AUDE OLIVA: Thank you very much for the introduction. Good morning, everyone. So I'm very pleased to be here. It's the first time I visit.

So I would like to give you a tour during this lecture about how you can predict human visual memory, and actually an interdisciplinary account of the methods you can use together in order to have basically a view of what people can remember or forget.

So the specific question we are going to ask today is, well, we are experiencing and seeing all the time a lot of digital information. First you see in the real world, but also you are exposed to many, many videos and images. So vision and visual memory is one of the core concept of cognition.

And the question we ask is, can we predict which image or graph or face or words or videos or piece of information or event is going to be memorable or forgettable for a group of people, and eventually for a given individual?

So let's take a moment to imagine that we will be able to predict accurately the memory of people. Well, this will be very useful to understand the mechanism of human memory, both at the cognitive and neuroscience level, system level, as well as possibly diagnose memory problem, short-term visual memory, long-term visual memory, that may arise possibly in an acute way or developing over time, as well as design mnemonic aids in order to recall better informations.

But beside basic science, if you could predict information that are memorable or forgettable, there is a realm of application that you could work on or basically propose that lie really everywhere between the data visualization or the slogan, make basically slogan better, as well as all the realm of education, the individual differences that may arise between people.

Someone will not learn very well. Well, what can we do in order to increase the memory of visual information that this person can grasp? As well as various applications-- social

networking, faces, retrieve better images, and so on.

So understanding what make an information memorable or forgettable basically is really a very inter-disciplinary question, something very exciting for us to work on, because there's a lot, a lot of future in working in that topic.

So this is a topic we started in my lab a few years ago. And the best for you to get a sense of how you start working, for instance, on memory is to do the kind of game and the experiment we had people doing.

So welcome to the Visual Memory Game. A stream of images will be presented on the screen for one second each. And your task is to-- If you are in front of your computer you will press a key. But what you're going to do is play the game with me and clap your hand anytime you see any image that you saw before.

So you're going to have to be attentive, because images, they go by fast in this rapid stream of images. And so you will be getting feedback. So it's a very straightforward memory experiment. And this is the first step in order to get some score on a lot of images regarding the type of information that people will naturally forget or naturally remember.

So let's do the game. So are you ready? All right. So clap your hand whenever you see a repeat. So this is what will rerun. Very simple images.

[CLAPPING]

Fine. Excellent. So images are shown one second. There's a one-second interval.

[CLAPPING]

You're good. No false alarm. Excellent. All right. So that was one of the level, level 9, out of 30 complete. And so here's the game that people had. They could play this game for five minutes, had a break, and come back. And this game had a lot of success, and we run it.

I'm going to show many results about this. So you could see your score, the amount of money you have done. This was run on Amazon Mechanical Turk. And this allow us to collect all around the world a lot of data regarding many, many images.

So those visual memory experiment were set up by Phillip Isola from MIT. And you can

basically play that game for any kind of information displayed visually. So we did it for pictures, faces, and words. And you're going to see the result.

So let's start with the pictures. In the first experiment, we presented 10,000 images, and about 2,200 we repeated many, many time. So those were the one where we collected the score. So for a given subject, those images were actually seen only once. Twice, sorry.

So you have the stream of images at the top. So, for instance, an image will be shown after 90, 100, 110 images or so. And if this image was one that the subject recognized, then he will press a key. So it's exactly the design you just did.

And when first you look at the type of images that are highly memorable or forgettable, well, there's a trend that we all expect, is images who are kind of either funny or have something distinctive or something different, or if people are doing various action, or of some object that are kind of out of context, those tend to be memorable. And, in general, landscape or images that don't have any activity tend to be forgettable. So we have those that score for more than 2,000 images from that experiment.

So one of the first thing we need to know is, well, everyone is playing that game, and we can have their own memory score. But in order to know if some images are indeed memorable for all of us or a good amount of the population, we need to see if there is consistency between people.

So here is a simple measure we do. You have a group of people looking at those images, and we have the memory score. And you can split the group into two, and rank according to the average of the first group, the images, from the most memorable to the less memorable. And then you can also rank the same images in the second group.

