ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 11, November 2017

# Capsule Networks for Vision Intelligence Systems

[1] Anuroop Shannu , [2] Pragnya Kondrakunta
Keshav Memorial Institute of Technology

***Abstract:* — The Convolutional Neural Network is a powerful field of study that has been the go-to solution to analyze and classify visual imagery in the recent phase. However, the current state-of-the-art CNN's are facing a reliability crisis when the images are rotated or tilted. There is a drastic fall in their performance when an image isn't close to the ones in the training data-set. A change in orientation and spatial positioning of features, also, has a negative impact on its working. In this paper, we seek answers to these problems by using Capsule Networks, a recent advancement in deep learning. Instead of adding layers, the Capsule Networks work on nesting the layers of a neural network. These nested layers are called capsules. This increases the internal structure of each neuron rather than the depth of the network. Such a structure helps to improve the learning process of machines and renders them close to human intuition. In our work, we explain Capsule networks for Image classification systems and throw light on their robustness.**

***Index Terms* — Capsule nets, Capsules, CNN, Deep Learning, Machine Learning, Neural networks.**

## I. DEEP LEARNING

Machine learning is a technique which allows the computer to learn without explicit task-specific programming. Deep learning is a sub-category of machine learning which tries to mimic the thinking abilities of humans. The deep neural networks are an imitation of the actions performed by neurons in the brain. They learn to recognize the existing patterns in a presented database. Neural nets are extremely flexible in different environments and can adapt to constantly changing information. The neural networks build models that contemplate the structure of the data in nominal time. The program maps out a set of virtual neurons and then assigns arbitrary numerical values called "weights," for the connections between them. These weights determine how each simulated neuron responds which are compared to the actual response. This feedback minimizes the cost (error) and adjusts the weights to achieve high accuracy on the testing data. Your data is constantly being updated, which means your learning models will be too. One of the most significant uses for deep learning is to understand patterns in a way that humans can't – and then trigger actions.

## II. CONVOLUTIONAL NEURAL NETWORKS

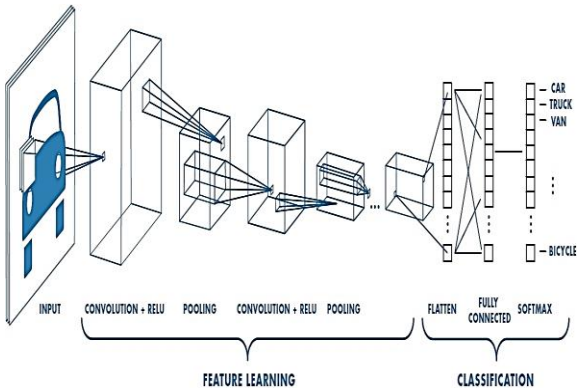
The Convolutional Neural Networks are extensively used for image classification. They were inspired by the architecture of the animal visual cortex. When we look at an image, we look for distinct features that allow us to determine what it is. The identification of an image is based on the combination of various features put together. This entire process occurs within a fraction of time. Similar to the cortical neurons in our brain, the CNN's hidden layer consists of a set of fundamental tiers : the convolutional layer, the pooling layer, the ReLU layer, fully connected layer and the loss layer. Basically, convolution is the matrix multiplication and its summation. When an input image is passed through the CNN, a set of filters scan the image and break it down into overlapping tiles. The convolution layers perform convolution on the matrices of overlapping tiles. This leads us to a feature map in the network. This feature maps are passed through activation functions like, the ReLU. The pooling layer help us to reduce the training time. They include values from the feature maps which are combined in the next layer. For example, max pooling is a type of pooling that returns the highest value in the feature map of the previous layer. Another type is the average pooling. This provides us with the average of the values in the feature map. This creates 'summaries' of sub-regions of the image and includes the distinct qualities traced.We need to consolidate these unique features traced and characterize the entire image. The fully connected layer integrates all the elementary details brought in by each neuron (convolutional layer) to predict the final classification of the image.

## III. PROBLEM STATEMENT

The Convolutional Neural Networks are extremely skillful and have an exceptional performance for classification of data close to their training set. However, small changes in orientation and rotation are adversely affecting their dexterity. Though the Convolutional Neural Networks were built keeping in mind the working of the visual cortex in human brain, they fail to take notice of a few key points. Spatial translation and invariance is one such detail. Invariance ability with which you can

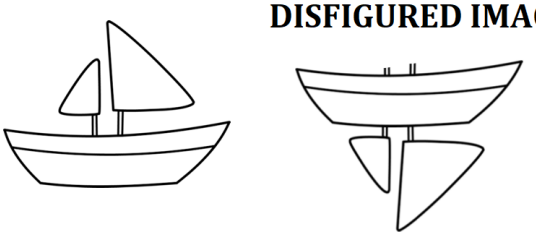
ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 11, November 2017



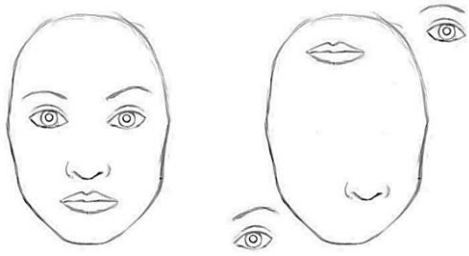

recognize an object even when its appearance *varies* in some way. It can include translation, rotation , size and illumination invariance. Pooling layer in the ConvNet was introduced such that minute changes in view point or size should not change the identification process. We use pooling after each layer to make computation in reasonable time frames. This invariance also leads to triggering false positive for images which have the components of a classification but not in the correct order. The pooling layer also adds this sort of invariance. Also, it loses out the positional data and this

**DISFIGURED IMAGE**



method we use is very crude.

