Chapter IX. Towards a world with intelligent machines that can see

Imagine a world where machines can truly see and interpret the visual world around us. A world where machines can pass the vision Turing test.

9.1. The vision Turing test

Alan Turing (1912-1954) was one of the greatest minds of the twentieth century and a pioneer in the development of the theory of computer science. In this seminal 1950 study, he proposed the "Imitation game", whereby a series of questions is posed both to a human and to a computer (Turing, 1950). Turing proposed that if we cannot distinguish which answers came from the human and which ones came from the computer, then we should call the computer intelligent. In the domain of vision, we imagine that we present the human or the computer with an image (or a video) and we are allowed to ask *any* question about the image. Again, if we cannot distinguish whether the answers come from the human or from the computer, we can claim victory, we can claim that we have solved the problem of vision, at least to human levels.

There are many visual problems where computers are already significantly better than humans. A simple example of such a problem is the ability to read bar codes, such as the ones used in a supermarket to label each item. In most supermarkets around the globe, there is still a need for a human to turn the product, locate the bar code, and position it in such a way that the scanner will be able to read it. This minimal human intervention will probably vanish soon. The task may seem somewhat limited: the number of possible "questions" about these images is rather limited. And part of the challenging invariance problem is solved by the human by positioning the image in the right place.

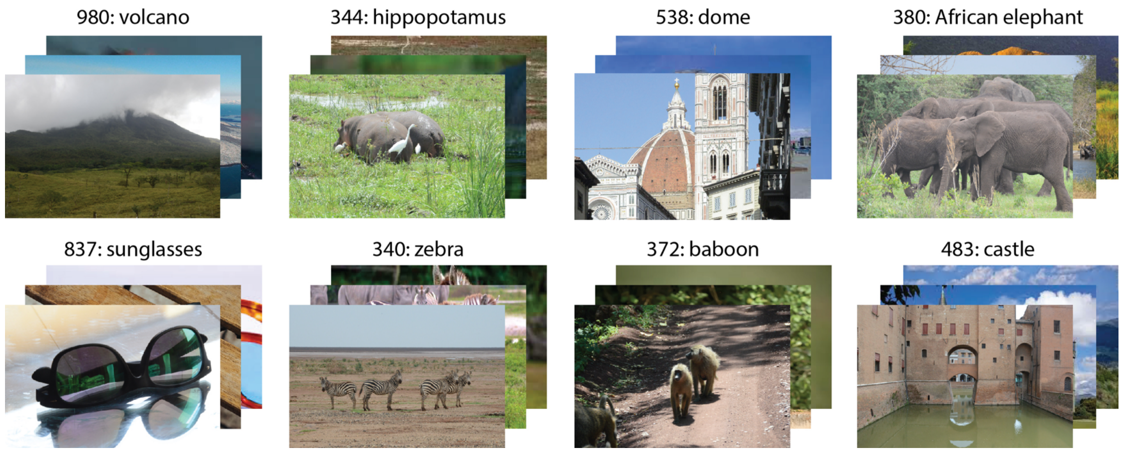
9.2. Computer vision competitions

There has been significant progress in a variety of image categorization tasks in the Computer Vision community. This progress has been fueled by a combination of increase in computational resources, access to a large number of digital images, and interesting competitions in academic conferences.

The last decade has seen an explosion in the number of digital images available on the web. In 2018, it is estimated that users upload on the order of a few billion digital images every day. In addition, many users provide more and more content in the form of "tags", brief captions, "likes" and other commentary. Every minute, humans take more photos than ever existed in total 100 years ago. There is also a rapid increase in the amount of video material being uploaded. In parallel to the availability of imagery, there are also accessible platforms such as Mechanical Turk where users can answer queries on images for a small fee, leading to more content annotation and labels. Images, content, and the concomitant exponential growth in computational power, have opened the doors to use networks with millions of tunable parameters for recognition tasks.

