and we cannot expect a learning system to learn an invariant representation.

Even if the training set is not invariant, we can make it invariant using *data augmentation*. Will this make the CNN learn an invariant prediction? First, we note that we used pretrained networks and according to the author's description of the training procedure, all three networks were trained using data augmentation. Obviously, not any data augmentation is sufficient for the networks to learn invariances. To understand the failure of data augmentation, it is again instructive to consider the subsampling factor. Since in modern networks the subsampling factor is approximately 45, then for a system to learn complete invariance to translation only, it would need to see $45^2 = 2025$ augmented versions of each training example. If we also add invariance to rotations and scalings, the number grows exponentially with the number of irrelevant transformations.

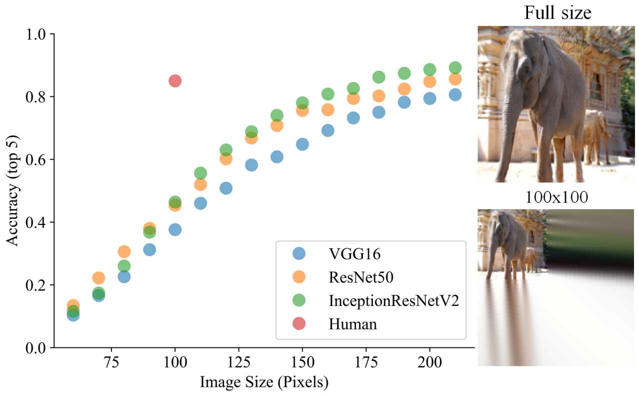

Figure 7: The performance of modern CNNs on test images from ImageNet that are embedded in a random location in a larger image is quite poor (less than 50% accuracy). Human performance is not affected. Right: An example of a full sized image and the same image resized to 100x100.

5 Implications for practical systems

Although our results show that modern CNNs fail to generalize for small image transformations, their performance on the ImageNet test set is still amazingly good and far better than previous techniques. This is related to the fact that the ImageNet test set contains the same photographer's biases as the training set, so generalization to very different sizes and locations is not required. To highlight this point, we created a new test set in which ImageNet images were embedded in a larger image in a random location (and the missing pixels were filled in using a simple inpainting algorithm). Figure 7 shows that human performance is not affected by the rescaling and random translations, while the performance of modern CNNs deteriorates dramatically. In fact, when images are scaled to half their original size and randomly translated, the accuracy of modern CNNs is less than 50%. For comparison, the shallow methods that preceded CNNs achieved over 70% accuracy on the original ImageNet data, and this was considered poor performance.

One way in which modern systems partially address the lack of invariance is using *test time augmentation* in which the system output on a given image is computed by a majority vote among many random crops of the image. Clearly this is wasteful in resources and still only provides partial invariance.

6 Discussion

CNN archictectures were designed based on an intuition that the convolutional structure and pooling operations will give invariance to translations and small image deformations "for free". In this paper we have shown that this intuition breaks down once subsampling, or "stride" is used and we have presented empirical evidence that modern CNNs do not display the desired invariances since the architecture ignores the classic sampling theorem. This still leaves open the possibility of a CNN learning invariance from the data but we have shown that the ImageNet training and testing examples include significant photographer's bias so that it is unlikely that a system will learn invariance using these examples. We have also discussed the fact that the failure to learn invariance holds even when "data augmentation" is used since the number of augmented examples needed to learn the desired invariance grows exponentially.

In addition to pointing out these failures, the sampling theorem also suggests a way to impose translation invariance by ensuring that all representations are sufficiently blurred before subsampling. However, such blurred representations may lead to a decrease in performance, especially in datasets and benchmarks that contain photographer's bias. Alternatively, one could use specially designed features such as SIFT in which invariance is hard coded [7, 29, 30, 31] or neural network architectures that explicitly enforce invariance [36, 19, 5, 6, 11, 12, 47, 46, 8]. Again, as long as the datasets contain significant photographer's bias, such invariant approaches may lead to a decrease in performance.

Acknowledgments

We thank Tal Arkushin for the helpful comments. Support by the ISF and the Gatsby Foundation is gratefully acknowledged.

References

[1] Anish Athalye and Ilya Sutskever. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[2] Tamara L Berg and Alexander C Berg. Finding iconic images. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2009.

[3] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Exploring the space of black-box attacks on deep neural networks. *arXiv preprint arXiv:1712.09491*, 2017.

[4] Charlotte Bunne, Lukas Rahmann, and Thomas Wolf. Studying invariances of trained convolutional neural networks. *arXiv preprint arXiv:1803.05963*, 2018.

[5] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016.

[6] Gong Cheng, Peicheng Zhou, and Junwei Han. Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2884–2893. IEEE, 2016.

[7] Olivier Chomat, Vincent Colin de Verdière, Daniela Hall, and James L Crowley. Local scale selection for gaussian based description techniques. In *European Conference on Computer Vision*, pages 117–134. Springer, 2000.

[8] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016.

[9] Taco S Cohen and Max Welling. Transformation properties of learned visual representations. *arXiv preprint arXiv:1412.7659*, 2014.

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

[11] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *arXiv preprint arXiv:1602.02660*, 2016.

[12] Sander Dieleman, Kyle W Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly notices of the royal astronomical society*, 450(2):1441–1459, 2015.

[13] Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both human and computer vision. *arXiv preprint arXiv:1802.08195*, 2018.

[14] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. *arXiv preprint arXiv:1709.01889*, 2017.

[15] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 1, 2017.

[16] Alhussein Fawzi and Pascal Frossard. Manitest: Are classifiers really invariant? *arXiv preprint arXiv:1507.06535*, 2015.

[17] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[18] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[19] Robert Gens and Pedro M Domingos. Deep symmetry networks. In *Advances in neural information processing systems*, pages 2537–2545, 2014.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Olivier J Hénaff and Eero P Simoncelli. Geodesics of learned representations. *arXiv preprint arXiv:1511.06394*, 2015.

[23] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

[24] Eric Kauderer-Abrams. Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450*, 2017.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[26] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[27] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[28] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 991–999. IEEE, 2015.

[29] Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.

[30] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[31] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[32] Elad Mezuman and Yair Weiss. Learning about canonical views from internet image collections. In *Advances in Neural Information Processing Systems*, pages 719–727, 2012.

[33] Rahul Raguram and Svetlana Lazebnik. Computing iconic summaries of general visual concepts. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[34] Erik Rodner, Marcel Simon, Robert B Fisher, and Joachim Denzler. Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches. *arXiv preprint arXiv:1610.06756*, 2016.

[35] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540. ACM, 2016.

[36] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1233–1240. IEEE, 2013.

[37] Ian Simon, Noah Snavely, and Steven M Seitz. Scene summarization for online image collections. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[38] Eero P Simoncelli, William T Freeman, Edward H Adelson, and David J Heeger. Shiftable multiscale transforms. *IEEE transactions on Information Theory*, 38(2):587–607, 1992.

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

[40] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017.

[41] Fnu Suya, Yuan Tian, David Evans, and Paolo Papotti. Query-limited black-box attacks to classifiers. *arXiv preprint arXiv:1712.08713*, 2017.

[42] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[44] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.

[45] Tobias Weyand and Bastian Leibe. Discovering favorite views of popular places with iconoid shift. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1132–1139. IEEE, 2011.

[46] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[47] Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *arXiv preprint arXiv:1411.6369*, 2014.

[48] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.