AI/Bio Computer vision

Gonzalo Reynaga García

April 14, 2021

Currently the problem of object recognition depends on how large and diverse the data set is used to train a neural network. In this article we analyse this problem from AI perspective and how to create a more efficient neural network and compare it algorithmically with biological vision.

How Artificial Intelligence (AI) works?

Currently the AI is based in a union of large number of neurons, that is a sum of weight inputs within a function. This group of neurons just create a function approximation for the output we want for an input, and for each neuron we add we need to know what weight value should get.

To determinate the weights, it used a gradient method that is an iterative way to minimize the error of a certain value (the cost function). This is neural network is converted into differential equations and like any differential equations has initial values, that in this case is the initial values of neuron's weights.

Depth perception

There several ways the animals and humans can perceive depth and the main ones are:

- **Binocular vision**. This can be achieved by getting two images of the same scene and get the difference between a point in one image and the location the same point in the other image, this can be used to determinate the depth of that point.

- **Focus**. A person can perceive depth, even with one eye, due the ability to focus the eye into an object, if the eye is focusing a far object, the near objects would look blur; this is used as a hint to perceive depth.

- **Motion parallax**[1]. If the observer moves, the image section that is far away, it will move less that the image section that is closer, giving the perception of depth.

- **Depth from motion.** This occurs when the object moves, it is possible to be a perception of velocity rather than depth[1].

The story of Waty (one neuron bug)

Let´s suppose a bug with one neuron exists, sized 5 mm, named Waty, that is able to only walk and see the world in grayscale with his eyes.

His brain has extremely low capacity and the minimum to survive is search for food.

Waty is in front of a tree and wants to eat an apple. What should Waty do to eat the apple from the point of view of vision processing?

Figure 1. Waty and tree

Process image pixels

Waty is training to recognize an apple. He takes an image of an apple (already it acquired before) so he uses his brain to store this image and process a 2D convolution of the image into the scene to locate the position of the apple.

Then Waty walks towards the apple but his eyes see a bigger apple image because is closer, and doesn't match with the convolution he did before, and he cannot recognize it.

He has to store this bigger image into memory to do a 2D convolution to locate the apple. And the process repeats again, store a bigger image into memory due the same reason.

Now a shadow is casting over the over apple, causing Waty to see a different image; and he has to store the new image into his brain to locate the apple.

After storing all the variations of lighting and size of the apple image into his brain (that is an extremely large number of combinations), and in this case we are ignoring the object's orientation variation just to simplify, Waty walks towards the apple to finally stand below of it on the ground.

What did he miss? Does he have to recognize the apple is on the branch? Does he have to know the branch is attached to the trunk of the tree? And the trunk is attached to the ground?

Then, suppose that he has to recognize the branch, trunk and all the variation of light and sizes. What if the tree is in a pot? Does he have to recognize the pot as well? And if the apple is on a table?

Obviously, Waty has only one neuron and he doesn't have the amount of memory required to store all the apple images and he will stay below the tree staring the apple, hungry.

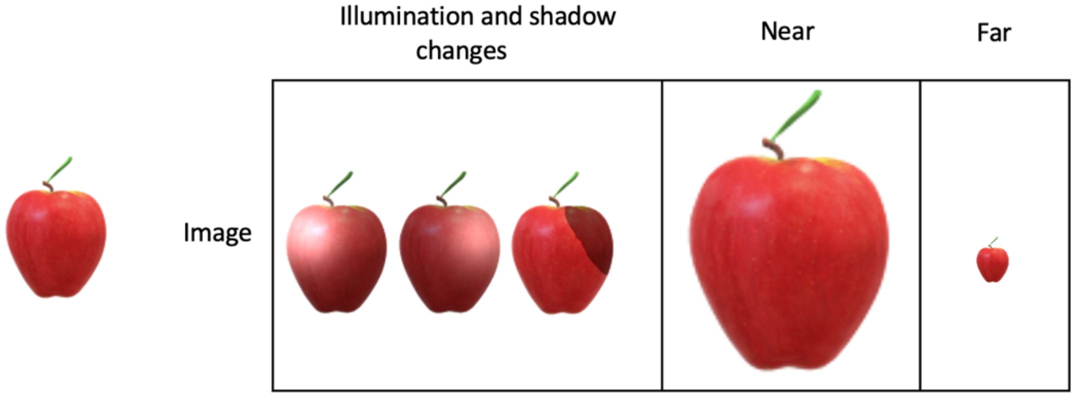

Figure 2. Apple image changes

Processing depth

Now Waty has the depth information to recognize an apple (already it acquired before) and stored it into his brain (imagine a sphere to simplify).