And if the two group were identical to each other, you will get a correlation between those two ranking of 1. And what we observe when we split repetitively the group into two like this is we observe a correlation of 0.75, which is pretty high.

And this give us basically the maximum performances that we can expect when we have a group of people that can predict the rank of images of another group of people.

And actually, here is the curve that shows what the 0.75 consistency look like. So on the x-axis, you have the image rank according to the group number one, so the images with the highest memory score. So I have a group of them that are above 90. And it's normal.

The group number one in blue decrease as images are less and less memorable. So that's like basically your ground truth curve. And the green show the group number two, totally independent people, and also the performances that they got for each bins along the image rank. And you can see the curve are pretty close by. So a correlation of 0.75 look like this.

So it means that, with independent group of people, there are some images that are going to be systematically more memorable or forgettable. You can see the full range going below 40% for the images that were forgotten, up to 95 for the one that were systematically remembered. But, importantly, a group of person is going to predict another group of person.

So there's several ways to test memory. The way we tested memory here to get ground truth was an objective measurement. You see an image again, and if you remember it, you press a key. So that's an objective measurement.

But you can also ask people, do you think you will remember an image? Do you think someone else will remember an image? We also run those subjective memory score.

And we observed this very interesting trend that the subjective judgment do not predict image memorability, which means that, if you ask yourself, am I going to remember this? Well, basically, maybe, maybe not. So subjective judgment of what you think your memory will be or what you think the memory of someone else will be is not correlated with the true memory, with whatever you're going to remember or not.

So this was very interesting, because it shows that objective measurement should be needed here in order to really get a sense of what people will remember or forget.

We basically have many papers on this topic since 2010. They are all on the web site. And then some of them will look at the correlation existing between memory, so the fact that you're going to remember certain kind of images, and other attributes, for instance, aesthetic. And, again, we found that memorability is distinct from image aesthetic.

This means that basically you could have an image that is judged very beautiful, or, on the contrary, ugly or boring. And in those cases, you will still remember those two images. So we found this absence of a correlation between those two attribute in our values studies. We replicate this with other data set and faces as well.

So it looks like what you will remember is this notion of distinctiveness, but it can be beautiful

or ugly. It doesn't matter. You can still either remember it or forget it.

So you had this question about the notion of the lag. So the lag is in that you could test visual memory after a few seconds or a few intervening images, or you could test it a few minutes or even one hour later, or even days later.

So because we were running those experiments on Amazon Mechanical Turk, we did not do the dates. However, we did run some with a larger gap, up to 1,000 different images between the first and the second repeat.

And so here is the design, the one that I show you for the about 100 images intervening between the first and the second repeat. But what about a shorter and a longer time scale?

So all that work is also published. You can go and download the paper and see the details. But the basic idea is that the ranks were conserved. So if one image is very memorable, one of the top after, let's say, a few seconds, will still be memorable after hours. And if an image is forgettable or one of the most forgettable after a few seconds, will still be forgettable after hours.

So the fact that the magnitude, the percentage of images remembered decrease is normal. That is known from memory research for decades. So it is expected. However, memorability here is basically the rank, which is an image, is for a population independently basically of the row, the magnitude. One of the most memorable given this condition is forgettable.

And we did those experiment, both on the web as well as in the lab, because in the lab we could control for more factors. And it was very interesting to see in the lab experiment that only after 20 second there were images that were totally forgotten by a good group, a good amount of people.

So some image seems to really-- basically do not stick and be gone in I don't know how many second. We did not go really to short-term memory. We work starting in long-term memory at 20 seconds and so on. But there's this phenomenon.

So it suggests that there are some features into the visual information that are encoded in less details than others or with more details than others. And what's very interesting is then you can go to neuroscience and basically start studying the level of details or the quality of encoding of an image and see where in the visual pathway an image basically is gone after 20

seconds or something like this. So important point-- the rank is conserved.

So we also look at those principle of memorability that we found for images in faces. So faces is a very interesting material to work with, because basically it's all images that look alike. So you have basically one object and look at many, many exemplars.

And there's no reason to believe that this high consistency we will find for images as different as amphitheater and the parking and so on, and landscape, this high consistency will be found with faces.