A typical example is the "ImageNet" large scale visual recognition challenge (Russakovsky et al., 2014). This dataset consists of color images from the web, each one associated with a label. In a typical instantiation of this competition, those labels can be any one of 1,000 classes including "goldfish", "coffee mug", "power drill", or "strawberry" (**Figure 9.1**). There are a few thousand examples of each class. The fact that the images are downloaded from the web is a blessing and a curse. A blessing because they encompass a wide diversity of image properties where the target object can appear in multiple positions, at multiple scales, rotations, colors, illumination, degrees of occlusion, etc. To some coarse approximation, this may reflect the natural distribution of objects in the world. This is not exactly true because those images are filtered through the lenses and biases of human photographers. For example, there are probably very few images of a hippopotamus that are at 45 degrees and in the middle of the night. Images taken from the web are also a curse because of their uncontrolled nature and the large number of other somewhat miscellaneous contextual factors that contribute to classification. For example, in the 3 pictures of "Domes" in Figure 9.1 (top row, third column), the pixels in the upper left are mostly blue. It seems likely that when people take pictures of Domes, they are set against the sky and there will be a higher propensity of blue in the top. In contrast, none of the "Baboon" pictures (bottom row, third column) contain blue at the top. Blue at the top is not a unique identifying feature of Domes, though. Many other pictures also contain blue at the top (e.g. volcanos, elephants, castles, and even sunglasses in the example below). There are also probably lots of pictures of Domes without blue at the top and there are probably pictures of baboons with blue at the top. The point is that there are lots of correlations in the images that are only tangentially related to the object labels themselves. Depending on the particular task and objective, these contextual correlations can represent a confound or a useful property. Another curious property of this particular dataset is that several of those categories are rather intriguing. In fact, there are many category labels that I would have to look up in the dictionary to figure out what they are (e.g. tench, junco) and many of those 1,000 classes correspond to rather specialized and refined groups of animals (how many humans can distinguish between the whiptail lizard, the alligator lizard, the green lizard, the komodo lizard, and the frilled lizard?). Yet, computers are trained to recognize these categories from scratch, and the distinction between whiptail lizards and frilled lizards may be as arbitrary as the separation between sunglasses and domes).

Figure 9.1: Example images from the ImageNet dataset. *The availability of datasets consisting of millions of labeled images provided a big boost to supervised learning algorithms for object categorization.*