He sees the environment with his two eyes creating a stereo image of his surroundings. This time is using the 2 images acquired by his eyes to calculate the depth of the environment.

Once he has the depth of the environment, he calculates the 3D convolution of the environment depth with the apple depth; this 3D convolution gives us an indication on how correlates the boundary of the environment depth with the boundary of the apple depth stored in memory.

He has a position of the apple and the depth information, he notices with this information, the apple is attached to something, and that something is attached to something else, and he can walk into that something without having to recognize that something; in this case is the trunk of the tree, but can be anything, like a pot, table, etc.

Let's imagine there is a ramp that goes directly to the apple and Waty cannot climb vertical surfaces, even so, just using depth information can determinate the slope of the surface and take the ramp without recognizing any object except for the apple.

When Waty walks to the apple, the image of the apple is bigger, but he doesn't process the image, he processes the depth and the size remains constant, so he doesn't need to store more images in his brain for the apple.

The illuminations changes don't affect the detection of the apple due the images acquired is just used to process depth and it doesn't matter if the image change due illumination.

Waty only requires processing the environment depth and store one single depth of an apple (we are ignoring orientation variations to simplify). This will only detect the apple of that specific size and it will be very difficult to confuse that volume with other similar, unless there is a rock with that shape; in that case, Waty will bite it, and then after losing a tooth, he will start to look for other apples, repeating the process again.

After Waty found his way to survive, he has gained more neurons to store more objects to detect, starting to detect his own specie to reproduce, then identify depredators, etc. following the same depth system, is there a reason to change it?

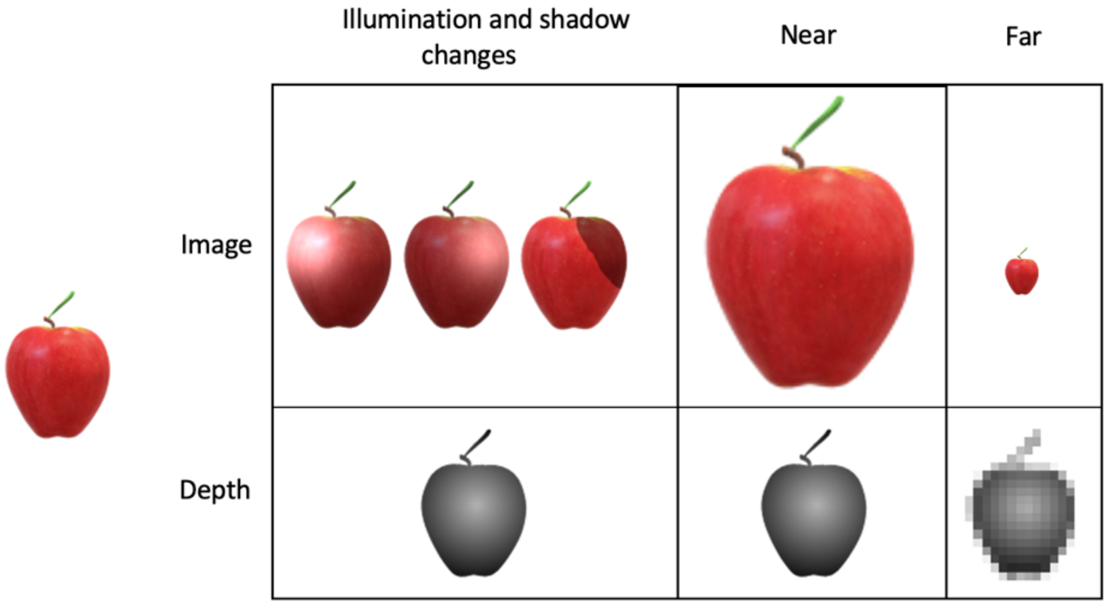

Figure 3. Apple image and depth changes (in depth, the volume will remain the same regardless the object distance)

After millions of years, Waty has billions of neurons and can detect hundreds of thousands of objects and now he can see color. Now he has a computer science degree, and he thinks is using image data to recognize objects rather than depth.

Waty takes a camera and takes a picture of an apple, he printed the picture and recognize the apple in the picture.