So we gather a data set of 10,000 faces. The paper is published, as well as the entire data set is available on the web. So you can go and download those 10,000 faces as well as all the attribute that we found where we study with the data set.

And we also found the same phenomenon, very high consistency, both in the correct positive responses, when people remember seeing a face, as well as in the false alarm, when people did not see a face but falsely thought that they saw a face and basically pressed the key.

So this very high consistency for both measurement suggests that, again, in the facial features of people, or at least the way a photo is taken, there is something at the level of the image that will make a face highly memorable or highly forgettable for most people.

So all the details of this study are actually on the web. It was a pretty complex study to run, because while we have very different sensitivity to faces, so the race effect on basically where we grew up, so we know there's a lot of individual differences.

So the way this study was run is the collection of faces we have followed the US census in terms of male, female, race, and age. We started at 18 years old and older. So did our population as well.

So on the group, we show a collection of faces that did match the population, the people who were running the study. And, as I say, all the data are available on the web if some of you want to go back and do additional analysis.

So we kept going with this. So we have the consistency in the visual material. So now what about words? So words is a very interesting case, because now you do know the words. But are there some kind of words that we can predict are more forgettable or memorable?

Again, there's no reason to believe a high consistency will be found. But we've run the study twice with two different data set, again, ten thousand item, and two different set of words. And we found, again, very high consistency, which is that a collection of words were systematically remembered by people and others were forgettable.

So this work that is done in collaboration with Mahowald, Isola, Gibson, and Fedorenko is under submission. But let me give you a taste of the words that we found.

So what make words more memorable or forgettable? So here is a cartoon that give you the basic idea is, well, if there is one word for one meaning, so basically a word has a single meaning, it will have a tendency to be much more memorable than if a words has many meanings.

So in the paper, we also look at the correlation between memorability and image ability or frequency and so on. And all this is describe, but really the main factor is this one-to-one referent between a word and its meaning or concept.

So let's look at some of the example. The paper will come with the two data set and thousand of words that were found memorable and forgettable. So I think if you write a letter, you should not say that your student is excellent but that she is fabulous. And our research is not a blast. It's in vogue. And the idea of a team is not irrational, they are grotesque.

So those are just a few example of the words that on average, of the three first, had more referent and the tendency to be more forgettable because they might be used for many more things than the one. I also notice a lot of the French words tend to be memorable.

So we do find this stable principle that you can predict the content, can predict what type of images, faces, words that are memorable. And we also did it for visualization, starting working on the topic of education.

So this is very useful, because at least at the level of a group, you can start making that prediction. Oh, I also have massive and avalanche. Forgot about it.

So now that we were able to have all those data and see that there is this consistency, then one of the next question you can look at is, OK, well, if it seems that we all have a tendency to remember the same image, can we find a neural signature into the human brain?

So the question of memorability, is it a perceptual or memory question? Because in all our

experiment, the images are shown for a short time. And then, when they are repeated, you see them a second time.

But basically, all the action is at the perception level. Whenever you perceive this image for half a second or one second, there is something going on here at the perception level that is going to bias if this image is going to basically go into memory or not.

So, knowing this, if we want to look at the potential neural framework of memorability, we have to look at the entire brain. We have to look at all the region that have been found to be related to perception, faces perception, picture perception, object, space, and so on, as well as the medial temporal lobe region, more in the middle of the brain, that have been related to memory.

So this is what we did with Wilma Bainbridge. This is her PhD, basically having a look at all those region. And here is the very simple experiment we run.

So we took a collection of faces and scene from the thousands we have, and where every one, we'll basically separate those two between looking at the region that are more activated for seeing versus the region more activated for faces. We split it that way, between the memorable and the forgettable set.

So in those set, every images is novel. So exactly like in the memory experiment, we could show them one time for half a second. So you had the perception level. You're in a perception experiment. You saw those image one after the other only one time. All those images are novel.

So we're going to look at the contrast from novel image minus novel image, from scene minus scene, from faces minus faces, except that some images are highly memorable and some are highly forgettable.