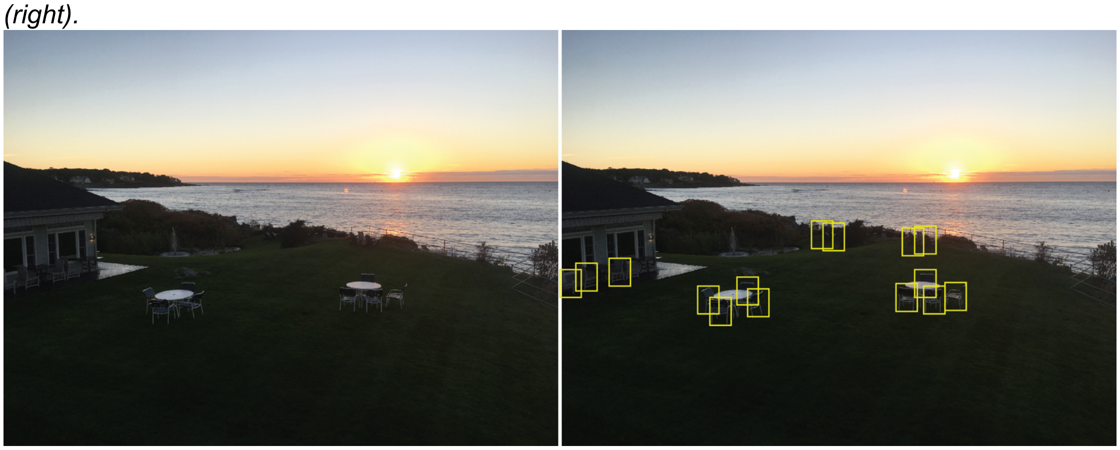


Armed with such a large dataset of images, the next step is to train a computational algorithm to label them. To ensure that we are not merely memorizing each image, it is critical to use cross-validation by separating the images within each category into a training set and a test set. All the model parameters can be modified *ad libitum* only while examining the training set. In deep convolutional network models, this step typically amounts to modifying the weights in a supervised fashion via back-propagation (see **Chapter 8**). However, it may also be possible to explore other aspects of the model including its architecture, number of layers, size of each layer, computational motifs, etc., as long as we limit ourselves to the training set. After training, the algorithm is tested with new images and the fraction of images that are correctly labeled is reported. A family of algorithms discussed in the previous Chapter have yielded increasingly higher performance in this type of task (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014b; He et al., 2015; Szegedy et al., 2015). For example, the "Inception-v2" architecture reported a top-1 performance of ~80%, which is quite impressive considering that chance levels are 0.1%.

Labeling an image to indicate whether it contains a particular object class is known as object identification (or object categorization). Beyond assigning a label, in many applications it may be of interest to localize where a particular object is in an image, a task known as object localization (or object detection). For example, the task may be to draw a bounding box around each chair in an image (**Figure 9.2**). Multiple algorithms have been developed for object localization tasks (Girshick, 2015; Redmon et al., 2016; He et al., 2018), including recent fast implementation that can work at frame rates compatible with video cameras running at 30 frames per second, therefore opening the doors to essentially being able to locate objects in real-time.

Most for historical and practical reasons, computer vision has disproportionately focused on single images, as opposed to videos. One particular domain that has been gaining traction in the computer vision community is action recognition, where the goal is to assign a label describing the action portrayed in a video. Several datasets have been developed and continue to emerge to evaluate action recognition capabilities (Soomro et al., 2012; Kay et al., 2017). Several of the comments above about image datasets are also pertinent in the case of videos. For example, contextual influences can also play an important role in action recognition: a lot of green pixels are more likely to be correlated with the action "playing soccer" than "swimming" whereas a lot of blue pixels are more likely to be correlated with "swimming". Additionally, in many cases, single frames can be sufficient for action recognition, without the need to invoke the temporal dimension. Several models have been proposed for action recognition, extending existing 2D image categorization architectures by incorporating trainable spatiotemporal filters (Simonyan and Zisserman, 2014a; Cheron et al., 2015; Feichtenhofer et al., 2016; Tran et al., 2017).

Figure 9.2: Object localization. *The task involves taking an image (left) and localization all instances of a given object class (e.g., chairs in this example) by drawing a box around them (right).*



9.3. Computer-vision applications in the real world

Deep convolutional network algorithms have had an enormous impact in a wide variety of vision applications. One of the earliest real-world applications was in algorithms for optical character recognition (OCR), which rapidly became mainstream in sorting letters based on their zip codes. There are even neat applications that can translate handwritten traces into mathematical formulae. On the one hand, some mathematical symbols are relatively simple; on the other hand, they are probably less stereotyped and there is less training data than in other OCR applications.

There are several situations where there is a very large number of images (or video) that needs to be classified. For example, one of the challenges in Astrophysics is to classify vast amounts of imagery to understand the shape of galaxies. One of the ways in which this was achieved was via crowd-sourcing by engaging the public in looking at images and learning to categorize galaxies. This is an ideal setting to apply pattern recognition techniques from computer vision (Kim and Brunner, 2017). A conceptually similar example is the categorization of plants and animals (Van Horn et al., 2018).