How Waty can recognize the apple in a picture if he has been recognizing objects using depth for millions of years?

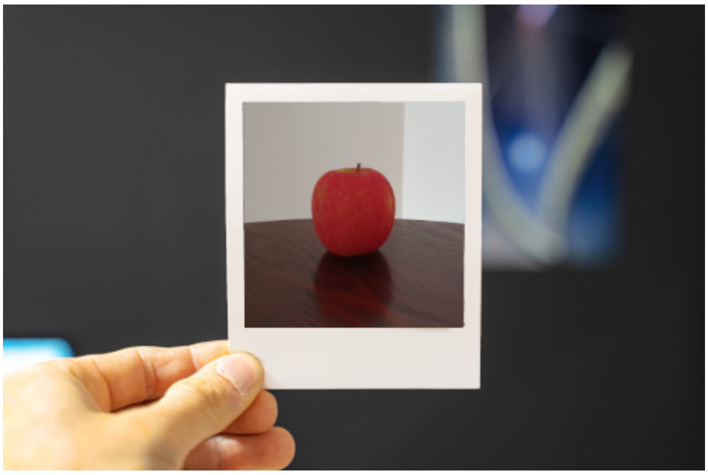

Figure 4. Apple picture

Recognizing objects from pictures

In this section we try to answer the question, how Waty can recognize the objects from pictures? And the answer is he cannot, at least not initially.

There is a deep study and evidence that exist difficulty to recognize objects from pictures in animals and humans[2].

To quote a line in the conclusion in this article:

> *"picture recognition seems to present greater difficulties for adults who are unaccustomed to seeing photographs and drawings."*
>
> *-(Bonet, Dalila, et al.)[2]*

Waty cannot recognize the object in the picture but he can learn to recognize it, but how?

My hypothesis about it is if he spent millions of years recognizing objects using depth, then the simplest way to do it is creating a depth estimation in the picture and then run the same object recognition algorithm he has been using from millions of years.

Creating a depth estimation from a picture is a skill gained and we all do it without notice it.

Using this hypothesis let's try to explain the following optical illusions:

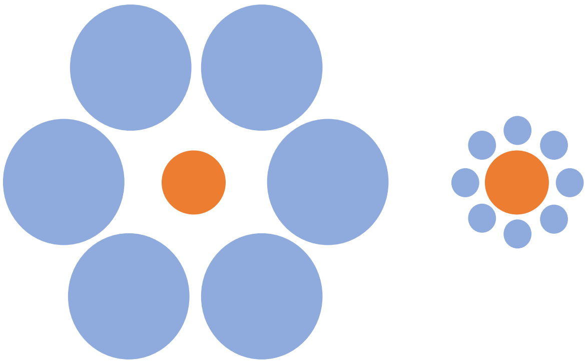

Figure 5 Ebbinhaus illusion

Which of the central circles is bigger? *there are the same size.*

Now, the same illusion but with motion:

https://www.youtube.com/watch?v=hRIWqfd5pn8

The effect is even stronger.

My explanation is that the brain creates a depth perception of the image, and according with the brain the central circle in the left is far, but as is a flat image and it is not far, then we perceive it smaller.

And in the case of the dynamic *Ebbinhaus illusion* from *YouTube* video, I think the brain could use the motion to *"correct"* the depth perceived as the surrounding circles are bigger perceiving the central circle is *"further away",* it considers the size should be *"far"* comparing with the initial considered depth creating an effect that is constantly becoming smaller.

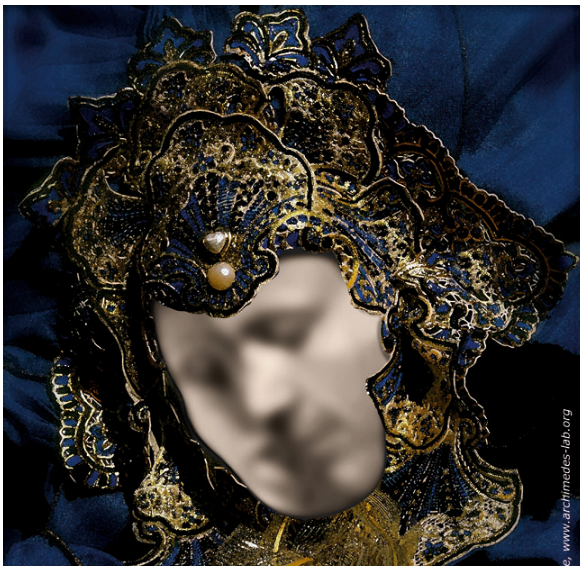

Figure 6. Mask of Love. (© 2011 Gianni Sarcone, Courtney Smith & Marie-Jo Waeber)

http://illusionotheyear.com/2011/05/mask-of-love

In *Figure 6*, there are two ways to see the image, one way shows one face and the other way show two faces.

