Understanding the role of individual units in a deep neural network

David Bau[a,1], Jun-Yan Zhu[a,b], Hendrik Strobelt[c], Agata Lapedriza[d,e], Bolei Zhou[f], and Antonio Torralba[a]

[a]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Adobe Research, Adobe Inc., San Jose, CA 95110; [c]Massachusetts Institute of Technology–International Business Machines (IBM) Watson Artificial Intelligence Laboratory, Cambridge, MA 02142; [d]Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139; [e]Estudis d'Informàtica, Multimèdia i Telecomunicació, Universitat Oberta de Catalunya, 08018 Barcelona, Spain; and [f]Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved July 7, 2020 (received for review August 31, 2019)

Deep neural networks excel at finding hierarchical representations that solve complex tasks over large datasets. How can we humans understand these learned representations? In this work, we present network dissection, an analytic framework to systematically identify the semantics of individual hidden units within image classification and image generation networks. First, we analyze a convolutional neural network (CNN) trained on scene classification and discover units that match a diverse set of object concepts. We find evidence that the network has learned many object classes that play crucial roles in classifying scene classes. Second, we use a similar analytic method to analyze a generative adversarial network (GAN) model trained to generate scenes. By analyzing changes made when small sets of units are activated or deactivated, we find that objects can be added and removed from the output scenes while adapting to the context. Finally, we apply our analytic framework to understanding adversarial attacks and to semantic image editing.

machine learning | deep networks | computer vision

Can the individual hidden units of a deep network teach us how the network solves a complex task? Intriguingly, within state-of-the-art deep networks, it has been observed that many single units match human-interpretable concepts that were not explicitly taught to the network: Units have been found to detect objects, parts, textures, tense, gender, context, and sentiment (1–7). Finding such meaningful abstractions is one of the main goals of deep learning (8), but the emergence and role of such concept-specific units are not well understood. Thus, we ask: How can we quantify the emergence of concept units across the layers of a network? What types of concepts are matched, and what function do they serve? When a network contains a unit that activates on trees, we wish to understand if it is a spurious correlation or if the unit has a causal role that reveals how the network models its higher-level notions about trees.

To investigate these questions, we introduce network dissection (9, 10), our method for systematically mapping the semantic concepts found within a deep convolutional neural network (CNN). The basic unit of computation within such a network is a learned convolutional filter; this architecture is the state of the art for solving a wide variety of discriminative and generative tasks in computer vision (11–19). Network dissection identifies, visualizes, and quantifies the role of individual units in a network by comparing the activity of each unit with a range of human-interpretable pattern-matching tasks such as the detection of object classes.

Previous approaches for understanding a deep network include the use of salience maps (20–27): Those methods ask where a network looks when it makes a decision. The goal of our current inquiry is different: We ask what a network is looking for and why. Another approach is to create simplified surrogate models to mimic and summarize a complex network's behavior (28–30), and another technique is to train explanation networks that generate human-readable explanations of a network (31). In contrast to those methods, network dissection aims to directly interpret the internal computation of the network itself, rather than training an auxiliary model.

We dissect the units of networks trained on two different types of tasks: image classification and image generation. In both settings, we find that a trained network contains units that correspond to high-level visual concepts that were not explicitly labeled in the training data. For example, when trained to classify or generate natural scene images, both types of networks learn individual units that match the visual concept of a "tree" even though we have never taught the network the tree concept during training.

Focusing our analysis on the units of a network allows us to test the causal structure of network behavior by activating and deactivating the units during processing. In a classifier, we use these interventions to ask whether the classification performance of a specific class can be explained by a small number of units that identify visual concepts in the scene class. For example, we ask how the ability of the network to classify an image as a ski resort is affected when removing a few units that detect snow, mountains, trees, and houses. Within a scene generation network, we ask how the rendering of objects in a scene is affected by object-specific units. How does the removal of tree units affect the appearance of trees and other objects in the output image?

Finally, we demonstrate the usefulness of our approach with two applications. We show how adversarial attacks on a classifier can be understood as attacks on the important units for a class. Also, we apply unit intervention on a generator to enable a human user to modify semantic concepts such as trees and doors in an image by directly manipulating units.

Results

Emergence of Object Detectors in a Scene Classifier. We first identify individual units that emerge as object detectors when training a network on a scene classification task. The network we analyze is a convolutional neural network (CNN) with the VGG-16

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "The Science of Deep Learning," held March 13–14, 2019, at the National Academy of Sciences in Washington, DC. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019 colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/science-of-deep-learning.

Author contributions: D.B., J.-Y.Z., H.S., A.L., B.Z., and A.T. designed research; D.B., J.-Y.Z., H.S., A.L., and B.Z. performed research; D.B. and B.Z. contributed new analytic tools; D.B. and J.-Y.Z. analyzed data; A.T. was the supervising advisor; and D.B., J.-Y.Z., and A.T. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: davidbau@csail.mit.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1907375117/-/DCSupplemental.