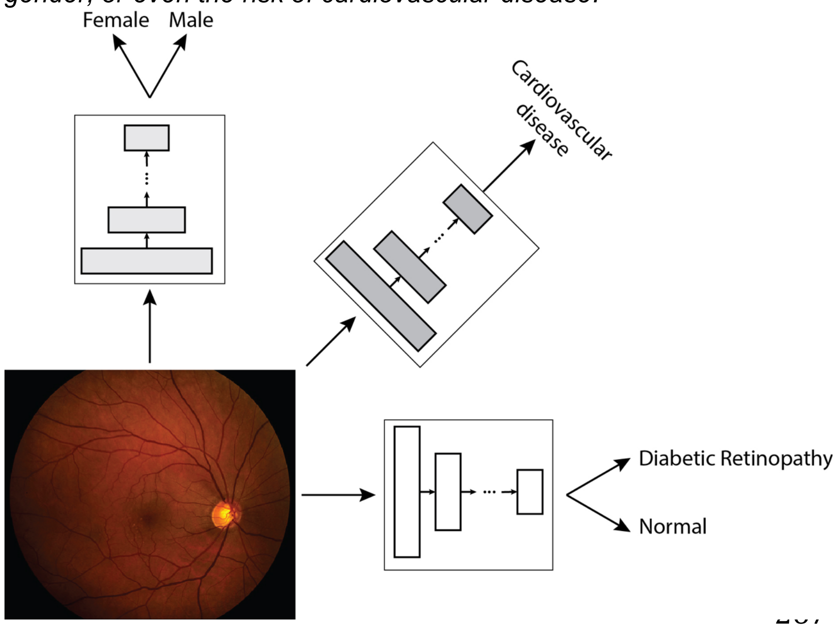
The exciting progress in self-driving cars has also been fueled by progress in computer vision, building sensors to localize pedestrians, other cars, brake lights, traffic lights, other signs, lanes, the sidewalk, even animals, bicycles, or anomalous objects on the road. While the majority of computer vision applications rely on video or camera feeds from regular cameras, images do not have to be restricted to such sensors. For example, self-driving cars can simultaneously use information from multiple cameras as well as multiple other sensors. There has been so much progress in terms of vision that most engineers trying to build self-driving car think that the main challenges ahead depend on how to intelligently use such information to rapidly make informed decisions rather than localizing specific objects and object types.

Computer vision is making enormous strides in the domain of clinical image analyses, so much so that there are many examples of problems where machines are on par or better than humans. Humans are capricious creatures, doctors do not always agree with each other in diagnosis. Sometimes doctors do not even agree with themselves when tested on the same image recognition problem on different days! One example problem is breast cancer detection from mammograms (Lotter, 2017). The American Cancer Society recommends obtaining a mammogram, generally consisting of two x-ray images of each breast, to all women, once or twice a year depending on age. This is a lot of images, early diagnosis can have a critical impact on deciding the course of action, and there is clear documentation of the variability among radiologists (Elmore et al., 2009). Current algorithms can achieve performance comparable to expert radiologists. In other words, computer vision algorithms can pass the Turing test in terms of discriminating whether a mammogram is likely to represent a tumor or not.

While this is the main question of interest in the vast majority of breast exams, occasionally, there may be other relevant questions clinicians may want to ask about an image. For example, sometimes there are incidental findings where a person is scanned to diagnose a given condition X. The scan does not reveal any finding regarding X but the radiologist detects other anomalies that lead to a different diagnosis Y. Such incidental findings may be complex for current computer vision algorithms because they may be extremely rare and the algorithms are ultra-specialized in detecting condition X. One possible initial solution would be for computer vision systems to flag such images as anomalous and route them back to a human for further inspection. In the future, it is quite possible that future generations will regard humans trying to diagnose images in the same way that we would now imagine a human trying to interpret a bar code in the supermarket.