My explanation for this illusion is that the brain set an initial depth estimation for the image to see and recognize one face (if this is the initial perception). But the brain has the ability to change the depth, a sensation of change *"focus"*, and this new depth match with the recognition of two faces.

Complete algorithm to recognize objects

Waty now is driving a car and wants to recognize a traffic signal from the road, a speed limit. It has the same vision capability as a human, he can see color, he has the same object recognition based in depth, but now, he can process the image pixels, but just only a very tiny part of the image.

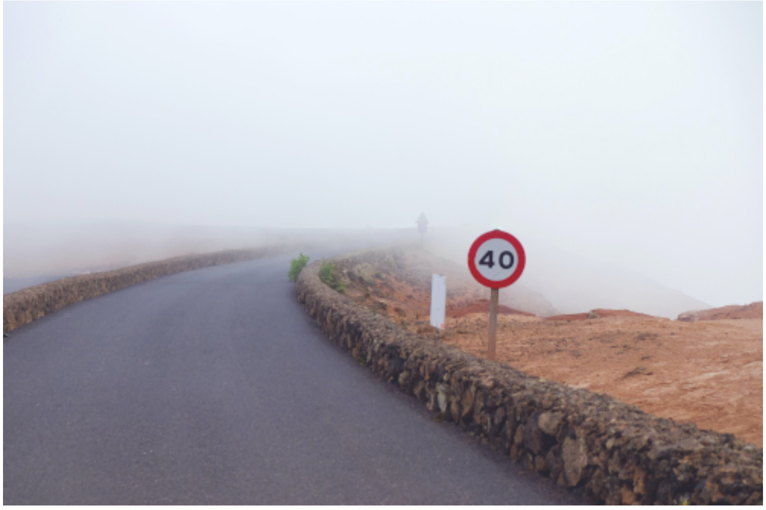

Figure 7. Traffic sign

His brain has adapted to select a portion of the image to be processed, this selection is based in objects that keep their attention in movement and high contrast colors.

He has his eyes looking to the road; without looking, he detects and recognize a traffic sign at the side of the road based in depth but be doesn't know which sign it is, due the object depth perception doesn't give him that information. As he knows for certain is a traffic sign, he then turns his eyes to this sign to process image on it.

Based in the image processing and comparing with all the trained traffic signs the has in memory, he recognizes is a speed limit sign. Note that he uses only a very specialised neural network image processing just to process traffic signs, discarding processing other objects like elephants, pandas, beds, cars, etc. therefore is a smaller neural network.

But what if the sign is on the wall? If the sign is on the wall the depth perception won't let him recognize a traffic sign, but he can recognize the wall based on depth, and on the wall can be match with any kind of flat objects, like traffic signs, posters, advertisement, billboards, etc.

He can only find the traffic sign on the wall based in high contrast color or if the deliberately looks for something in the wall, and the same as before he cannot be confused in process other objects like car, persons, etc. due is a flat surface, even if it has an image of a car or a person in a poster or billboard.

As a summary, the depth perception recognition throws a list of possible objects, and the image processing specify the object, in that way the depth discard a lot of objects with a very simple processing and the heavy specialised image process give the final object recognition. In that way, Waty from millions of years ago wouldn't confuse the rock with an apple.

Implementation of object recognition based in depth.

You would think, well doesn't matter how is done, the neural function should learn how to do it, yes, I agree with that, but the problem is how we train it.

The current neural networks to recognize objects works different that the human perception of the objects[3]. My interpretation of the motive of this, it's the neural function is large enough and it is just memorizing the input to give the correct result for the specific input, that means the neural function converge into a local minimum.

If the neural network is big enough, it is easier to just transform the input into the output desired, if the network is keeping memorizing the inputs, we have better result in output. But in this way, it doesn't really learn anything.

And if the neural function reaches the maximum capacity of memorizing the objects, the additional object we force to learn, the function won't try to change its recognizing method, it just spread the error a across the other objects, because it's difficult to get out of the local minimum.