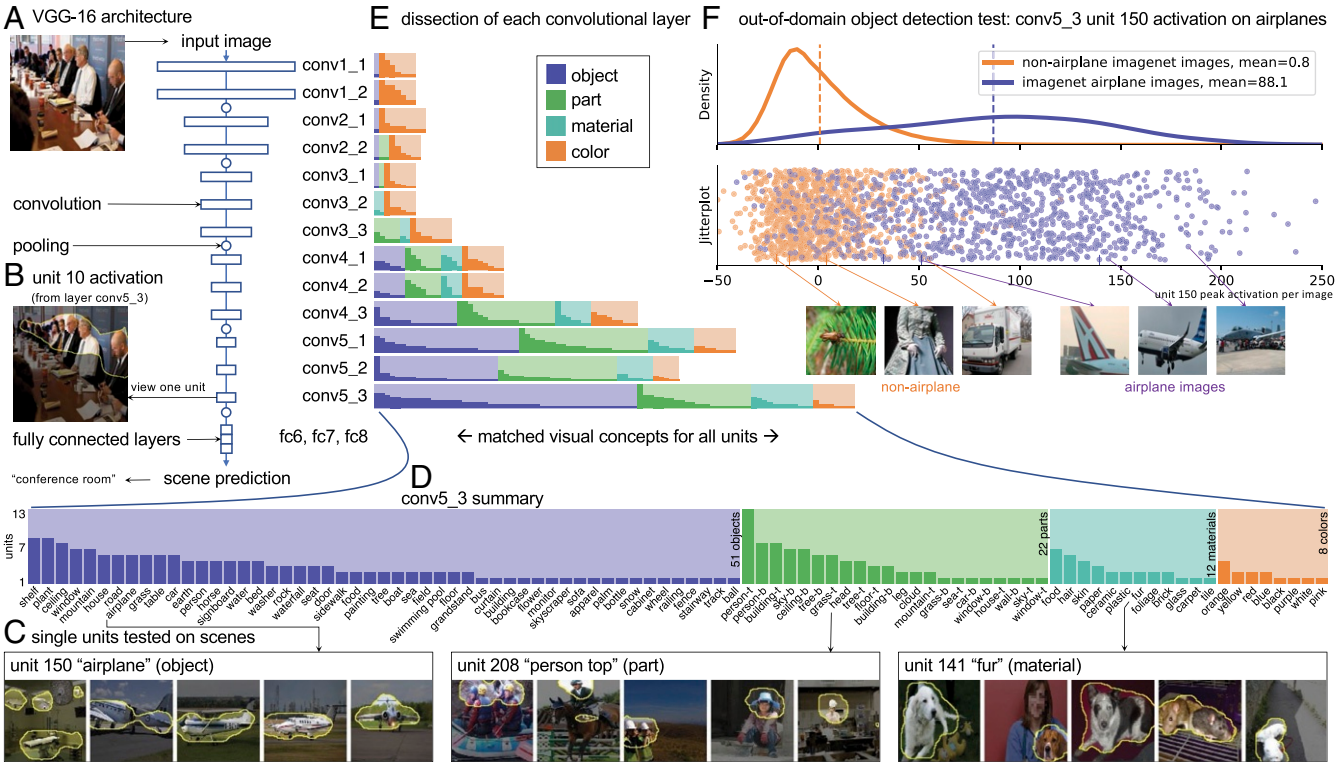

Fig. 1. The emergence of single-unit object detectors within a VGG-16 scene classifier. (*A*) VGG-16 consists of 13 convolutional layers, conv1_1 through conv5_3, followed by three fully connected layers: fc6, -7, -8. (*B*) The activation of a single filter on an input image can be visualized as the region where the filter activates beyond its top 1% quantile level. (*C*) Single units are scored by matching high-activating regions against a set of human-interpretable visual concepts; each unit is labeled with its best-matching concept and visualized with maximally activating images. (*D*) Concepts that match units in the final convolutional layer are summarized, showing a broad diversity of detectors for objects, object parts, materials, and colors. Many concepts are associated with multiple units. (*E*) Comparing all of the layers of the network reveals that most object detectors emerge at the last convolutional layers. (*F*) Although the training set contains no object labels, unit 150 emerges as an airplane object detector that activates much more strongly on airplane objects than nonairplane objects, as tested against a dataset of labeled object images not previously seen by the network. The jitter plot shows peak activations for the unit on randomly sampled 1,000 airplane and 1,000 nonairplane Imagenet images, and the curves show the kernel density estimates of these activations.

architecture (named after the Oxford Visual Geometry Group) (13) trained to classify images into 365 scene categories using the Places365 dataset, from the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory Scene Recognition Database (32). We analyze all units within the 13 convolutional layers of the network (Fig. 1*A*). *Materials and Methods* has further details on networks and datasets.

Each unit u computes an activation function $a_u(x,p)$ that outputs a signal at every image position p given a test image x. Filters with low-resolution outputs are visualized and analyzed at high-resolution positions p using bilinear up sampling. Denote by t_u the top 1% quantile level for a_u: That is, writing $\mathbb{P}_{x,p}[\cdot]$ to indicate the probability that an event is true when sampled over all positions and images, we define the threshold $t_u \equiv \max_t \mathbb{P}_{x,p}[a_u(x,p) > t] \geq 0.01$. In visualizations, we highlight the activation region $\{p \mid a_u(x,p) > t_u\}$ above the threshold. As seen in Fig. 1*B*, this region can correspond to semantics such as the heads of all of the people in the image. To identify filters that match semantic concepts, we measure the agreement between each filter and a visual concept c using a computer vision segmentation model (33) $s_c : (x,p) \to \{0,1\}$ that is trained to predict the presence of the visual concept c within image x at position p. We quantify the agreement between concept c and unit u using the intersection over union (IoU) ratio:

$$\text{IoU}_{u,c} = \frac{\mathbb{P}_{x,p}[s_c(x,p) \wedge (a_u(x,p) > t_u)]}{\mathbb{P}_{x,p}[s_c(x,p) \vee (a_u(x,p) > t_u)]}. \quad [1]$$

This IoU ratio is computed on the set of held-out validation set images. Within this validation set, each unit is scored against 1,825 segmented concepts c, including object classes, parts of objects, materials, and colors. Then, each unit is labeled with the highest-scoring matching concept. Fig. 1*C* shows several labeled concept detector units along with the five images with the highest unit activations.

When examining all 512 units in the last convolutional layer, we find many detected object classes and relatively fewer detected object parts and materials: within layer conv_3, units match 51 object classes, 22 parts, 12 materials, and eight colors. Several visual concepts such as "airplane" and "head" are matched by more than one unit. Fig. 1*D* lists every segmented concept matching units in layer conv5_3, excluding any units with IoU ratio $< 4\%$, showing the frequency of units matching each concept. Across different layers, the last convolutional layer has the largest number of object classes detected by units, while the number of object parts peaks two layers earlier, at layer conv5_1, which has units matching 28 object classes, 25 parts, nine materials, and eight colors (Fig. 1*E*). A complete visualization of all of the units of conv5_3 is provided in *SI Appendix*, as well as more detailed comparisons between layers of VGG-16, comparisons with layers of AlexNet (12) and ResNet (16), and an analysis of the texture vs. shape sensitivity of units using a stylization method based on ref. 34.

Interestingly, object detectors emerge despite the absence of object labels in the training task. For example, the aviation-related scene classes in the training set are "airfield," "airport

terminal," "hangar," "landing deck," and "runway." Scenes in these classes do not always contain airplanes, and there is no explicit airplane object label in the training set. Yet, unit 150 emerges as a detector that locates airplanes, scoring IoU = 9.0% agreement with our reference airplane segmentations in scene images. The accuracy of the unit as an airplane classifier can be further verified on Imagenet (35), a dataset that contains 1,000 object classes; its images and classes are disjoint from the Places365 training set. Imagenet contains two airplane class labels: "airliner" and "warplane," and a simple threshold on unit 150 (peak activation > 23.4) achieves 85.6% balanced classification accuracy on the task of distinguishing these airplane classes from the other object classes. Fig. 1*F* shows the distribution of activations of this unit on a sample of airplane and nonairplane Imagenet images.

Role of Units in a Scene Classifier. How does the network use the above object detector units? Studies of network compression have shown that many units can be eliminated from a network while recovering overall classification accuracy by retraining (36, 37). One way to estimate the importance of an individual unit is to examine the impact of the removal of the unit on mean network accuracy (38, 39).

To obtain a more fine-grained understanding of the causal role of each unit within a network, we measure the impact of removing each unit on the network's ability of classifying each individual scene class. Units are removed by forcing the specified unit to output zero and leaving the rest of the network intact. No retraining is done. Single-class accuracy is tested on the balanced two-way classification problem of discriminating the specified class from all of the other classes.