Incidental findings may represent one arena where humans may still surpass machines in clinical image diagnosis. The reverse is also true. Machines may be able to discover aspects of how to reason about images that were never conceived by humans before. An intriguing example of this phenomenon arose when investigators were examining retinal fundus photographs (Poplin et al., 2018). They were interested in using a computer vision approach to diagnose diabetic retinopathy, a condition that may arise in diabetic patients when high blood sugar levels cause blood vessels in the retina to swell and leak. These blood vessels can be examined in fundus photographs, images of the back of the eye, used by ophthalmologists to diagnose the disease. After collecting hundreds of thousands of labeled images, a computer vision algorithm quickly learned to match clinicians in diagnosis, a feat that comes as no surprise at this stage (**Figure 9.3**). The diagnosis label is only one of the questions that one can ask about those images. The investigators decided to turn their machine learning algorithms to ask other questions on the same images. First, they asked whether they could guess the subject's age and, voila, they could do so quite precisely, with an absolute error of less than 3.5 years. Next, they asked whether they could predict the subject's gender. To everyone's surprise, they were able to do so extremely well, with an area under the receiver operating curve of 0.97. The curve refers to the plot of the probability of correct detection versus the probability of false alarm: it is trivial to achieve high detection rates at the expense of very high false alarm rates (by claiming that every image shows disease), or very low false alarm rates without any correct detection (by claiming that no image shows disease). A good algorithm will have low false alarm rate and high probability of detection; the best that an algorithm could achieve is an area of 1.0. Trained ophthalmologists had never been able to guess somebody's gender from fundus photographs. Perhaps they never cared to ask that question, after all, they have the subject right in front of them. However, even after telling them that the information was there and asking doctors to guess the gender, they were unable to do it. It is not entirely clear what exact image features the algorithm uses to discriminate gender. Some people have hypothesized that perhaps doctors, both male and female, position themselves slightly closer to female patients than to male patients on average and this slight bias is captured by the algorithms. Or perhaps there are real subtle differences between female and male blood vessels in the retina. Regardless of whether this explanation holds true or not, this example shows that computer vison can discover image features that are not apparent even to experts in the field. In this case, those features (age and gender) are perhaps not that interesting (doctors always have access to both without a fundus photograph). The most enigmatic finding appeared when the investigators decided to ask an even more daring question. Would it be possible to predict the risk of cardiovascular disease from fundus photographs? Intriguingly, the answer was yes, with an area the curve of 0.7, which is comparable to the best predictors such as the Framingham score based on decades of clinical work. Computer vision algorithms can not only learn to diagnose images like doctors, they can teach us novel things from those images.

Figure 9.3: Clinical applications of computer vision. *Example clinical application of computer vision, taking a photograph of the back of the eye (fundus photograph) and using a deep convolutional network to diagnose diabetic retinopathy (Poplin et al., 2018). In addition, computer vision algorithms can be trained to ask other questions from the same image, including predicting the subject's gender, or even the risk of cardiovascular disease.*



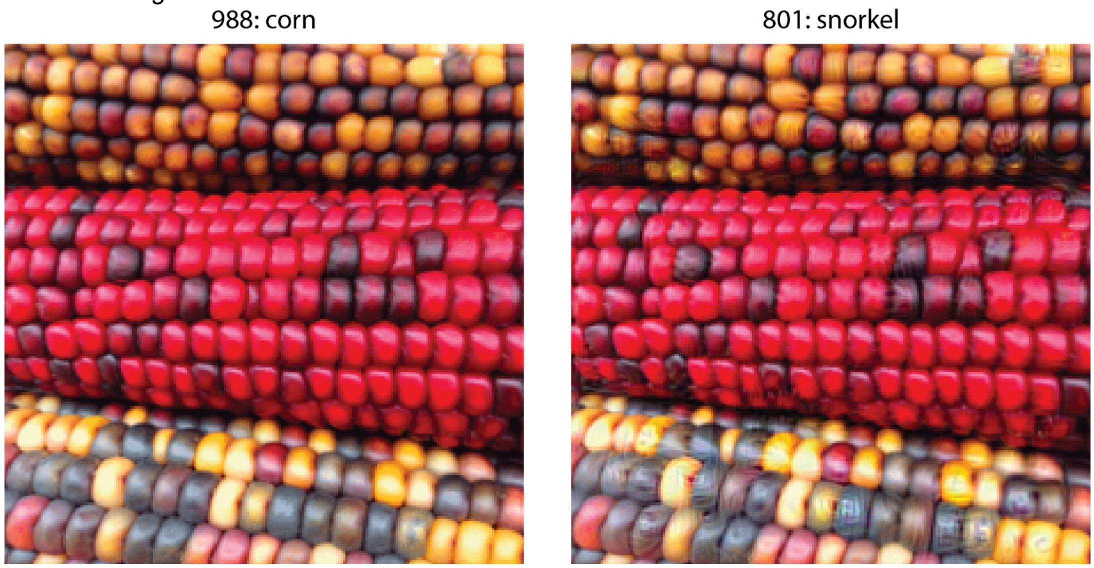
There is a wide variety of applications for automatic face recognition algorithms. The current version of the iPhone can use an image of the user's face to log in. Facebook can now search for photos that include a particular person when that person is not tagged. State-of-the-art algorithms for face recognition surpass expert human performance including forensic facial examiners, facial reviewers, and so-called superrecognizers (Phillips et al., 2018). There is also a growing industry of security applications based on facial recognition capabilities. Security applications in the near future may also rely on action recognition classification algorithms. Concomitant with advances in face recognition, there are vigorous and interesting discussions about concepts of privacy. It is quite likely that very soon, it will be rather challenging to walk down the street without being recognized.

