# Part C: capsule networks

CNN is a network of neurons that uses Convolutions to decide. CNN is a depiction of brain visual cortex working, to make decisions. CNN is majorly used for image recognition. The concept of CNN is derived by the amazing work by David H. Hubel[1] and Torsten Nils Wiesel[2] that confirmed that the brains visual cortex contains neurons that help the brain to recognize the signals from the eyes in the form of images. The work by David H. Hubel and Torsten Nils Wiesel was further used to developed "neocognitron" [3], artificial neural network which led the foundation for various type of neural networks including Convolutional neural network.

The foundation of CNN was kept and CNN shown great results when it comes to image classification. But the implementation of CNN in AI has it's problem. CNN in AI uses layers of make the network deeper and also provides us with great results when we choose a good CNN model structure with a good amount of data. But every layer in CNN focuses on specific sub-area or a portion of the image and train itself. This method is great, but there are 2 issues.

- CNN becomes sensitive to placement of the features on an image
- A special relationship of features is not captured properly (due to sub-sampling)

### CNN becomes sensitive to placement of the features on an image

Say we train a CNN model for detecting a person in an image (Say with high number of images – all standing), the model will be able to predict a validation image with a person in it with high accuracy, but the model may fail to identify the same validation image when passed flipped 90 degrees or passed as upside down. This is because, the model is only trained on images with people all standing, and no flipped images. The solution to this issue could be data augmentation. But still, with data augmentation, we are simply training the model with different versions of data, the CNN model is still sensitive to the placement of features in the images.

### The spatial relationship of features is not captured properly

Now let's discuss a yet anther issue with CNN, for example, we have a trained model to identify a human face and used data-augmentation as well. Now the CNN is trained to identify features with eyes, nose, lips, facial structure, ears, hairs, etc. Now CNN doesn't capture spatial relationships due to sub-sampling (pooling), and we have trained the model with fabricated data using image augmentation, now the network is trained to identify specific features.

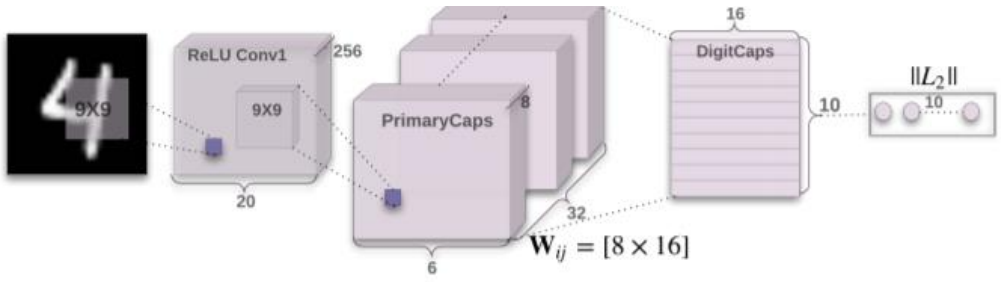
Now say we pass an image to this CNN, where we have eyes, nose, lips, ears, hairs in the image, but not on the face, but somewhere else. There is a high probability, that our CNN model will identify the image as a face.

Now the whole idea behind CNN is to depict the work of brain visual cortex working, but due to lack of Spatial information, a CNN model can be easily fooled. This same concept of fooling the CNN is very well explained in a recent paper "One-pixel attack for fooling deep neural networks" published [4] by Jiawei Su, Danilo Vasconcellos Vargas and Kouichi Sakurai.

## Capsule Network:

Geoffrey Hinton, a well-known name suggested an alternate network architecture "Capsule Network"[5][6] to tackle the issues with pooling and suggested a way to also capture the directions of the features of the image. Capsule network (CapsNet) is a different type to a network where we use a different type of activation method "Squash" and instead of making the network deep, we use a group of neurons in a single layer. This group of neutrons is known as the capsules. Now in-order to train the model, we have to have a communication mechanism in a capsule. The

communication is done by “dynamic routing”. In-fact, “dynamic routing” and “Squash” are the hard and soul of the CapsNet




1. Architecture of CapsNet [6]

## Activation function in CapsNet

Unlike the activation function in CNN, CapsNet doesn’t use methods like ReLU, sigmoid, etc. It uses the squash method instead. The squash method returns Vj value. The value is been calculated using the Sj. Now Cij are the values from the capsules. There is no need for bias here. After discussing the variables, let’s Discuss the important variable U the variable value contains the value and the direction values as well. When it comes to capsule networks the calculations are a vector-based not the linear type when compared to traditional neural networks. The squash activation function formula is given below along with a couple of more formulas that are used in the activation function of the capsule network.

$$\mathbf{v}_j = \frac{||\mathbf{s}_j||^2}{1 + ||\mathbf{s}_j||^2} \frac{\mathbf{s}_j}{||\mathbf{s}_j||}$$