The relationships between objects and scenes learned by the network can be revealed by identifying the most important units for each class. For example, the four most important conv5_3 units for the class "ski resort" are shown in Fig. 2*A*: These units damage ski resort accuracy most when removed. The units detect snow, mountains, houses, and trees, all of which seem salient to ski resort scenes.

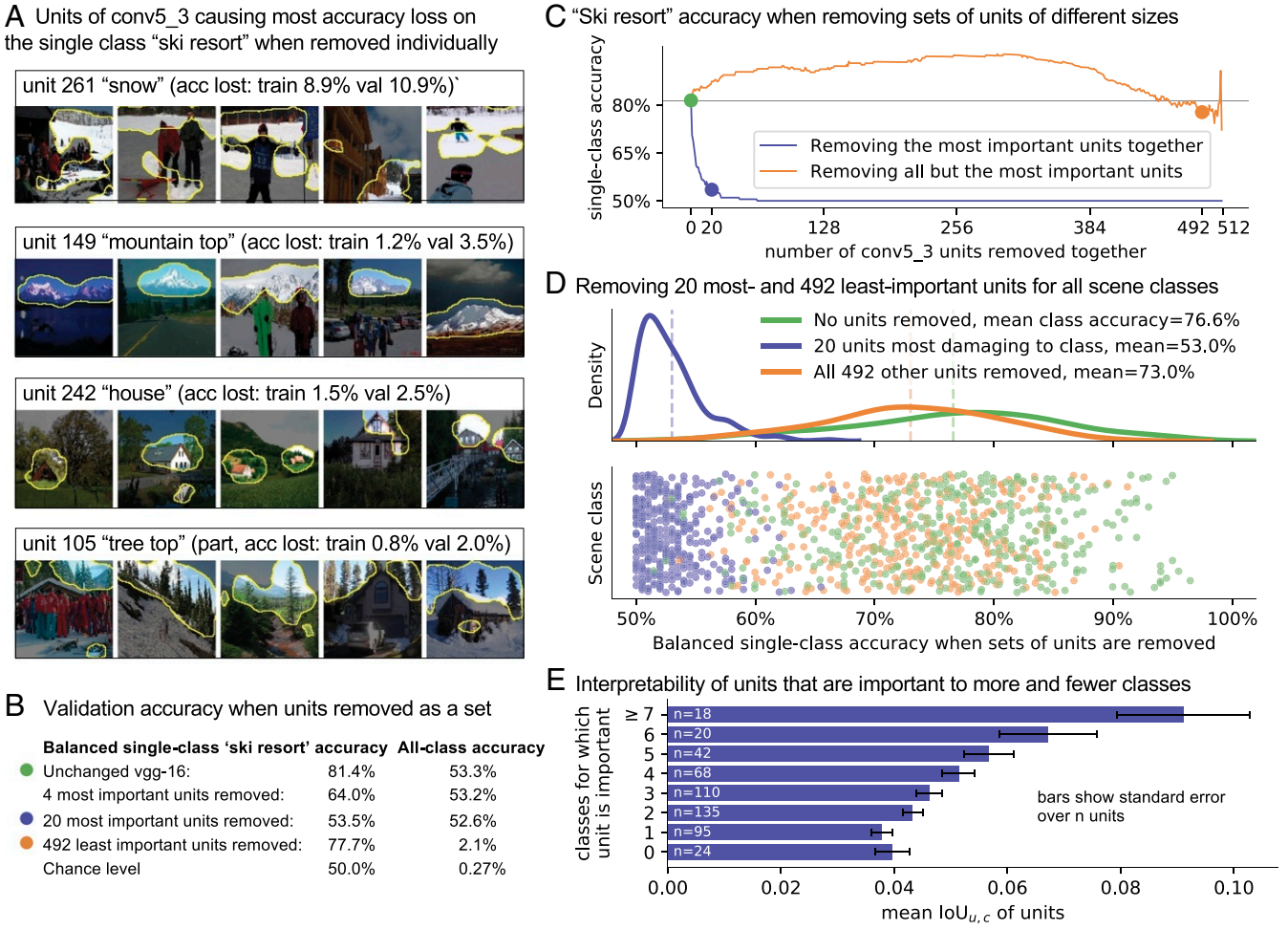

Fig. 2. A few units play important roles in classification performance. (*A*) The four conv5_3 units cause the most damage to balanced classification accuracy for ski resort when each unit is individually removed from the network; dissection reveals that these most-important units detect visual concepts that are salient to ski resorts. Accuracy lost (acc lost) is measured on both training data and held-out validation (val) data. (*B*) When the most-important units to the class are removed all together, balanced single-class accuracy drops to near-chance levels. When the 492 least-important units in conv5_3 are removed all together (leaving only the 20 most-important units), accuracy remains high. (*C*) The effect on ski resort prediction accuracy when removing sets of units of successively larger sizes. These units are sorted in ascending and descending order of individual unit's impact on accuracy. (*D*) Repeating the experiment for each of 365 scene classes. Each point plots single-class classification accuracy in one of three settings: the original network, the network after removing the 20 units most important to the class, and with all conv5_3 units removed except the 20 most-important ones. On the *y* axis, classes are ordered alphabetically. (*E*) The relationship between unit importance and interpretability. Units that are among the top four important units for more classes are also closer matches for semantic concepts as measured by IoU$_{u,c}$.

To test whether the ability of the network to classify ski resorts can be attributed to just the most important units, we remove selected sets of units. Fig. 2*B* shows that removing just these 4 (of 512) units reduces the network's accuracy at discriminating ski resort scenes from 81.4 to 64.0%, and removing the 20 most important units in conv5_3 reduces class accuracy further to 53.5%, near chance levels (chance is 50%), even though classification accuracy over all scene classes is hardly affected (changing from 53.3 to 52.6%, where chance is 0.27%). In contrast, removing the 492 least-important units (leaving only the 20 most important units in conv5_3) has only a small impact on accuracy for the specific class, reducing ski resort accuracy by only 3.7% (to 77.7%). Of course, removing so many units damages the ability of the network to classify other scene classes: Removing the 492 least-important units reduces all-class accuracy to 2.1% (chance is 0.27%).

The effect of removing varying numbers of most-important and least-important units upon ski resort accuracy is shown in Fig. 2*C*. To avoid overfitting to the evaluation data, we rank the importance of units according to their individual impact on single-class ski resort accuracy on the training set, and the plotted impact of removing sets of units is evaluated on the held-out validation set. The network can be seen to derive most of its performance for ski resort classification from just the most important units. Single-class accuracy can even be improved by removing the least important units; this effect is further explored in *SI Appendix*.