Similar ideas have also expanded well beyond vision and are making rapid strides in fields as diverse as speech recognition, predicting voting patterns or predicting consumer choices. Interestingly, in a vast majority of cases, the basic architecture of the algorithm is the same, a deep convolutional network that mimics the cascade of processing along the ventral visual stream. What changes is the basic input: instead of using pixels in RGB space, one can use a spectrogram of the frequencies of sound as a function of time to process sounds. However, subsequent processing steps and the procedure to train those algorithms is remarkably similar if not exactly the same in many applications. In Neuroscience, this idea is sometimes phrased as "Cortex is cortex", alluding to the conjecture that the same basic architectural principles are followed in the visual system, the auditory system, etc. Without doubt, there are important differences across modalities, and engineers will also fine tune their algorithms for each application, but as a first approximation, some of the basic ingredients seem to hold across multiple tasks.

9.4. Challenges ahead

Exciting and rapid progress in computer vision may lead us to think that we have almost solved the problem of vision. Yet, I would argue that we are still extremely far. And the best is yet to come.

Figure 9.4: Adversarial examples. *The two images below appear to be indistinguishable to humans. Yet, state-of-the-art computer algorithms classify the one on the left as corn and the one on the right as snorkel.*



One interesting current challenge that has led to a lot of discussion is the notion of adversarial examples (**Figure 9.4**). These are images that appear similar to humans but that receive different labels by a computer vision system (Szegedy et al., 2014). Given any algorithm that is forced to assign a binary label to an image, A versus B, it is inevitable that there will be a boundary where you can move from A to B with small image changes. It's like standing in the arbitrary border between two states or two countries. These adversarial images were created by using knowledge about the categorical boundaries and astutely changing a few pixels to push the image into the opposite side. Humans also suffer from such adversarial examples and many other visual illusions that deceive us into seeing things that don't exist (Chapter 4). And there are many cases of images that deceive humans but do not confuse computer vision systems. What is intriguing about these adversarial examples is the profound difference between machines and human perception. In many real-world applications, seeing the world the way humans do may be quite relevant. In fact, there has been a whole industry of investigators designing "adversarial attacks" to confuse computer vision systems, together with a similarly vigorous community of defenses against such adversarial attacks. For example, one may ask whether the image on the right in **Figure 9.4** would revert back to a corn if it is scaled, or its color changed, or using different versions of the same network (e.g. starting from different random initial conditions), or using different architectures. These examples clearly illustrate that even if current algorithms can label lots of images correctly, they do not necessarily see the world the way humans do.