So other factor that you have to look at is, it's still possible that within those group of images and faces that are highly memorable or forgettable, that there's a lot of image features that basically correlate with those.

So if you take a collection of images like I shown you before, and did the environment or the photo that have people or action tend to be memorable versus a landscape tend to be forgettable? So here you have a lot of visual features that will co-vary with the dimension of memorability.

So in that study for the brain, we equalize for that, because we had enough images. So here you have a sample of the two groups and sample of images and the type of statistic we look at.

And the two, for instance, for this scene were equalized in term of the type of category you had-- outdoor, indoor, beach, landscape, house, kitchen, and so on-- as well as a collection of low-level features.

And you can see some of the average signature that are actually identical on a lot of low to mid to higher level image features that were equalized between the two group. So whatever we find is not going to be due to simple statistic due to the image or the type of object those have.

We could play the same game for the faces, so we did. Here are the numbers again, memorable and forgettable faces that were also equalized for various attribute like attractiveness, emotion, kindness, happiness, and so on, as well as male, female, race, as well as expression, and so on. And you can see that the statistic, the average faces for both the memorable and the forgettable group, are actually also identical.

So with those group, what's left is hopefully only the factor of something else in the image at the level of a higher image statistic, because only image statistic will explain the fact that very different people will remember the same faces and forget the same faces. But certainly not some of obvious low-level image statistic. Those cannot explain any result.

So two years and four study later, we replicated basically this study four times with many different matter. Then, I'm just going to show you here one snapshot of the result.

So this is a Multi-variate Pattern Analysis looking at the memorable versus the forgettable groups, searchlight analysis, MVPA looking for the region of the brain that are more active. Well, that have a different pattern. They are also more active, but at a different pattern for memorable faces.

And we find signature in the hippocampus, the parahippocampal area, as well as the perirhinal that are typical for memorable faces and scene, and a new signature in the visual area or even the higher visual area, because we did equalize for those.

So it seems to show that those MTL region play a role in a kind of higher-order statistical

perception, a notion of distinctiveness that is within those image.

So this suggest that at the perception of a new image, could be a face or a scene or a collection of object and so on, well, there's already a signature of this that's going to guide or bias if this is going to be put into memory or not. And we have those at the level of the group, looking now at the level of individuals.

All right. So, well, it looks like we're going, given your question is now trying to model this notion of memorability. So we have a good case that there is some information into the image of a higher-level status-- we don't know which one-- that cannot be explained by simple features that make all of us reacting the same way. Even our brain do react the same way. We also have images, signature of memorability coming up.

So we have this intrinsic information that make all of us remember or forget the same kind of visual information. So can we now in a way imitate or model those result into an artificial system?

All right. So you have all heard about the revolution in computer vision a couple of years ago and deep learning and those neural networks that are now able to outperform a lot of-- that were able to recognize and perform a lot of task, some of them at the level of humans, so recognizing various object and so on.

And one of the aspect of those neural networks, and I'm going to talk about them, is that they require a lot of information. So you need to teach them the classes you want to distinguish. And they can and need a lot, a lot of data.

So everything we did so far, we are getting that on a couple of thousand images. And that's really not enough to even start scratching computational modeling. So with Aditya Khosla, we run recently a new large-scale visual memorability study on Amazon Mechanical Turk, but this time getting score for 60,000 photograph.

And you have a set of a sample over here. 60,000. They are all going to be available in a few weeks. The paper is under revision. It's looking good. So as soon as we have a citation, we are going to give away all the data as well as the score and the images and so on.

And in this experiment, images were presented for 600 milliseconds or shorter time, but they really did not change much. So as a snapshot, because the 60,000 images really cover a lot of

type of photo-- faces, event, action, and even a graph and so on-- well, you either hear some that were highly memorable or forgettable. There's also the website. It's not populated yet with many things, but it will be very shortly.

And the correlation we got on that data set was pretty high again, 0.68. That was expected. And the paper explain the various split we did. But, again, we find this very high correlation. So we do know that there's something to model there.

And other-- again, a very quick summary. The type of images that seems to be the most memorable are the one which have a focus and close settings, that show some dynamics and something distinctive, unusual, a little different, whereas the less memorable one seems to have no single focus, distant view, static, and more commonalities.

