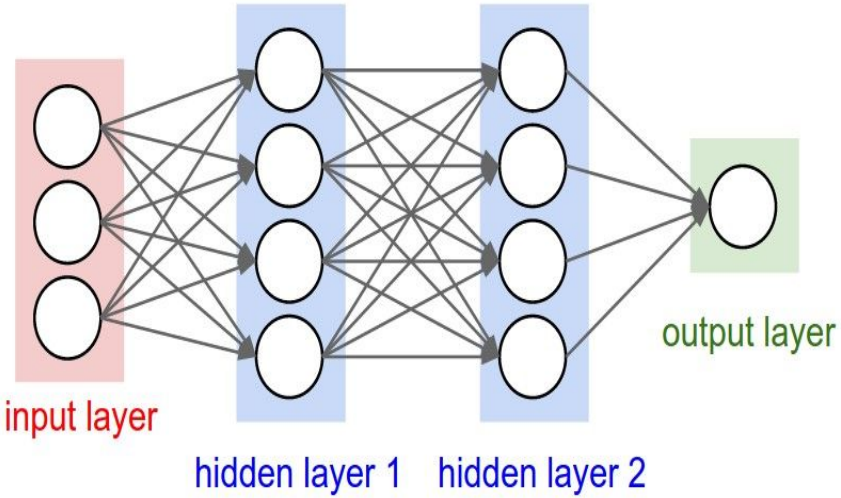
Machine-learning system tackles speech and object recognition, all at once

Devyan Biswas and Rishab Sukumar

General Overview

- Correlate segments of spoken caption with objects within an associated image.
  - Highlighting relevant regions of images *in real time*
- Demo: https://vimeo.com/290377425

Object Recognition Today: What are CNN's?

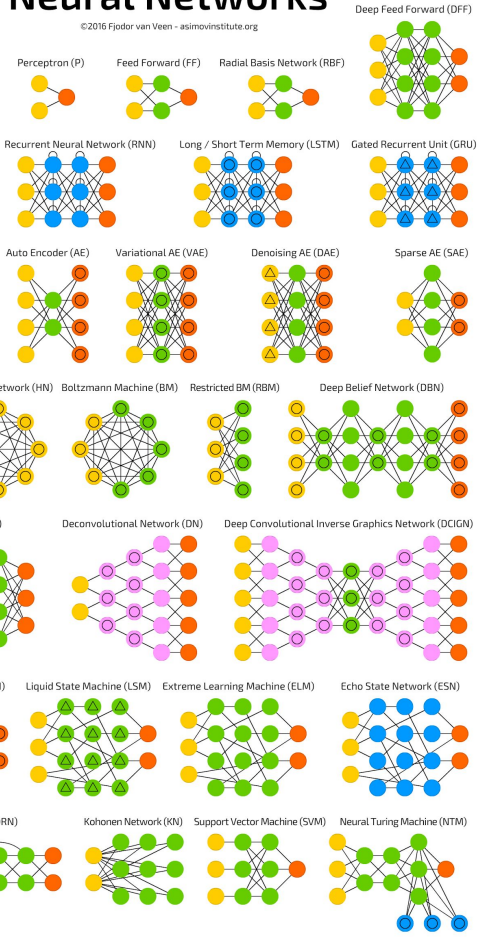


- Convolutional Neural Networks
  - So...what are neural networks?
- Fundamentals: Linear Algebra, Calculus(depending on model)
- Neural networks are models based on the human brain, have nodes that act like neurons
  - Like how humans learn: training and association
  - Interconnected nodes in layers, feed forward; not necessarily 1:1
  - Uses weighting with incoming connections to get value; sends or doesn't send data.
  - Weighting is adjusted as it is training