The solution to actually force a neural network to learn is use the same method as Waty, he started with one neuron and trained with thousands of data samples. Obviously, this will be impossible to get good results, after that intense train, add other neuron and train thousands of data samples again, add other neuron and repeat until reach a considerable large number of neurons.

This method would work, just because with a scarce number of neurons, the neural function won't have other choice that converge into the best method possible to get the best result with the current number of neurons, that in this case I think it have to process depth internally rather than the image itself. And once it has the best method is likely to continue with that method in the next generations (when a neuron is added).

Gradually adding neurons to the approximation function (neural network), it's likely we get a global minimum and what is really it do the job are the initial weights values passed into generations, those initial values very important because it could decide if converges into local minimum, therefore to overfitting.

Conclusion

It's very evident that the recognition of object based on depth it is more efficient than processing the image itself, the challenge will be creating a precise depth of scene enough to run the algorithm.

It is so efficient than the human and animals it is more likely to use depth for recognize objects, that will explain the problem of mosquitos have when landing in zebras due the stripes patterns[5] and the fact that the mosquito's eyes cannot focus an image to determinate depth.

Even the process of image is not totally unused, the depth has more priority (weight) when decide an object classification, I suppose this will be more evident with smaller brains with the correct experiment, like mosquito for instance.

I think Waty's story is simple enough that can be modelled and simulated in a computer and analysed what he can see and how.

References

[1] Depth perception https://en.wikipedia.org/wiki/Depth_perception

[2] Bovet, Dalila & Vauclair, Jacques. (2000). Picture recognition in animals and humans. Behavioural brain research. 109. 143-65. 10.1016/S0166-4328(00)00146-7.

[3] Out of shape? Why deep learning works differently than we thought https://blog.usejournal.com/why-deep-learning-works-differently-than-we-thought-ec28823bdbc

[4] Stanford I Convolutional Neural Network for Visual Recognition I Image Classification I Lecture 2 https://www.youtube.com/watch?v=W4td7raCidQ

[5] How Do Zebra Stripes Stop Biting Flies? https://www.youtube.com/watch?v=JyDa8SQ0l3I

[6] 9 Weird Ways Animals See the World https://www.youtube.com/watch?v=z4rxineIFFE

[7] How Computer Vision Works https://www.youtube.com/watch?v=OcycT1Jwsns

[8] MIT 6.S094: Computer Vision https://www.youtube.com/watch?v=CLOAswsxudo

[9] Fish Vision https://www.offthescaleangling.ie/the-science-bit/fish-vision/

[10] Scientists Velcroed 3-D Glasses to Cuttlefish to Study Their Depth Perception https://www.smithsonianmag.com/science-nature/scientists-velcroed-3d-glasses-cuttlefish-study-depth-perception-180973918/

[11] Praying Mantises Don Tiny Goggles to Help Us Understand 3-D Vision https://www.smithsonianmag.com/smart-news/praying-mantises-have-their-own-version-3d-vision-180968123/

[12] HOW ANIMALS SEE THE WORLD? https://allyouneedisbiology.wordpress.com/2015/11/29/animals-vision/

[13] YOLO Algorithm and YOLO Object Detection: An Introduction https://appsilon.com/object-detection-yolo-algorithm/

[14] Insect Vision: Ommatidium Structure and Function https://www.youtube.com/watch?v=TU6bgQnTi18

[15] Insect Vision Part 1: Apposition Eye https://www.youtube.com/watch?v=Lpt0XN_G8Tc

[16] Where We See Shapes, AI Sees Textures https://www.quantamagazine.org/where-we-see-shapes-ai-sees-textures-20190701/

[17] Neuroscientists find a way to make object-recognition models perform better https://news.mit.edu/2020/object-recognition-v1-1203

[18] Su, Jiawei et al. “One Pixel Attack for Fooling Deep Neural Networks.” IEEE Transactions on Evolutionary Computation 23 (2019): 828-841.

[19] Sheep Can Recognize Human Faces | National Geographic https://www.youtube.com/watch?v=Rcl_25XZleY

[20] Godard, C. et al. “Unsupervised Monocular Depth Estimation with Left-Right Consistency.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 6602-6611.

[21] Predicting 3D Volume and Depth from a Single View https://www.youtube.com/watch?v=5uMGMZ05LC4