All photos, you can still find two images that will be focus and dynamics, and one of them will be more memorable than the other because it will have something unusual that now our system can capture.

So we have this new data set. We have all the memory score. We have the high consistency. How do we even start thinking about the computational model of visual memory or memorability?

Well, in order to give you a sense of one of the basic we needs, in order to even start thinking of a model, I'm going to show and run another demo. So in this demo, you're also going to see some images. And clap your hand whenever you see an image that repeat. OK? Exactly the same game than before.

Ready? All right. If everyone play the game, it will be fun. OK.

[CLAPPING]

[CLAPPING]

A little false alarm.

[CLAPPING]

False alarm.

[CLAPPING]

Good. More energy.

[CLAPPING]

Good.

[CLAPPING]

Sorry.

[CLAPPING]

No

[LAUGHTER]

[CLAPPING]

No.

[CLAPPING]

Yes.

[CLAPPING]

No.

AUDIENCE: Too close.

[CLAPPING]

AUDE OLIVA: No. Yeah, that was-- yes.

[CLAPPING]

[LAUGHTER]

All right. So for the sake of the demo, I put here really images that are very different, some you are familiar with. You have a concept. You know what it is. This is a restaurant. This is an alley. This is a stadium, and so on. And some for which either you don't have a specific

concept, or you have the same concept-- texture, paintings, texture, texture, texture.

And the basic idea of memory is you need to recognize, to put a unique tag or collection of tag in order to remember that individual image. So the fact that you saw a collection of texture or paintings or whatever you want to call them, you'll remember that as a group.

But to go to the individual memory of one, you're going to need to have a specific concept. And in order to remember it, this is going to be an abstraction, a format, a collection of words, or a coding that make it unique.

So you need to recognize to remember, which means that if you want a model of memory which start from the raw image-- I'm not in toy world here. I really start from the raw image, like the retina. Well, you're going to need to build a visual recognition system first.

So first you need a model that recognize object and scene and event and so on. And then, from this, there can be a base to start modeling memory.

So, fortunately, the field of computer vision made a lot of progress in the past couple of years. So now we do have visual recognition system that works pretty well. I'm going to describe them.

We need to first a visual recognition system. All right. So, what does a visual recognition system needs to do? So, well, it's your Sunday morning. You're going to the picnic area, and you're faced with that view. You take a photo. This photo actually became viral on the web.

And here is the state of the art of computer vision system. When it comes to recognize the object-- I know it's a different view, but it works very well for any view-- object recognition for about 1,000 object category is reaching human performances so far.

And so this will tell you this is a black bear, there's a bench, a table, and so on, and trees in the background. But it's missing the point, that this is a picnic area.

So you do need at least two kind of information in order to reach visual scene understanding. You need the scene and the context, you need to know the place, and you need to know the object.

So, as I said, so far on the challenge that's called the ImageNet Challenge in computer vision, computer vision model average human performances, which is 95% correct on exemplars of

objects that have never been trained on, a new one, for 1,000 object category.

And recently, we basically published a few papers for the other part of visual scene understanding, the place and the context. And this is an output of our system. And you can go on the web-- I'm going to give you the address-- and play with it and see the performances of recognizing the context and the place.

So just to put into context what the field of computer vision have been doing for the past 15 years is, well, the number of data set, the number of images by data set has been increasing, so that there's more exemplar to learn from.

And, in perspective, you can see that two-years-old kids. Of course, it will depend on the sampling you use in order to have an estimate of the number of visual information that the retina sees. But it sees much, much more variety and numbers of visual input.

But right now, both ImageNet, that is, a data set of objects, and Places, which is a data set of scene I'm going to present to you now, have about 10 million label images of many categories. So label means that for the places, it will tell you, this is a kitchen or this is a conference room, and so on for hundred of categories.

So 10 million is largely enough to start building very serious visual recognition system, but it's no near the human brain. However, we might be getting there. So how do we even start to build a visual scene recognition data set?