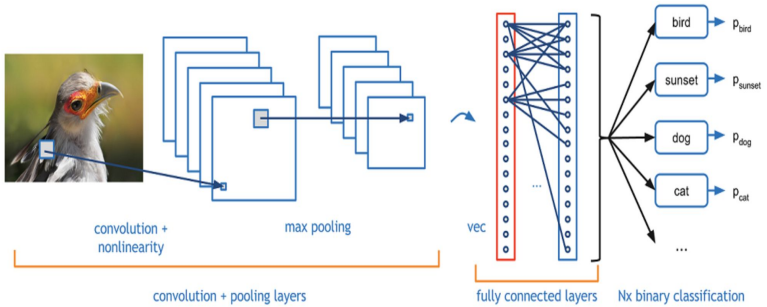
A mostly complete chart of

Neural Networks

- Backfed Input Cell
- Input Cell
- Noisy Input Cell
- Hidden Cell
- Probablistic Hidden Cell
- Spiking Hidden Cell
- Output Cell
- Match Input Output Cell
- Recurrent Cell
- Memory Cell
- Different Memory Cell
- Kernel
- Convolution or Pool



What are CNN's? (cont...)

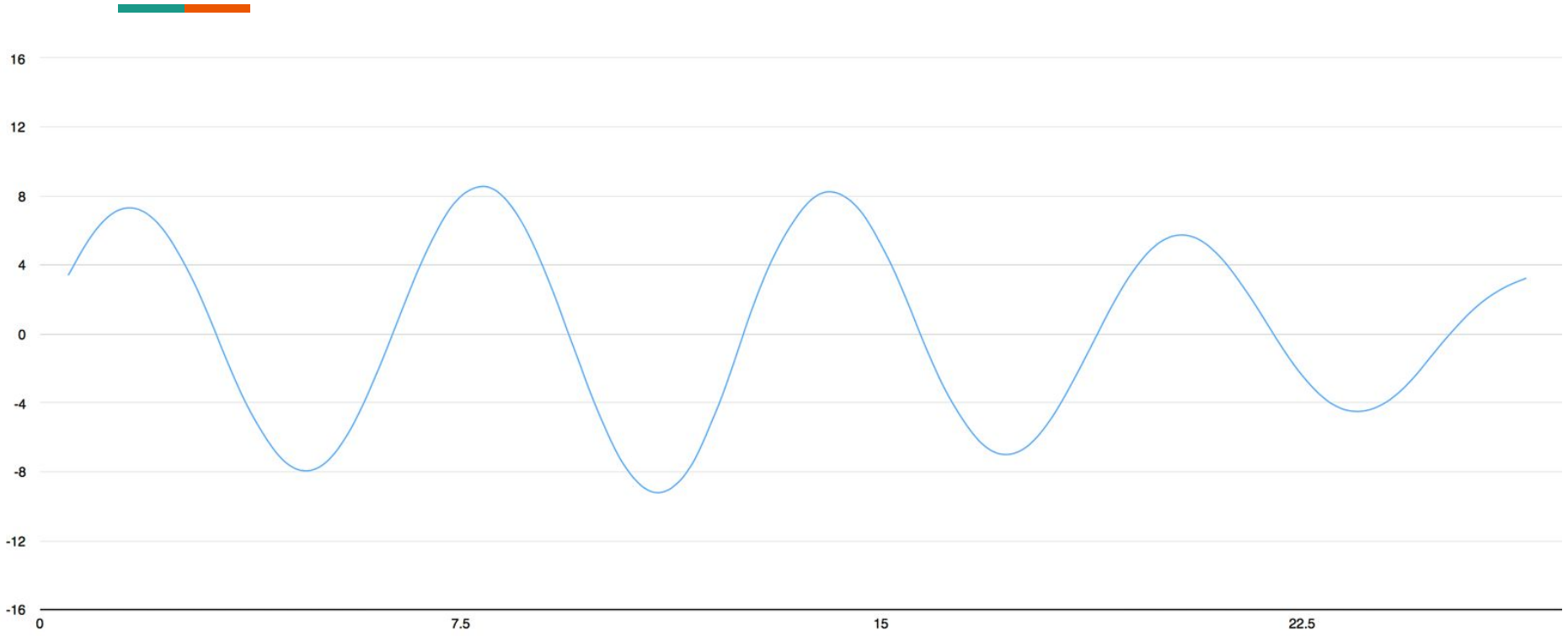


https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks/

- Used in image processing and object recognition
- CNN's are useful in 2 regards
  - Reduce total units (generally) because they translate many:1 , so quicker learning (generally)
  - Contextual
- What does "convolution" mean?
  - Using a filter (feature identifier/neuron), convolving over input image
  - Performs element-wise multiplication
  - Sums and returns to one point
  - Gives you activation/feature map
- Last output: Fully Connected Layer
- Training with backpropagation