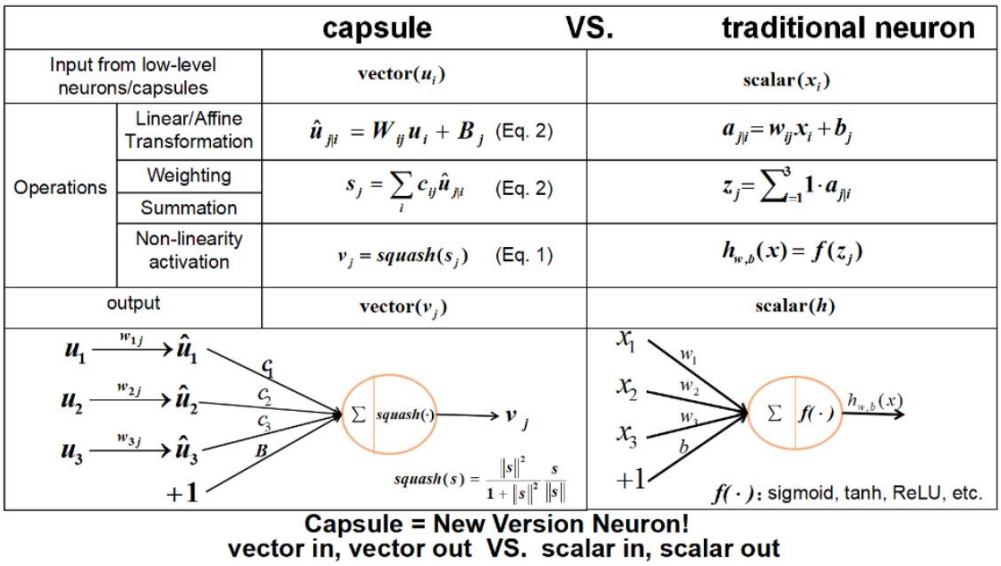
2. Squash formula [6]

$$\mathbf{s}_j = \sum_i c_{ij} \hat{\mathbf{u}}_{j|i} \,, \qquad \hat{\mathbf{u}}_{j|i} = \mathbf{W}_{ij} \mathbf{u}_i$$

3. Summation formula [6]

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}$$

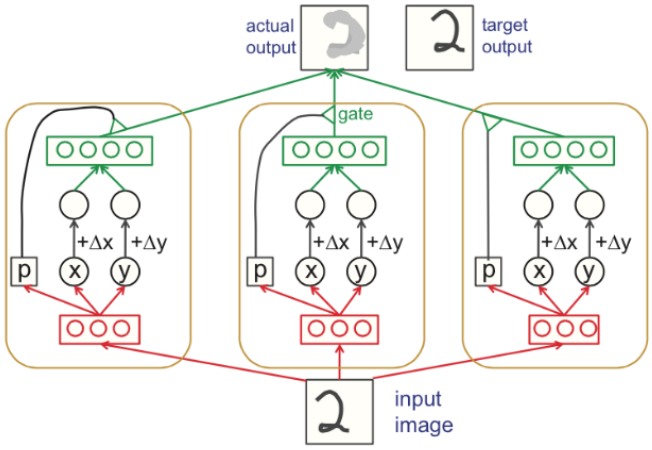
4. Capsule value formula [6]




5. Capsule Vs Traditional neuron [7]

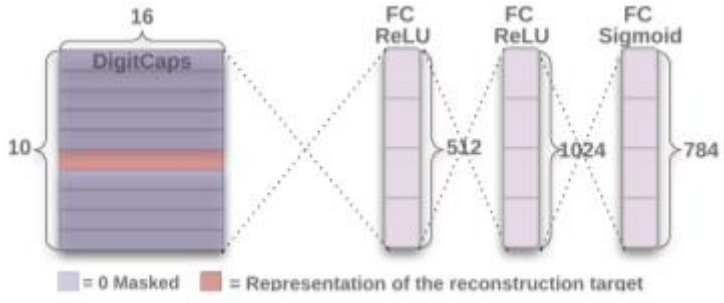
## Dynamic routing

Dynamic routing output from a capsule vector to higher-level capsules j. Shouting decisions are made by changing the scaler weights.




6. Dynamic routing

Now once we get a cab soon it is then decoded using a usual CNN mechanism and we use the activation methods like ReLU Sigma it and then we get output out of it.




7. Decoders [6]

1. https://en.wikipedia.org/wiki/David_H._Hubel
2. https://en.wikipedia.org/wiki/Torsten_Wiesel
3. K. Fukushima, S. Miyake and T. Ito, "Neocognitron: A neural network model for a mechanism of visual pattern recognition," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 5, pp. 826-834, Sept.-Oct. 1983, doi: 10.1109/TSMC.1983.6313076.
4. https://arxiv.org/abs/1710.08864
5. https://link.springer.com/chapter/10.1007/978-3-642-21735-7_6
6. https://arxiv.org/abs/1710.09829
7. https://github.com/naturomics/CapsNet-Tensorflow/
8. http://helper.ipam.ucla.edu/publications/gss2012/gss2012_10754.pdf