This is a work we did and published in 2010, the Scene Understanding, or SUN data set, where we collected the words from the dictionary that correspond to places at the subordinate level-- I'm going to give you example-- and then retrieve a lot of images from the web. And there was a total of 900 different categories, and about 400 of them have enough exemplars to build a artificial system.

So instead of going and only looking to build images, like to build a data set of a bedroom and kitchen and so on, what's happening for the human brain is that you have a different environment. You see there's many attribute you can put in. And you are forthcoming and storytelling and so on.

So most of the places, it's not only that this is, for instance, a bedroom, is that you will go more to a student bedroom or-- Well, I think those are student bedroom two doors from my colleague.

So there are many type of adjective that we can use in order to retrieve images that will give us a larger panorama of the type of concept that are used by the human brain in order to recognize environment.

So a simple bedroom. The tag were put automatically. Superior bedroom. Senior bedroom. Colorful bedroom. Hotel bedroom. And so on and so on.

So that was the retrieval we did, which means that now, for every category there is also a tag in term of the subtype of environment this can be. Messy bedroom.

And a couple of years later, 80 million images later, and a lot of Amazon Mechanical Turk experiment, we are launching this week the Places2 data set, with 460 categories, different categories of environment, and 10 million images label.

So this is a larger data set of label images, with a label to be used right away for artificial system learning, deep learning, and so on.

So here is just a snapshot of the differences of the places in term of the number of exemplars with other a large data set. So the Places data set is actually part of the ImageNet challenge this year, which means that you can go to ImageNet and register for the challenge and download right now, tonight, eight million images of places to use for the learning of your neural networks, as well as of a set that will be used for the testing, and participate to the challenge.

So this was launch last week, and the website associated with Places will be launched this week. And we decided to just give this away to everyone right away. We are finishing up the paper now. It will be an archive paper.

No time to wait for month and month. This is a data set that can be used by a lot of people to make progress fast. And so that's what we are doing.

So, as I said, the computer vision model now require, if you use deep learning, a lot of data. And we hope that with this data set, fast progress are going to be made.

So what we specifically did is using the AlexNet deep learning architecture-- If you don't know what this is, I can tell you later how to basically access to it with the paper on so on. This is not my model. This is a model put together by Geoffrey Hinton and collaborator a few years ago.

And you can download the model or download the code and re-train.

So neural net now basically are based on the collection of operation that are call layers, convolution, normalization, simple image processing operation that you do in a sequence. You do it one time, then a second and third and so on. And then you have those multi-layers models.

And the number of layers is still a question of research. How does layer correspond to the brain? I'm going to say a little bit about that.

And using this simple-- well, this AlexNet model, we built a scene recognition system. And now you can go to places.csail.mit.edu with a smartphone will work, will take a photo. And it should tell you the type of environment the photo represent.

It will give you several possibilities, because basically environment are ambiguous. They can be of different type. So I don't know if you can read. Let me read a few. The first one, it says, restaurant, coffee shop, cafeteria, food court, restaurant patio. I guess they all fit.

The second one, parking lot and driveway. The third one, conference room, dining room, banquet hall, classroom. That was a difficult one. And the fourth environment is patio, restaurant patio, or restaurant.

So if you go there, you can also give feedback if one of the label match the environment that you're looking at. And it should be above 80% correct. And this model use 1.5 million images and 200 categories. So soon with a great-- we hope that things will be even more interesting and accurate.

And I took this morning a couple of photo at breakfast. So you may all recognize the scenery here. So from the breakfast area looking outside, it's an outdoor harbor, dock, boat deck. Yeah, it could be actually on a boat deck looking at the harbor. And otherwise, the breakfast area was restaurant, cafeteria, coffee shop, food court, or bar. All those, again, fits.

So those model now works very, very well. So why? Well, let me tell you why. So those neural networks, you can go to any layers and open them and look at what every single artificial neuron do. So what we call the receptive field of every single unit in the layer one, the layer two, the layer three, and so on.

So the first the layer-- here, four layers are shown-- basically learn simple features, contours

and simple texture. That's called pool1. And those looks like the type of responses of the visual cells, and possibly V1, V2. I'm going to tell more about this.