An interesting application of computer vision would be to restore visual functionality to people with severe visual impairment. By restoring "visual functionality", we do not necessarily mean getting a blind person to *see* in the same way that a sighted person does, but rather, the ability to convey information that they can use. Digital cameras are extremely good and relatively cheap. A blind person could easily where a camera on their forehead, or in a pendant. Imagine an algorithm that can label every object in an image. How can we convey such rich information to a blind person? An image is worth a thousand words, they say. In a glimpse, we get a rich representation of our surroundings, which is quite different from labeling every object. This representation highlights certain aspects of the image while ignoring others, it allows us to discern distances, relationships between objects, even actions and intentions. Even if we could accurately label all the objects in an image, there is a much more to visual understanding. While we are discussing blind people, we could easily extend these ideas to enhancing the visual capabilities of sighted people as well. It would be easy to wear a camera that would give us immediate access to a 360-degree view of the world, or cameras that grant us real-time access to other parts of the spectrum that our eyes are not sensitive to such as infrared. Computer vision systems could help us parse those images.

Another area that is advancing rapidly, and yet there is also plenty to improve, is image captioning (also related to question-answering systems on images). Given an image, the goal is to provide a brief and "relevant" description. In contrast to categorization tasks, it is more challenging to quantitatively evaluate the results. An example of state-of-the-art in image captioning is shown in **Figure 9.5**, which is based on results obtained on https://www.captionbot.ai/ (circa November 2018). It is important to emphasize the date because I suspect that we will see major improvement in the years to come. The captions provided by this algorithm are quite impressive. The system is pretty good at detecting people, even whether it is one person (9.5A) or multiple people (9.5B, D). The system can also detect the gender in 9.5A and it makes a reasonable guess that people are happy in 9.5D. The system also correctly infers that the person is sitting in 9.5A, and standing in 9.5B, D. The system also detects other important aspects of the scene including the presence of a table in 9.5A, water in 9.5B and a building in 9.5D. There are many other objects that are not described, which is perhaps not too bad, given that the goal is to caption and not mark every single object. It is a bit surprising that the system does not describe the Tower of Pisa in 9.5D, given that such monuments have an exorbitant amount of training data.

There is a rather salient spoon in 9.5A that was not described. And it seems likely that a lot of humans would describe the bride in 9.5B. The system is not able to describe line drawings (9.5C), but it is quite nice that it was able to realize that this is a light drawing. Differentiating line drawings from photographs is perhaps not too difficult, particularly if the image has a huge number of white pixels, a few black pixels and essentially no textures. It is relatively easy for humans to recognize that there are 3 people in the drawing in 9.5C, though it is not clear exactly how this deduction happens. Current algorithms such as this one probably have minimal, if any, training with drawings. In contrast, most humans have had exposure to the basic symbolism behind line drawings.

Figure 9.5: Image captioning. *Four example results from the www.captionbot.ai image captioning system.*

One easy way to break these captioning systems is to scramble the image. For example, we can divide the image into four quadrants, and rearrange the quadrants randomly. The image largely loses its meaning, yet the caption remains largely unchanged. If we present the fundus photograph from Figure 9.3 (only the fundus photograph, without the rest of the Figure), the system responds with "I can't really describe the picture but I do see light, sitting, lamp". It's commendable that the system realizes that it cannot quite describe the image, that it realizes that it is different from its training set. And there is indeed a light in there. The system probably saw many examples where the word "light" is correlated with the word "lamp", throwing it into the description. It is a bit harder to deduce where the word "sitting" comes from, a characteristic that many people have criticized: given the large number of parameters in the system, it is not always easy to put in words why the system produces a given output. Of note, the same type of architectures can be trained to outperform doctors in interpreting the same fundus photographs. Doctors can evaluate fundus photographs and also understand what is happening in Figure 9.5 where as many current systems are ultra-specialized for specific tasks.

To end on a light tone, I would like to highlight an example of a problem that I consider to be extremely challenging: understanding the human sense of humor. It is clear that one can ask a large number of questions about the images in **Figure 9.5**. As impressive as those captions are, they do not come even close to solving the Turing test for vision. The captions completely miss to grasp fundamental aspects of scene, what is happening, who is doing what, to whom, and why. Humans can look at these images and understand the relationships between the different objects, what is their relative positions, why they are where they are, and even make inferences about happened before or what may happen next.