This internal organization, in which the network relies on a small number of important units for most of its accuracy with respect to a single output class, is seen across all classes. Fig. 2*D* repeats the same experiment for each of the 365 scene classes. Removing the 20 most important conv5_3 units for each class reduces single-class accuracy to 53.0% on average, near chance levels. In contrast, removing the 492 least important units only reduces single-class accuracy by an average of 3.6%, just a slight reduction. We conclude that the emergent object detection done by units of conv5_3 is not spurious: Each unit is important to a specific set of classes, and the object detectors can be interpreted as decomposing the network's classification of individual scene classes into simpler subproblems.

Why do some units match interpretable concepts so well, while other units do not? The data in Fig. 2*E* show that the most interpretable units are those that are important to many different output classes. Units that are important to only one class (or none) are less interpretable, measured by IoU. We further find that important units are predominantly positively correlated with their associated classes, and different combinations of units provide support for each class. Measurements of unit-class correlations and examples of overlapping combinations of important units are detailed in *SI Appendix*.

Does the emergence of interpretable units such as airplane, snow, and tree detectors depend on having training set labels that divide the visual world into hundreds of scene classes? Perhaps the taxonomy of scenes encodes distinctions that are necessary to learn about objects. Or is it possible for a network to infer such concepts from the visual data itself? To investigate this question, we next conduct a similar set of experiments on networks trained to solve unsupervised tasks.

Emergence of Object Detectors in a Generative Adversarial Network. A generative adversarial network (GAN) learns to synthesize random realistic images that mimic the distribution of real images in a training set (14). Architecturally, a trained GAN generator is the reverse of a classifier, producing a realistic image from a random input latent vector. Unlike classification, it is an unsupervised setting: No human annotations are provided to a GAN, so the network must learn the structure of the images by itself.

Remarkably, GANs have been observed to learn global semantics of an image: For example, interpolating between latent vectors can smoothly transform the layout of a room (40) or change the texture of an object (41). We wish to understand whether the GAN also learns to decompose local semantics, for example, if the internal units represent the generation of a scene as a hierarchy of meaningful parts.

We test a Progressive GAN architecture (19) trained to imitate LSUN kitchen images (42). This network architecture consists of 15 convolutional layers, as shown in Fig. 3*A*. Given a 512-dimensional vector sampled from a multivariate Gaussian distribution, the network produces a 256×256 realistic image after processing the data through the 15 layers. As with a classifier network, each unit is visualized by showing the regions where the filter activates above its top 1% quantile level, as shown in Fig. 3*B*. Importantly, causality in a generator flows in the opposite direction as a classifier: When unit 381 activates on lamp shades in an image, it is not detecting objects in the image because the filter activation occurs before the image is generated. Instead, the unit is part of the computation that ultimately renders the objects.

To identify the location of units in the network that are associated with object classes, we apply network dissection to the units of every layer of the network. In this experiment, the reference segmentation models and thresholds used are the same as those used to analyze the VGG-16 classifier. However, instead of analyzing agreement with objects that appear in the input data, we analyze agreement with segmented objects found in the generated output images. As shown in Fig. 3*C*, the largest number of emergent concept units does not appear at the edge of the network as we saw in the classifier but in the middle: Layer 5 has units that match the largest number of distinct object and part classes.

Fig. 3*D* shows each object, part, material, and color that matches a unit in layer 5 with $\text{IoU} > 4\%$. This layer contains 19 object-specific units, 41 units that match object parts, one material, and six color units. As seen in the classification network, visual concepts such as "oven" and "chair" match many units. Different from the classifier, more object parts are matched than whole objects.

In Fig. 3*D*, individual units show a wide range of visual diversity: The units do not appear to rigidly match a specific pixel pattern but rather, different appearances for a particular class: For example, various styles of ovens or different colors and shapes of kitchen stools.

In Fig. 3*F*, we apply the window-specific unit 314 as an image classifier. We find a strong gap between the activation of the unit when a large window is generated and when no large window is generated. Furthermore, a simple threshold (peak activation > 8.03) can achieve a 78.2% accuracy in predicting whether the generated image will have a large window or not. Nevertheless, the distribution density curve reveals that images that contain large windows can be often generated without activating unit 314. Two such samples are shown in Fig. 3*G*. These examples suggest that other units could potentially synthesize windows.

Role of Units in a GAN. The correlations between units and generated object classes are suggestive, but they do not prove that the units that correlate with an object class actually cause the generator to render instances of the object class. To understand the causal role of a unit in a GAN generator, we test the output of the generator when sets of units are directly removed or activated.

We first remove successively larger sets of tree units from a Progressive GAN (19) trained on LSUN church scenes (42). We rank units in layer 4 according to $\text{IoU}_{u,\text{tree}}$ to identify the most tree-specific units. When successively larger sets of these tree units are removed from the network, the GAN generates