And as you go higher up in the layers, then you start having some texture, some patches, that make more sense. And higher up in the layers, like layer number five, then you start having artificial receptive field that are specific to part of object, part of scene, or an entire object by itself. Like we can see here some kind of [INAUDIBLE] coffee, as well as the tower and so on.

So it seems that the system are able to recognize environment of object. But what they learn are the part and the object that the environment contain. And I'm going to show example of those.

So jumping, just giving you a result in neuroscience. And so there's a lot of debate out there about, OK, well, there's those model with different layers. And you have different models out there. To which extent they correspond to the visual hierarchy of the human brain?

Well, first, the computational model were inspire by the visual hierarchy, the V1, V2, V4, and [INAUDIBLE] and so on, knowing that more complex features are built over time and space.

So what you can also do is run through a network and run in an fMRI experiment the same image, and then look at the correlation existing between the responses of the cells, let's say in layer one, and the responses you may have on the human brain in different part of the brain.

And what we find is that the layer one will correspond more, will have a higher correlation, with responses in the visual area, literally V1, V2. And as you move up through the layers, then there is correlation, higher [INAUDIBLE] correlation, with part of the ventral and the parietal part of the brain.

And I know you had the lecture by Jim DiCarlo that actually must have explained this. So Jim DiCarlo team did it, Jack Gallant as well in Berkeley. And we also did it with other type of images and other network, and all the result really corroborate each other with this nice correlation between low to high visual areas between the brain, the human brain, and those multi-layers model.

All right. Let me show you some of the receptive field, the artificial receptive field that we find in the higher layers. So this network was trained for scene categorization. So the only thing that this network, the one we are using here, learn was to discriminate between, in this case, 200 categories-- the kitchen, the bedroom, the bathroom, the alley, the living room, the forest, and

so on and so on. 200 of those. So that's the task.

What the network learn and what we observe is that object discriminant and diagnostical information between those category form the emerging representation that are automatically, naturally learn by the networks.

So the network has never learned a wheel, but the representation emerge, as you can see here. So this is one artificial neurone. And it is receptive field and its responses, the highest response it got for a collection of images.

And as you can see, those higher pool5 receptive field are more independent to the location. They are built that way. But the network never learn the parts. This is something that emerged naturally by learning different environment.

The other thing that's very interesting in this model, when you can open up and look at the receptive field using various method, here is one, is, well, this model has never learned the notion of shape or object. So it's going to basically become sensitive to discriminant and diagnostic information.

So you have this unit that is discriminant to the bottom part of either the legs of animate, or even you can see the trees over there. So this is a unit. It seems that it was needed in order to classify environment to have units that we might not have a word for it. But the human brain might as well have many of those unit that do not correspond to necessarily a word.

So, basically, with this model you can have a lot of new object emerging that you might not have thought of, but they become parts of the code needed in order to identify an environment or an object.

So we also have another unit for this bottom parts of a collection of chairs. We also have chair, of course, showing up. Faces. The model never learned faces. Only learn kitchen, bathroom, street. We have several unit emerging for faces. Why? Because those are correlated with a collection of environment.

Then, entire object shapes, like bed, very diagnostic of a bedroom in this case. So that will be a unit that is very, very specific. That would be really only for beds, that one.

And then, others like lamps and so on. There's thousand and thousand here. Another unit never learned, screen monitors. And here is specific unit for that that emerge. Also, collection

of object or space, collection of chairs over here.

The network found that this is a discriminant information to classify environment, crowds. It's a very interesting unit because it's really independent of the location, as well as basically the number of people and if they are closer or further away. But it does capture this notion of crowd.

So the model doesn't have the word crowd. So it has never known the crowd. It's one of the unit emerging that now can be used as an object detector to enhance the recognition of what's going on in the scene. So this is an ice skating area, and there is a crowd.

And also unit that are more specific to space and useful for navigation, for instance. We have many of those. So in that case, just the fact that there is lamps up or perspective, so we have a unit like this specific to this.

So it's not the only object as physical object. There's also a collection of unit that are related to spatial layout and geometry that are also discriminant for environment. And many, many other that are showing up. So those object detector naturally emerge inside this kind of network trained for scene understanding.