Even more mysteriously, all these images are meant to be somewhat curious or funny. Let us consider 9.5C as an example. Why is it funny? To grasp what's happening, we may need to incorporate not just the pixels, not just the specific objects, but also their symbolism and relative interactions. The scale at the center, together with the few traces that represent the attire of the person in the center, plus his relative position with respect to the other person leads us to think that he is a judge. Note that it is the combination of many of these labels and their interactions that lead us to this understanding. Each one piece of information on its own would not necessarily be sufficient. The person sitting below the judge is probably the accused (or less likely a witness). This inference is perhaps partly based on the person's shirt with horizontal stripes, but mostly based on his relative position and an understanding of the arrangement of the judge and the accused in a court of law. We can infer that the third person is a policeman, which is consistent with his outfit but also with the fact that he is standing, and that he is behind the accused. After deciphering that the person in the center is a judge, we guess that is his holding a gavel, that he is shouting, and that he is hitting the table with his gavel. The accused is also angry, making eye contact with the judge. Curiously, the accused also seems to be holding a gavel. This is unusual: the accused is not supposed to hold a gavel, let alone use it, as he is doing. This is the essence of why the image is funny: it portrays an unexpected scenario. If we take out the few pixels that represent the accused's gavel, the image immediately becomes less interesting. There is a very large amount of world knowledge that we need to have to be able to understand to interpret Figure 9.5C. What's more, predicting humor is further complicated by the fact that, even if you trained an algorithm to understand all the symbolism in Figure 9.5C, that would be of no help whatsoever to understand why Figure 9.5A is intriguing, nor to deduce what probably happened in Figure 9.5B.

References

Cheron G, Laptev I, Schmid C (2015) P-cnn: Pose based cnn features for action recognition. In: IEEE international conference on computer vision

.

Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, Yankaskas BC, Kerlikowske K, Onega T, Rosenberg RD, Sickles EA, Buist DS (2009) Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 253:641-651.

Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Girshick R (2015) Fast R-CNN. In: (ICCV, ed).

He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. arXiv 1512.03385.

He K, Gkioxari G, Dollar P, Girshick R (2018) Mask R-CNN. In: IEEE Trans. Pattern Anal. Mach Intell.

Kay W, Carreira J, Simonyan K, Zhang B, Hiller C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A (2017) The kinetics human action video dataset. arXiv:arXiv:1705.06950v06951

Kim E, Brunner R (2017) Star–galaxy classification using deep convolutional neural networks Monthly Notices of the Royal Astronomical Society 464:4463-4475.

Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: NIPS. Montreal.

Lotter WE (2017) Prediction as a Rule for Unsupervised Learning in Deep Neural Networks. In: Program in Biophysics. Cambridge: Harvard University.

Phillips PJ, Yates AN, Hu Y, Hahn CA, Noyes E, Jackson K, Cavazos JG, Jeckeln G, Ranjan R, Sankaranarayanan S, Chen JC, Castillo CD, Chellappa R, White D, O'Toole AJ (2018) Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences of the United States of America 115:6171-6176.

Poplin R, Varadarajan A, Blumer K, Liu Y, McConnell M, Corrado G, Peng L, Webster D (2018) Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering 2:158-164.

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: CVPR.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang S, Karpathy A, Khosla A, Bernstein M, Berg A, Fei-Fei L (2014) ImageNet Large Scale Visual Recognition Challenge. In: CVPR: arXiv:1409.0575, 2014.

Simonyan K, Zisserman A (2014a) Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems.

Simonyan K, Zisserman A (2014b) Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556.

Soomro K, Zamir A, Shah M (2012) UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. arXiv 1512.005673v3.

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. In: International Conference on Learning Representations.

Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2017) A Closer Look at Spatiotemporal Convolutions for Action Recognition. arXiv preprint.

Turing A (1950) Computing Machinery and Intelligence. Mind LIX:433-460.

Van Horn G, Oisin M, Song Y, Cui Y, Chen S, Alex S, Hartwig A, Perona P, Belongie S (2018) The inaturalist species classification and detection dataset. In: CVPR.