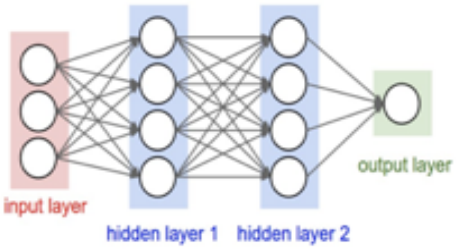
Machine-learning system tackles speech and object recognition, all at once

Humans have always taken for granted how easily we are able to associate spoken words with the world around us. As children, we have no way of communicating. However, over time, we pick up the languages that we hear and gain the ability to express ourselves. This phenomenon is especially impressive as children are unaware of written language. Even without the ability to read or write, we learn how to converse coherently.

To date, computer scientists have been unable to emulate how humans learn new languages. Hours of manual labor and large amounts of textual data are required to successfully train a Machine Learning model today. This process can be extremely time consuming and wasteful. To avoid the tedious process of collecting textual data and edge closer to mirroring how humans absorb new information, MIT computer scientists have developed a new system for speech and object recognition. This system correlates segments of spoken captions with objects within an associated image. Relevant regions within these images are highlighted in real time as a spoken caption is played.

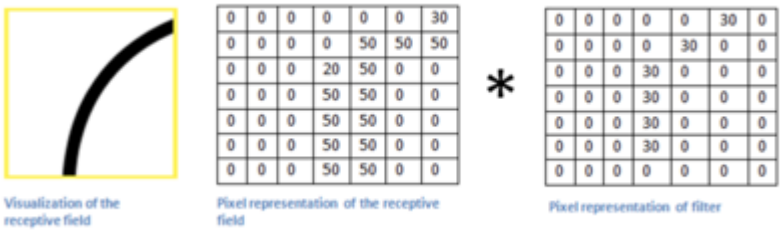
The main concept behind this system is neural networks, specifically convolutional neural networks. A neural network is, essentially, a computational and algorithmic attempt to imitate how the human brain works and learns: through reinforcement, association, etc…. At its core, neural networks need only some basic linear algebra and calculus to work. The essence is this: An input layer takes in a series of inputs, that are passed into nodes, which are the fundamental units of neural networks. These nodes can take in multiple input channels, and each node does a weighted sum of the inputs (note, each input channel to a node can have different weights). Then, these results can be passed to any number of nodes in any number of layers, resulting in a singular value or a list at the end.



The way a neural network, therefore, “learns” is to update the value of its weights. The main method of training is by checking error margin(anyway) and back-propagating through the weights and updating in the direction of minimizing error, found by gradient descent. Now, onto CNN’s(convolutional neural networks).

The basic layout of a CNN is similar to a basic neural network, except it is specially tailored to detecting features(more on this later) within an image. Here is how it works. An image is

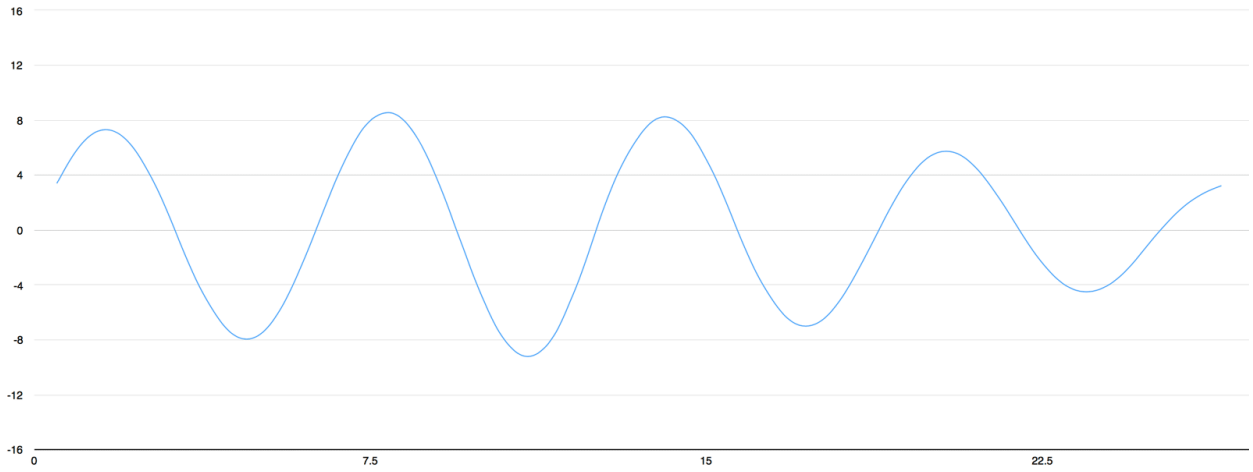
essentially a 3-D matrix, with NxNx3, with N being the number of pixels, and 3 representing the 3 values: R for red, G for green, and B for blue. An image, after all, is any combination of these 3 colors on a scale from 0-255 for each pixel, hence the 3-D array/matrix. The CNN takes into it as input a region of an image, called a receptive field, for every instance. This smaller 3-D matrix is multiplied by another matrix of weights, which are designed to detect a particular feature. Much like neural networks, the value of the element-wise multiplication can tell us whether a particular feature is present. Look at the example below for a better idea of how this works(https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks/) :