So now I only have a couple of minutes to wrap up, 20 minutes, so I'm going to just give you the hint of the next part. So with the Places challenge that is starting this week, certainly in less than a year the computational vision model of scene recognition are going to be very close to human performances.

And then there's a long way to go and many more things to match, like the error. So know that the error look alike or when a category can have many type of object or what can happen next. So you can really expand.

But let's say we can consider now that we have a base of visual recognition into a model that works pretty well at the level of human, or close enough, or will be close enough. So, now that we have that, we can add the memory module.

How to add the memorability module is really open in the air. There's many ways to do. So we just did it one way, to have a ground very first model of visual memorability at the level of human.

And this is going to be out-- the paper is in revision-- in a few weeks. And you're going to be

able to download the images, the model, and so on. Again, this is model number one, and we hope that then a better model can be done.

So we went for the Occam razor approach. The most simple one given the model is we took AlexNet, we feed AlexNet with both ImageNet and Places. Because all the images that are memorable or unforgettable, they might have object, they might have places. OK, so let's put the two together so we have more power.

We train the model, so we have the visual recognition model. We do know that those units make sense. They recognize that this is a kitchen. And we do know why, because we have the parts. So that's a classical standard AlexNet. The output is scene and object categorization, so we are still in categorization land.

And we can remove the last layer. And then this is a procedure that has been well published in computer vision and computer science, this notion of fine tuning and back propagation. So you use the network on learning that are learn to recognize places, and you finely tune and adopt the feature so that the task has change.

So the task was recognition. And now, for the network, at the end there is a new task for that network that has learn all those object and scene. And the new task is now to learn that those element are of high, medium, or low memorability, which is a continuous value.

So by doing this, we have now a model where we give it a new image, and it's going to output a score between 0 and 1. And the human to human I'll show you is 0.68 correlation. The human to computer is 0.65, which mean now that there is a first model that can basically replicate human memorability of a group nearly at the level of a human.

And we use the data set of 60,000 image to finely tune the recognition model, as well as test with images that this has not shown.

And because we can open up this new network of memorability and look also now at the receptive field of the neuron that are related to high or low memorability image-- and we will publish every single receptive field on the web as well with this paper-- and thus now we can see the unit that are related, this higher level information of object or space or parts and so on that is related with images that are highly memorable in green, strong positive, or highly forgettable.

And we find, again, that if you have animate object or kind of object, roundish object, and so on, you can go [INAUDIBLE] those will make your images more memorable.

And so our last part is, so now that we do have this model that spill out responses at the level of human and indicate the parts that are related to higher memory, lower memory, at least as a guideline, then we can go back to a given image and emphasize the part that correspond to the high memorability receptive field and de-emphasize the part corresponding to the forgettable part of the receptive field.

And this give you images on the right like that that have been weighted by the element that are memorable and the element that are forgettable.

So maybe here it doesn't matter that the ground is forgettable. Here, the part I like to emphasize are the memorable one. But, for instance, in this imagine we have two person. And, well, she just happened to be more forgettable in this case, which will make a perfect CIA agent, if we think about it. And we did tested those.

So then, for a values part of navigation, I mean values scene, we do find that the element of exit or entrance, so where there's basically path and a 3D structure for navigation, tend also to be more memorable. So those are highlighted in those images.

Another example, where the kids will be more memorable than the features of the person. And I don't-- Well, I'm not an expert in game. You can explain that one to me.

So I have to stop because we need a break, and I might need to talk. However, here is a vision of where we are going. And maybe other people will be interested to go on this adventure. Because it's really, really just the beginning.

But if you now can have a model that at the level of a group of human predict which image are memorable, and as well, also match part of the visual region and even the higher level object recognition, then you can start having this similarity going on in this study between the human brain and all the part, from perception to memory and computational model, to characterize what the computation underlying those particular region. Because now you have model zero.

I'm not saying that the model I'm showing to you is a model that imitate the brain. But it's a model zero, and now it can be tuned to better learn the calculation that are done at different level along the visual to the memory hierarchy in order to characterize visual memory.