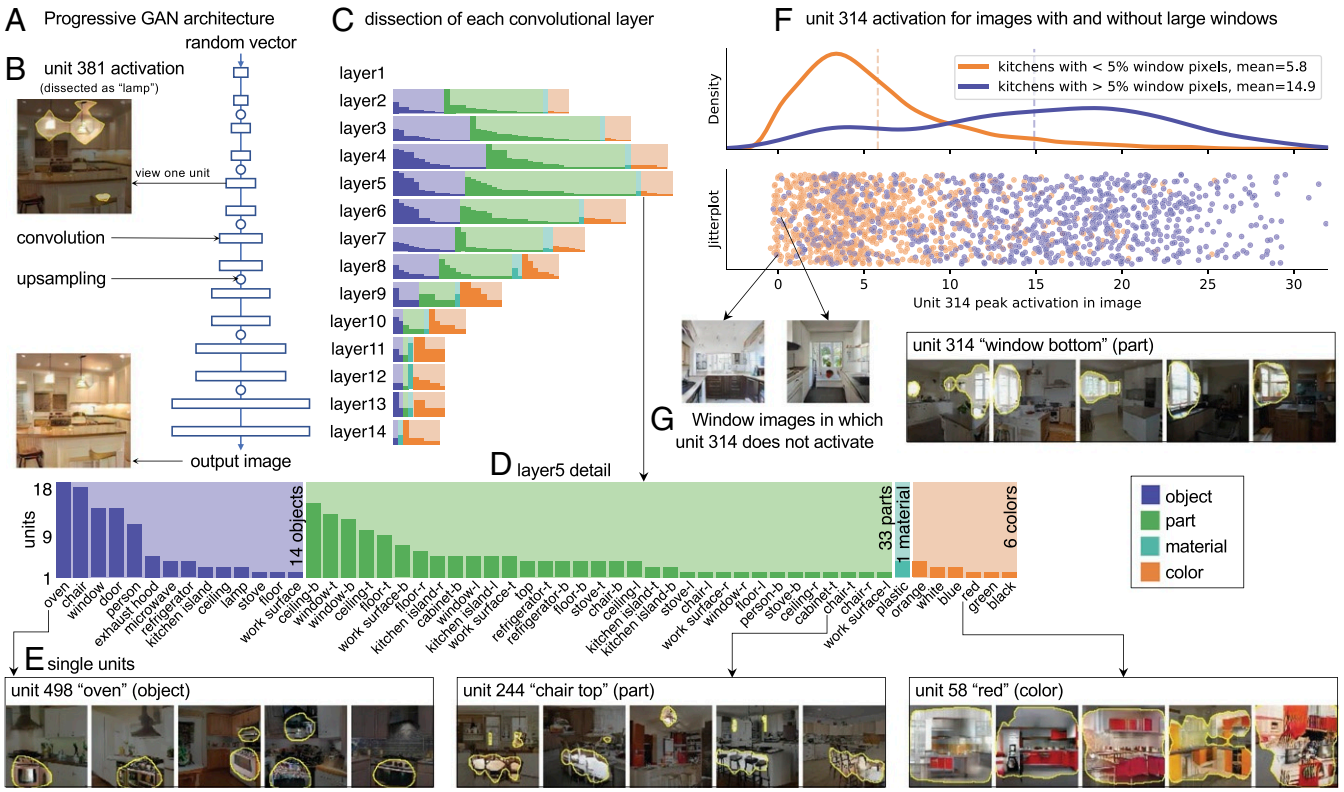

Fig. 3. The emergence of object- and part-specific units within a Progressive GAN generator (19). (*A*) The analyzed Progressive GAN consists of 15 convolutional layers that transform a random input vector into a synthesized image of a kitchen. (*B*) A single filter is visualized as the region of the output image where the filter activates beyond its top 1% quantile level; note that the filters are all precursors to the output. (*C*) Dissecting all of the layers of the network shows a peak in object-specific units at layer 5 of the network. (*D*) A detailed examination of layer 5 shows more part-specific units than objects and many visual concepts corresponding to multiple units. (*E*) Units do not correspond to exact pixel patterns: A wide range of visual appearances for ovens and chairs is generated when an oven or chair part unit is activated. (*F*) When a unit specific to window parts is tested as a classifier, on average the unit activates more strongly on generated images that contain large windows than images that do not. The jitter plot shows the peak activation of unit 314 on 800 generated images that have windows larger than 5% of the image area as estimated by a segmentation algorithm and 800 generated images that do not. (*G*) Some counterexamples: images for which unit 314 does not activate but where windows are synthesized nevertheless.

images with fewer and smaller trees (Fig. 4*A*). Removing the 20 most tree-specific units reduces the number of tree pixels in the generated output by 53.3%, as measured over 10,000 randomly generated images.

When tree-specific units are removed, the generated images continue to look similarly realistic. Although fewer and smaller trees are generated, other objects such as buildings are unchanged. Remarkably, parts of buildings that were occluded by trees are hallucinated, as if removing the trees reveals the walls and windows behind them (Fig. 4*B*). The generator appears to have computed more details than are necessary to render the final output; the details of a building that are hidden behind a tree can only be revealed by suppressing the generation of the tree. The appearance of such hidden details strongly suggests that the GAN is learning a structured statistical model of the scene that extends beyond a flat summarization of visible pixel patterns.

Units can also be forced on to insert new objects into a generated scene. We use $\mathrm{IoU}_{u,\mathrm{door}}$ to find the 20 most door-specific units identified in layer 4 of the same outdoor church GAN. At tested locations, the activations for this set of 20 units are all forced to their high t_u value. Fig. 4*C* shows the effect of applying this procedure to activate 20 door units at two different locations in two generated images. Although the same intervention is applied to all four cases, the doors obtained in each situation are different: In cases 1 to 3, the newly synthesized door has a size, style, and location that is appropriate to the scene context. In case 4, where door units are activated on a tree, no new door is added to the image.

Fig. 4*D* quantifies the context sensitivity of activating door units in different locations. In 10,000 randomly generated images, the same 20-door-unit activation is tested at every feature map location, and the number of newly synthesized door pixels is evaluated using a segmentation algorithm. Doors can be easily added in some locations, such as in buildings and especially on top of an existing window, but it is nearly impossible to add a door into trees or in the sky. By learning to solve the unsupervised image generation problem, a GAN has learned units for emergent objects such as doors and trees. It has also learned a computational structure over those units that prevents it from rendering nonsensical output, such as a door in the sky or a door in a tree.

Applications

We now turn to two applications enabled by our understanding of the role of units: understanding attacks on a classifier and interactively editing a photo by activating units of a GAN.

Analyzing Adversarial Attack of a Classifier. The sensitivity of image classifiers to adversarial attacks is an active research area (43–46). To visualize and understand how an attack works, we can examine the effects on important object detector units. In Fig. 5*A*, a correctly classified ski resort image is attacked to the target "bedroom" by the Carlini–Wagner optimization method