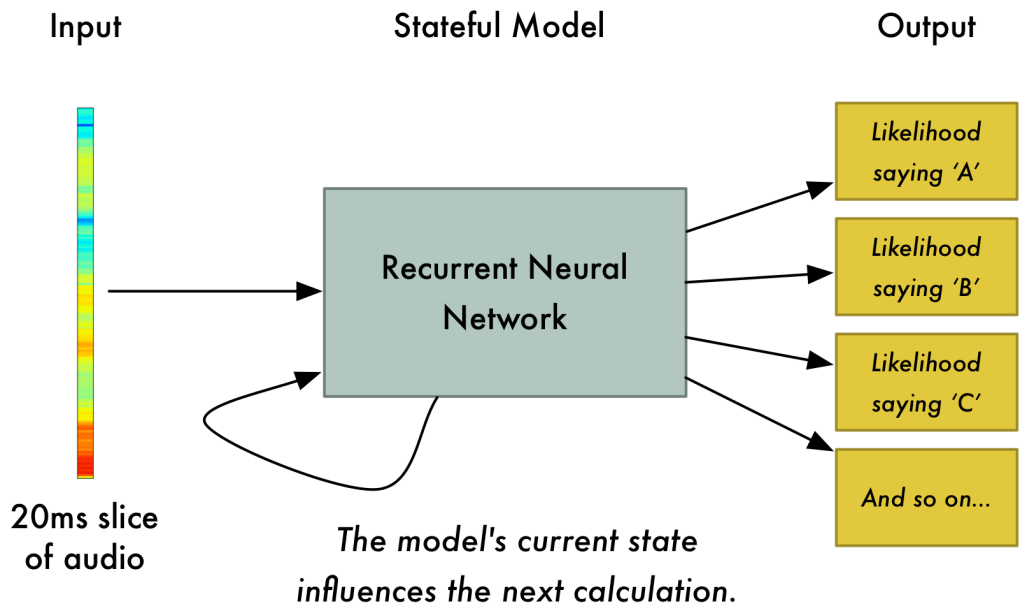


You then move over say one pixel and continue with the same process. The multiplication results in a singular value, but over the entire image, this can generate many data points passed to the nodes in the next layer. This overlap is what makes CNN's contextually aware. The reduction of the larger matrix into a summed up smaller matrix is then fed into several different layers like ReLU (Rectified Linear Units; help with quicker and more accurate weighing) and Pooling(kind of like downsizing or sampling) layers, but the general purpose of these is to simply help keep order to the images/information, like maintaining dimension or introducing nonlinearities. After all this, the last layer takes in information from whatever came before, and the last step for processing is the fully connected layer. This essentially takes in whatever input volume it has from the layer before and returns an N sized vector of probabilities, where each i in N represents the number of classes the program has to choose from for classification of the object. The way it does this is, in high level, is that it looks at the previous layer output (the feature/activation map) and determines which features correlates best with which classes. Training is done with an image and the correct result and is calculated as explained earlier.