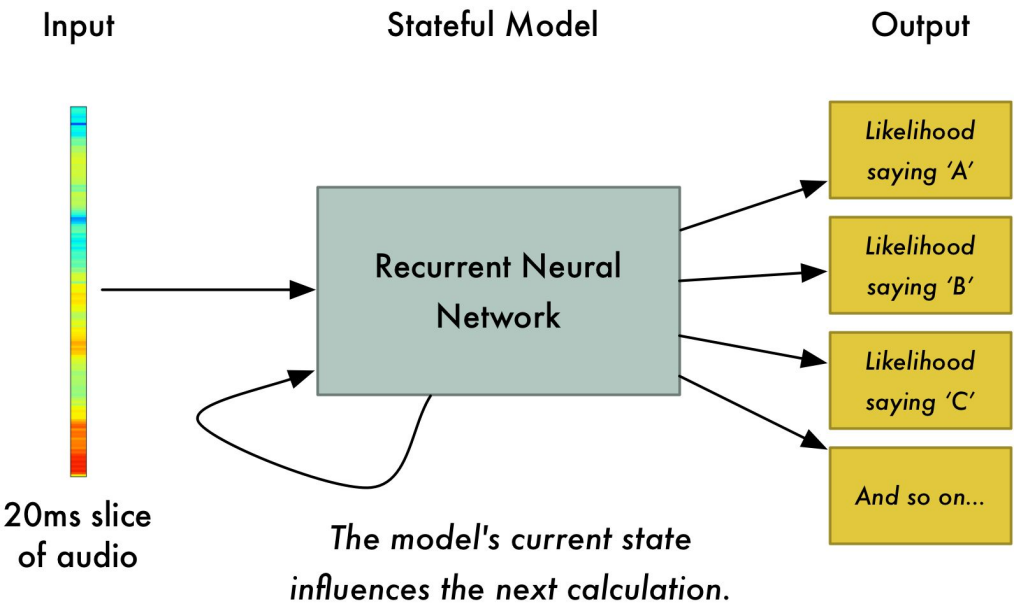
Current Speech-Recognition Technologies

- Require manual transcriptions and annotations of training set images.
- Systems such as Siri require transcriptions of many thousands of hours of speech recordings.
- Uses spectrograms of audio recordings.
- The spectrogram is converted into a series of values based on the height of the sound wave at fixed intervals.


16
12
8
4
0
-4
-8
-12
-16
0
7.5
15
22.5

Current Speech-Recognition Technologies (Cont…)

- Uses short samples of spectrograms as input for a Recurrent Neural Network (i.e. a neural network that has a memory that influences future predictions.)
- Output of RNN is an N-dimensional vector where each element denotes the likelihood of related classes.


Input
Stateful Model
Output
Recurrent Neural Network
Likelihood saying 'A'
Likelihood saying 'B'
Likelihood saying 'C'
And so on...
20ms slice of audio
The model's current state influences the next calculation.

How is MIT's model different?

- Eliminates need for manual transcriptions and annotations for training.
- Focuses on a more natural method of learning where knowledge of written text is not necessary.
- Training set consists of numerous audio recordings describing the contents of an image.
- Uses two CNN's. One to process images and the other to process spectrograms.
- Matches words to specific patches of pixels in the image.
- The model has **no access** to any true alignment information between the speech and the image.

The MatchMap Concept

- What is the matchmap?
  - Correlates sections of audio data(spectrogram) with image data
  - Able to directly learn what corresponds between speech frame and group of pixels
  - No need for annotated training data
  - “Reversible”; Unpool localized information analysis from pooled representation of input
- Results (from Academic Paper):
  - Speech-Prompted Localization
  - Image-Caption Retrieval
  - Image-Word Dictionary

Possible Future Applications

- Speech-object recognition in videos.
- Learning translations between different languages, without need of a bilingual annotator, possibly in real-time.
- Generate images given a spoken description.
- Generate artificial speech describing a visual scene.

Questions?

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