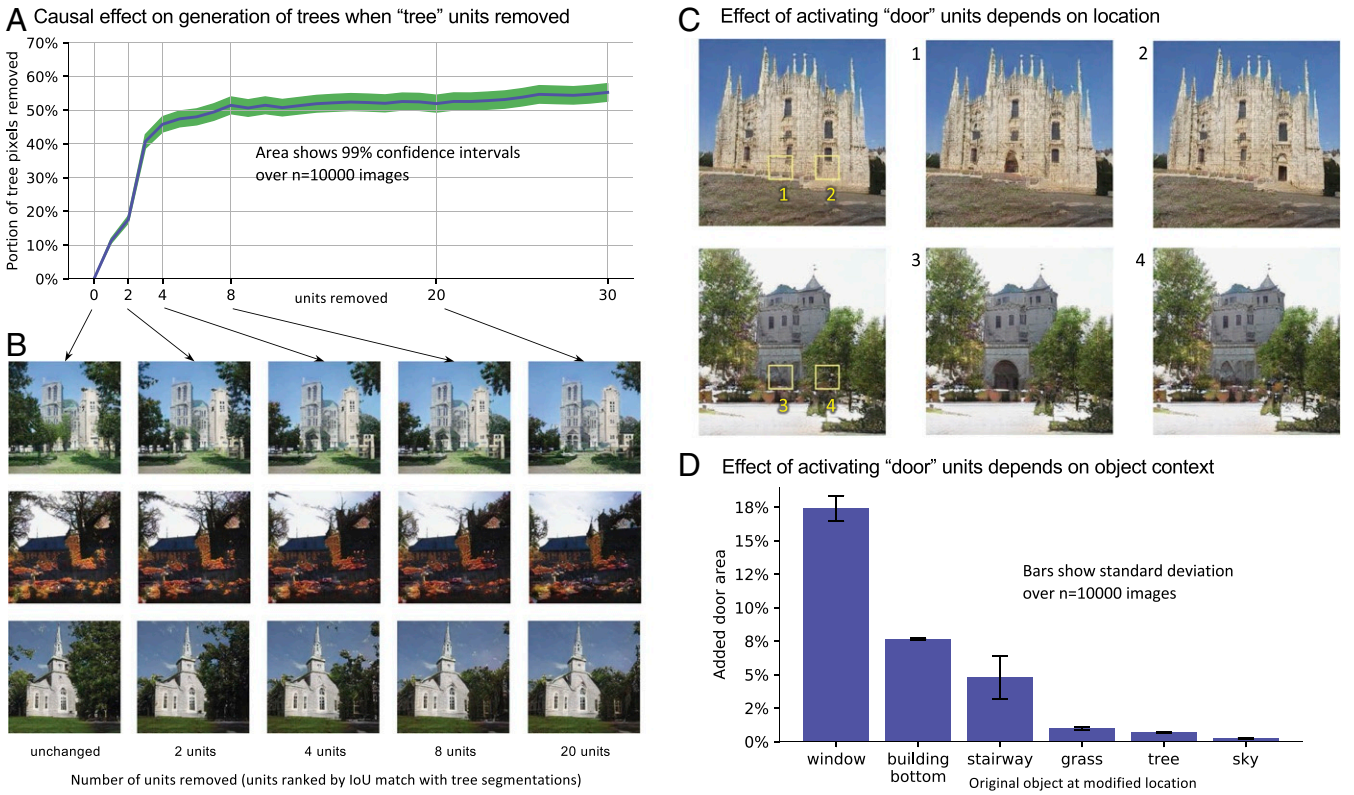

Fig. 4. The causal effect of altering units within a GAN generator. (*A*) When successively larger sets of units are removed from a GAN trained to generate outdoor church scenes, the tree area of the generated images is reduced. Removing 20 tree units removes more than half the generated tree pixels from the output. (*B*) Qualitative results: Removing tree units affects trees while leaving other objects intact. Building parts that were previously occluded by trees are rendered as if revealing the objects that were behind the trees. (*C*) Doors can be added to buildings by activating 20 door units. The location, shape, size, and style of the rendered door depend on the location of the activated units. The same activation levels produce different doors or no door at all (case 4) depending on locations. (*D*) Similar context dependence can be seen quantitatively: doors can be added in reasonable locations, such as at the location of a window, but not in abnormal locations, such as on a tree or in the sky.

(45, 47). The adversarial algorithm computes a small perturbation, which when added to the original, results in a misclassified image that is visually indistinguishable from the original image. To understand how the attack works, we examine the four most important units to the ski resort class and the four most important units to the bedroom class. Fig. 5*B* visualizes changes in the

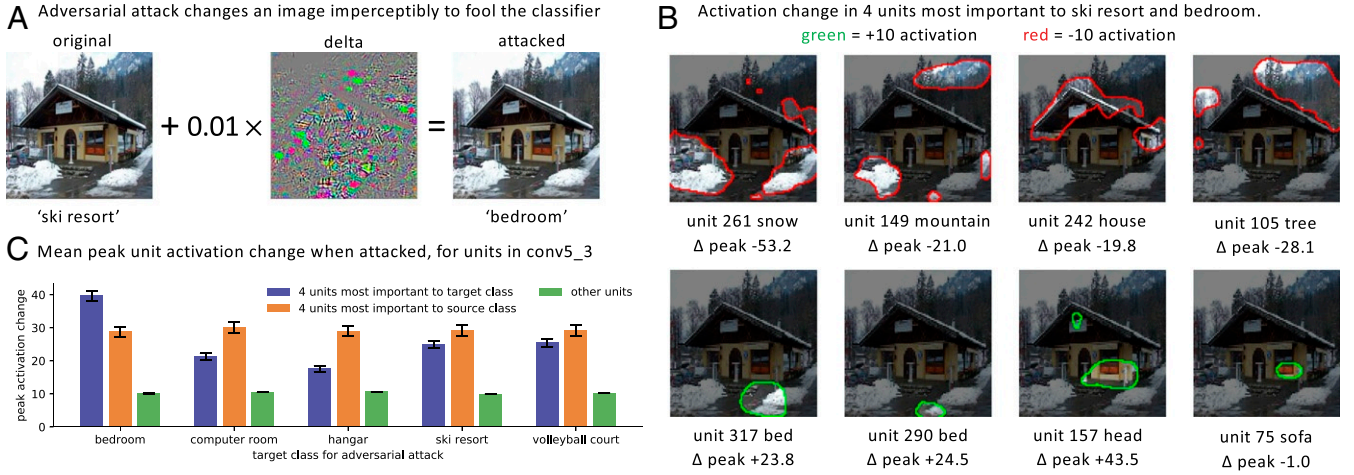

Fig. 5. Application: Visualizing an adversarial attack. (*A*) The test image is correctly labeled as a ski resort, but when an adversarial perturbation is added, the visually indistinguishable result is classified as a bedroom. (*B*) Visualization of the attack on the four most important units to the ski resort class and the four units most important to the bedroom class. Areas of maximum increase and decrease are shown; Δpeak indicates the change in the peak activation level for the unit. (*C*) Over 1,000 images attacked to misclassify images to various incorrect target classes. The units that are changed most are those that dissection has identified as most important to the source and target classes. Mean absolute value change in peak unit activation is graphed, with 99% CIs shown.

activations for these units between the original image and the adversarial image. This reveals that the attack has fooled the network by reducing detection of snow, mountain, house, and tree objects and by increasing activations of detectors for beds, person heads, and sofas in locations where those objects do not actually exist in the image. Fig. 5*C* shows that, across many images and classes, the units that are most changed by an attack are the few units that are important to a class.

Semantic Paint Using a GAN. Understanding the roles of units within a network allows us to create a human interface for controlling the network via direct manipulation of its units. We apply this method to a GAN to create an interactive painting application. Instead of painting with a palette of colors, the application allows painting with a palette of high-level object concepts. Each concept is associated with 20 units that maximize $\mathrm{IoU}_{u,c}$ for the concept u. Fig. 6*A* shows our interactive interface. When a user adds brush strokes with a concept, the units for the concept are activated (if the user is drawing) or zeroed (if the user is erasing). Fig. 6*B* shows typical results after the user adds an object to the image. The GAN deals with the pixel-level details of how to add objects while keeping the scene reasonable and realistic. Multiple changes in a scene can be composed for creative effects. Movies of image editing demonstrations are included in Movies S1–S3; online demonstrations are also available at the website http://gandissect.csail.mit.edu.

Discussion

Simple measures of performance, such as classification accuracy, do not reveal how a network solves its task: Good performance can be achieved by networks that have differing sensitivities to shapes, textures, or perturbations (34, 48).