Speech recognition today is dealt with in a manner very similar to object recognition using CNN's. Audio recordings are first converted into wave graphs call spectrograms. These spectrograms are then converted into a series of numbers which correspond to the height of the wave at different points of time. In order to make this data easier for a neural network to process, it is broken up into its component parts using a method known as the Discrete Fourier Transform. Different frequency bands in the spectrogram are isolated and the energy in each of these bands is eventually added up to obtain a fingerprint of sorts for the audio recording. This data is then divided into several segments at equal intervals of time and fed to a convolutional neural network. The neural network relates each segment of the spectrogram with letters or words in a transcription. It ultimately generates an N-dimensional vector with the likelihood of a specific segment of the spectrogram matching with a specific character or word. Thus, to some extent, such models make it possible for computers to understand human speech.

(https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a)





Current speech recognition technologies like Siri require transcriptions of thousands of hours of audio recordings. The new system introduced by MIT's computer scientists eliminates the need to train data on manual transcriptions. It seeks to achieve speech and object recognition through more natural means. The model attempts to directly relate snippets of audio with regions within a corresponding image. This is especially impressive as the model has no access to any true alignment information between the spoken caption and the image. The system essentially learns to relate speech with objects in a manner almost akin to humans. Just as in Speech Recognition models we use today, a spectrogram is created for an audio recording. The spectrogram is converted into numeric data, simplified using the Discrete Fourier Transform

method and fed into a Convolution Neural Network. MIT's model is unique because of the behavior of the neural network. The neural network relates each segment of the spectrogram with a patch of pixels in the corresponding image. The N-dimensional vector that is generated holds the likelihood of segments of the spectrogram matching with specific patches of pixels in the image. Thus, the system devised by MIT's computer scientists is trained directly on audio and visual data without any need for manual transcription.
(https://news.mit.edu/machine-learning-image-object-recognition-0918)

The MIT method discussed in the article has been coined match map by the team. In essence, it is running 2 different convolutional neural networks. One is on the image itself, and the other is on the spectrogram as discussed above. Essentially it is an exhaustive search and compares for the two segments: a sample of the spectrogram at a fixed interval with a receptive field of the image. The main discovery made is that the only reinforcement/training needed is a correct or incorrect mark for the two networks to be able to back-propagate and correct the error. From the paper: " The idea that a pooled representation over an entire input used for training can then be unpooled for localized analysis is powerful because it does not require localized annotation of the training data, or even any explicit mechanism for localization in the objective function or network itself, beyond what already exists in the form of convolutional receptive fields" (http://openaccess.thecvf.com/content_ECCV_2018/papers/David_Harwath_Jointly_Discovering_Visual_ECCV_2018_paper.pdf).

This model has a wide array of possible applications in the future. Although seemingly impossible, this system could potentially be used to generate translations from one spoken language to another without the need of a bilingual annotator. Future world conferences may not need translators as this system could simply convert one spoken language to another such that people from all parts of the world would be able to understand each other. This could be achieved simply by training the model with spoken captions in multiple languages. Finally, another possible application of this model is automatically generating spoken descriptions for images or alternatively generating an actual image based on a spoken description. This system could potentially change the way industries operate. It could be truly revolutionary.

Devyan Biswas (UID: 804988161)
Rishab Sukumar (UID: 304902259)

Bibliography

- https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks/
- https://news.mit.edu/machine-learning-image-object-recognition-0918
- http://openaccess.thecvf.com/content_ECCV_2018/papers/David_Harwath_Jointly_Discovering_Visual_ECCV_2018_paper.pdf
- http://www.ai.mit.edu/courses/6.899/papers/ForsythECCV-02-1.pdf
- https://news.mit.edu/2016/recorded-speech-images-automated-speech-recognition-1206
- http://news.mit.edu/2009/explained-fourier
- https://news.mit.edu/2015/learning-spoken-language-phoneme-data-0914
- https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a