Secondly, coincidences in high-dimensional spaces are rare,



*To a CNN, both pictures are similar, since they both contain similar elements. Source.*

and when they occur, it is because the data "agrees". Objects are composed by parts with specific arrangements. For example, let us consider a face. Breaking it down to elementary details : two eyes lay above a mouth. But if object parts are not in the correct position, it renders false recognition. One eye with a mouth on side and another eye above is not a face! The power of deep neural network is in how we connect layers together. We use fully-connected matrices to connect all features in one layer to all features in another layer, but this, beside for computational efficiency, makes no sense! If the "eye" and "mouth" neurons in a layer l connect to a "face" neuron in layer l+1, that makes sense. But if we connect "eyes", "wheels", "hands", etc to the neuron "face", this will lead to more confusion of information, and poorer performance. For this reason we seek an algorithm that can guide the connection between layers in a more meaningful way. In that situation, optimization algorithms will ,more frequently, find better and faster solutions. Also, there is a strong urge to make each neuron more capable to enhance the performance of the entire neural network. Keeping this in mind, Geoffrey Hinton, the father of deep-learning, released a paper on Capsule Networks. Instead of adding more layers to a neural network we will nest the neural layers.This is in conformation with the need to improve skills of each neuron.

## IV. CAPSULE NETWORKS

The CNN follows hierarchical detection of edges, shapes and then objects. Although CNNs have been built with human vision in mind, they do not notice some of the key perspectives which otherwise a human brain would consider before classifying an image. For example, the spatial information of the features present in an image is always lost.

The pseudo-code for a CNN can be depicted as below:

```
if (2 eyes && 1 nose && 1 mouth) {
 It's a face!
}
```

While the pseudo-code for Capsule Networks looks more like:

```
if (2 adjacent eyes && nose under eyes && mouth under nose) {
 It's a face!
}
```

It beat out the state-of-the-art CNN, reducing the number of errors by 45%. Now let's describe the idea of capsules. Like the basic neuron, they also represent the symbolic mathematizing of a cognitive idea using a somewhat naive assumption: higher up parts of our brains do more interpreting, understanding, and calculating of higher level features, with specific parts of the brain getting specific in what areas or topics they deal with. We don't take in data to all dimensions equally across the brain but instead we "feed in" lower level features for processing by higher level parts of the brain to take the cognitive load off the higher level

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 11, November 2017

processing. If the lower level feature is not relevant to some higher level part of the brain, it shouldn't be sent there. At the very least, its signal should be diminished somewhat. These capsules were conceived to handle the problem of identifying pose. This is when a model might be trained to identify a dog, but becomes reliant on the orientation of that dog within view. If you turn this dog around and try to take a picture from a different angle, the model might have trouble recognizing it. Capsules attempt to solve this by having a higher level part of the "symbolic mathematical brain", i.e. the network, handle the identification and post of complicated features while lower level ones handle "sub"-features. A higher level capsule might identify a face based on lower level capsules identifying a mouth and nose with agreeing orientation. Lower level capsules make "weak bets" on what an object could be by identifying simpler sub-parts of that object. A higher level capsule then takes these lower-level bets and tries to see if they agree. If enough of them agree, then its likely beyond coincidence that this object is Y. That's the essence of how these capsule networks work.

## V. APPLICATIONS

Living in the age of self driving cars, state of the art vision intelligence systems play a key role in today's technological society. As Elon Musk points out, it is important that machines and computers gain trust of humans in tasks which can alter a person's life in both good and bad ways and driving is not an exception.
In concern to the blind, this project urges them to move forward by performing routine activities -comfortably and at a much faster pace.
It helps monitor automobile drivers' alertness/drowsiness. We can design techniques for forensic identification too.

## VI. CONCLUSION

The capsule networks lay foundation to a completely new era of neural networks. They prove the traditional back-propagation algorithm wrong and rely on dynamic routing which is a more intuitive approach. Capsule networks will create a breakthrough in the field of deep learning and artificial intelligence. It is exciting to realize the various applications of this state of the art algorithm.

## REFERENCES

1. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images"- Anh Nguyen, Jason Yosinski, Jeff Clune

2. " Dynamic Routing Between Capsules" - Sara Sabour Nicholas Frosst,Geoffrey E. Hinton

3. "Capsule Networks Are Shaking up AI—Here's How to Use Them" - Nick Bourdakos

4. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." - Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo et al.

5. "Multiple object recognition with visual attention" - Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu.

6. "Batch-normalized maxout network in network." - a-Ren Chang and Yong-Sheng Chen.

7. "High performance neural networks for visual object classification." - Dan C Cire¸san, Ueli Meier, Jonathan Masci, Luca M Gambardella, and Jürgen Schmidhuber.

8. "Multi-digit number

9. recognition from street view imagery using deep convolutional neural networks." -Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet.

10. " Tagger: Deep unsupervised perceptual grouping. In Advances in Neural Information Processing Systems" - Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, and Jürgen Schmidhuber.

11. " Shape representation in parallel systems." -Geoffrey E Hinton. In International Joint Conference on Artificial Intelligence Vol 2, 1981a.

12. " A parallel computation that assigns canonical object-based frames of reference."-Geoffrey E Hinton.

13. " Learning to parse images."-Geoffrey E Hinton, Zoubin Ghahramani, and Yee Whye Teh.

14. "Transforming auto-encoders" - Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang.

15. "Spatial transformer networks."- Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu.

16. "A method for stochastic optimization."- Diederik Kingma and Jimmy Ba.