To develop an improved understanding of how a network works, we have presented a way to analyze the roles of individual network units. In a classifier, the units reveal how the network decomposes the recognition of specific scene classes into particular visual concepts that are important to each scene class. Additionally, within a generator, the behavior of the units reveals contextual relationships that the model enforces between classes of objects in a scene.

Network dissection relies on the emergence of disentangled, human-interpretable units during training. We have seen that many such interpretable units appear in state-of-the-art models, both supervised and unsupervised. How to train better disentangled models is an open problem that is the subject of ongoing efforts (49–52).

We conclude that a systematic analysis of individual units can yield insights about the black box internals of deep networks. By observing and manipulating units of a deep network, it is possible to understand the structure of the knowledge that the network has learned and to build systems that help humans interact with these powerful models.

Materials and Methods

Datasets. Places365 (53, 54) consists of 1.80 million photographic images, each labeled with 1 of 365 scene classes. The dataset also includes 36,500 labeled validation images (100 per class) that are not used for training. Imagenet (35, 55) consists of 1.28 million photographic images, each focused on a single main object and labeled with 1 of 1,000 object classes. LSUN is a dataset with a large number of 256 × 256 images in a few classes (42, 56). LSUN kitchens consist of 2.21 million indoor kitchen photographs, and LSUN outdoor churches consist of 1.26 million photographs of church building exteriors. Recognizable people in dataset images have been anonymized by pixelating faces in visualizations.

Tested Networks. We analyze the VGG-16 classifier (13) trained by the Places365 authors (32) to classify Places365 images (57). The network achieves classification accuracy of 53.3% on the held-out validation set (chance is 0.27%). The 13 convolutional layers of VGG-16 are divided into five groups. The layers in the first group contain 32 units that process image data at the full 224 × 224 resolution; at each successive group, the feature depth is doubled, and the feature maps are pooled to halve the resolution, so that at the final stage that includes conv5_1 and conv5_3, the layers contain 512 units at 14 × 14 resolution. The GAN models that we analyze are trained by the Progressive GAN authors (19, 58). The models are configured to generate 256 × 256 output images using 15 convolutional layers divided into eight groups, starting with 512 units in each layer at 4 × 4 resolution and doubling resolution at each successive group, so that layer 4 has 8 × 8 resolution and 512 units and layer 5 has 16 × 16 resolution and 512 units. Unit depth is halved in each group after layer 6, so that the 14th layer has 32 units and 256 × 256 resolution. The 15th layer (which is not pictured in Fig. 3*A*) produces a three-channel red–green–blue image. Code and pretrained weights for all tested networks are available at the GitHub and project website for this paper (59).

Reference Segmentation. To locate human-interpretable visual concepts within large-scale datasets of images, we use the Unified Perceptual Parsing image segmentation network (33) trained on the ADE20K scene dataset (53, 60) and an assignment of numerical color values to color names (61). The segmentation algorithm achieves mean IoU of 23.4% on objects, 28.8% on parts, and 54.2% on materials. To further identify units that specialize in object parts, we expand each object class into four additional object part classes, which denote the top, bottom, left, or right half of the bounding box of a connected component. Our reference segmentation algorithm can detect 335 object classes, 1,452 object parts, 25 materials, and 11 colors.

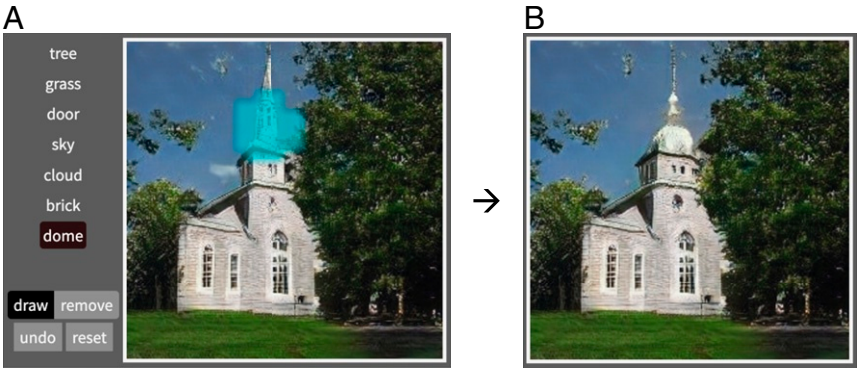

Fig. 6. Application: Painting by manipulating GAN neurons. (*A*) An interactive interface allows a user to choose several high-level semantic visual concepts and paint them onto an image. Each concept corresponds to 20 units in the GAN. (*B*) After the user adds a dome in the specified location, the result is a modified image in which a dome has been added in place of the original steeple. After the user's high-level intent has been expressed by changing 20 dome units, the generator automatically handles the pixel-level details of how to fit together objects to keep the output scene realistic.

Data Availability. The code, trained model weights, and datasets needed to reproduce the results in this paper are public and available to download from GitHub at https://github.com/davidbau/dissect and at the project website at https://dissect.csail.mit.edu/data/.

ACKNOWLEDGMENTS. We thank Aditya Khosla, Aude Oliva, William Peebles, Jonas Wulff, Joshua B. Tenenbaum, and William T. Freeman for their advice and collaboration. Also, we are grateful for the support of the Massachusetts Institute of Technology–IBM Watson Artificial Intelligence Lab, Defense Advanced Research Projects Agency Explainable Artificial Intelligence (DARPA XAI) Program FA8750-18-C-0004, NSF Grant 1524817 on Advancing Visual Recognition with Feature Visualizations, NSF Grant BIGDATA 1447476, Grant RTI2018-095232-B-C22 from the Spanish Ministry of Science, Innovation and Universities (to A.L.), Early Career Scheme of Hong Kong Grant 24206219 (to B.Z.), and a hardware donation from Nvidia.

1. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene CNNs. arXiv:1412.6856 (22 December 2014).
2. M. D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks" in *European Conference on Computer Vision* (Springer, Berlin, Germany, 2014), pp. 818–833.
3. A. Mahendran, A. Vedaldi, "Understanding deep image representations by inverting them" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, NY, 2015), pp. 5188–5196.
4. C. Olah *et al.*, The building blocks of interpretability. *Distill* **3**, e10 (2018).
5. A. Bau *et al.*, Identifying and controlling important neurons in neural machine translation. https://openreview.net/pdf?id=H1z-PsR5KX. Accessed 24 August 2020.
6. A. Karpathy, J. Johnson, L. Fei-Fei, Visualizing and understanding recurrent networks. arXiv:1506.02078 (5 June 2015).
7. A. Radford, R. Jozefowicz, I. Sutskever, Learning to generate reviews and discovering sentiment. arXiv:1704.01444 (6 April 2017).
8. Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intelligence* **35**, 1798–1828 (2013).
9. B. Zhou, D. Bau, A. Oliva, A. Torralba, Interpreting deep visual representations via network dissection. arXiv:1711.05611 (26 June 2018).
10. D. Bau *et al.*, Gan dissection: Visualizing and understanding generative adversarial networks. https://openreview.net/pdf?id=Hyg_X2C5FX. Accessed 24 August 2020.
11. Y. LeCun, Y. Bengio, "Convolutional networks for images, speech, and time series" in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. (MIT Press, Cambridge, MA, 1995), vol. 3361, pp. 255–258.
12. A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks" in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2012), pp. 1097–1105.
13. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (4 September 2014).
14. I. Goodfellow *et al.*, "Generative adversarial nets" in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2014), pp. 2672–2680.
15. O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and tell: A neural image caption generator" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, NY, 2015), pp. 3156–3164.
16. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, NY, 2016), pp. 770–778.
17. P. Isola, J. Y. Zhu, T. Zhou, A. A. Efros, "Image-to-image translation with conditional adversarial networks" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (IEEE, New York, NY, 2017), pp. 1125–1134.
18. J. Y. Zhu, T. Park, P. Isola, A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, NY, 2017), pp. 2223–2232.
19. T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation. https://openreview.net/pdf?id=Hk99zCeAb. Accessed 24 August 2020.
20. S. Bach *et al.*, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* **10**, e0130140 (2015).
21. T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, NY, 2016), pp. 117–126.
22. R. C. Fong, A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation" in *International Conference on Computer Vision* (IEEE, New York, NY, 2017), pp. 3429–3437.
23. S. M. Lundberg, S. I. Lee, "A unified approach to interpreting model predictions" in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2017), pp. 4765–4774.
24. R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, NY, 2017), pp. 618–626.
25. M. Sundararajan, A. Taly, Q. Yan, "Axiomatic attribution for deep networks" in *Proceedings of the 34th International Conference on Machine Learning* (JMLR, 2017), **vol. 70**, pp. 3319–3328.
26. D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: Removing noise by adding noise. arXiv:1706.03825 (12 June 2017).
27. V. Petsiuk, A. Das, K. Saenko, "Rise: Randomized input sampling for explanation of black-box models" in *British Machine Vision Conference* (BMVA Press, Malvern, UK, 2018).
28. M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier" in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2016), pp. 1135–1144.
29. B. Kim, J. Gilmer, F. Viegas, U. Erlingsson, M. Wattenberg, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). arXiv:1711.11279 (7 June 2018).
30. A. Koul, A. Fern, S. Greydanus, Learning finite state representations of recurrent policy networks. https://openreview.net/pdf?id=S1gOpsCctm. Accessed 24 August 2020.
31. L. A. Hendricks *et al.*, "Generating visual explanations" in *European Conference on Computer Vision* (Springer, Berlin, Germany, 2016), pp. 3–19.
32. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, "Learning deep features for scene recognition using places database" in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2014), pp. 487–495.
33. T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, "Unified perceptual parsing for scene understanding" in *Proceedings of the European Conference on Computer Vision* (Springer, Berlin, Germany, 2018), pp. 418–434.
34. R. Geirhos *et al.*, Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231 (14 January 2019).
35. J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, NY, 2009), pp. 248–255.
36. W. Wen, C. Wu, Y. Wang, Y. Chen, H. Li, "Learning structured sparsity in deep neural networks" in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2016), pp. 2074–2082.
37. H. Li, A. Kadav, I. Durdanovic, H. Samet, H. P. Graf, Pruning filters for efficient convnets. https://openreview.net/pdf?id=rJqFGTslg. Accessed 24 August 2020.
38. A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, M. Botvinick, On the importance of single directions for generalization. arXiv:1803.06959 (22 May 2018).
39. B. Zhou, Y. Sun, D. Bau, A. Torralba, Revisiting the importance of individual units in CNNs via ablation. arXiv:1806.02891 (7 June 2018).
40. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 (19 November 2015).
41. J. Y. Zhu, P. Krähenbühl, E. Shechtman, A. A. Efros, "Generative visual manipulation on the natural image manifold" in *European Conference on Computer Vision* (Springer, Berlin, Germany, 2016), pp. 597–613.
42. F. Yu *et al.*, LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv:1506.03365 (4 June 2016).
43. C. Szegedy *et al.*, Intriguing properties of neural networks. arXiv:1312.6199 (21 December 2013).
44. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. arXiv:1412.6572 (20 December 2014).
45. N. Carlini, D. Wagner, "Towards evaluating the robustness of neural networks in 2017" in *IEEE Symposium on Security and Privacy (SP)* (IEEE, 2017), pp. 39–57.
46. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks. https://openreview.net/pdf?id=rJzIBfZAb. Accessed 24 August 2020.
47. J. Rauber, W. Brendel, M. Bethge, Foolbox: A python toolbox to benchmark the robustness of machine learning models. arXiv:1707.04131 (20 March 2018).
48. A. Ilyas *et al.*, Adversarial examples are not bugs, they are features. arXiv:1905.02175 (12 August 2019).
49. X. Chen *et al.*, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets" in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2016), pp. 2172–2180.
50. I. Higgins *et al.*, β-vae: Learning basic visual concepts with a constrained variational framework. https://openreview.net/pdf?id=Sy2fzU9gl. Accessed 24 August 2020.
51. Q. Zhang, Y. Nian Wu, S. C. Zhu, Interpretable convolutional neural networks. arXiv:1710.00935 (2 October 2017).
52. A. Achille, S. Soatto, Emergence of invariance and disentanglement in deep representations. *JMLR* **19**, 1947–1980 (2018).
53. B. Zhou *et al.*, "Scene parsing through ade20k dataset" in *Computer Vision and Pattern Recognition* (IEEE, 2017).
54. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places365-Standard. http://places2.csail.mit.edu/download.html. Accessed 1 August 2020.
55. J. Deng *et al.*, ImageNet ILSVRC 2012 data set. http://www.image-net.org/download-images. Accessed 1 August 2020.
56. F. Yu *et al.*, LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. https://www.yf.io/p/lsun. Accessed 1 August 2020.
57. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Data from "Pre-trained CNN models on Places365-Standard." GitHub. https://github.com/CSAILVision/places365. Accessed 1 August 2020.
58. T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs. GitHub. https://github.com/tkarras/progressive_growing_of_gans. Accessed 1 August 2020.
59. D. Bau *et al.*, Code for understanding the role of individual units in a deep neural network. Github. https://github.com/davidbau/dissect. Deposited 24 August 2020.
60. B. Zhou *et al.*, ADE20K full dataset. https://groups.csail.mit.edu/vision/datasets/ADE20K/. Accessed 1 August 2020.
61. J. Van De Weijer, C. Schmid, J. Verbeek, D. Larlus, Learning color names for real-world applications. *IEEE Trans. Image Process.* **18**, 1512–1523 (2009).