

# 3-Level Residual Capsule Network for Complex Datasets

Sree Bala Shruthi Bhamidi

*Electrical and Computer Engineering  
Purdue School of Engineering, IUPUI  
Indianapolis, USA  
sbhamidi@iupui.edu*

Mohamed El-Sharkawy

*Electrical and Computer Engineering  
Purdue School of Engineering, IUPUI  
Indianapolis, USA  
melshark@iupui.edu*

**Abstract**—The Convolutional Neural Network (CNN) have shown a substantial improvement in the field of Machine Learning. But they do come with their own set of drawbacks. Capsule Networks have addressed the limitations of CNNs and have shown a great improvement by calculating the pose and transformation of the image. Deeper networks are more powerful than shallow networks but at the same time, more difficult to train. Residual Networks ease the training and have shown evidence that they can give good accuracy with considerable depth. Residual Capsule Network [15] has put the Residual Network and Capsule Network together. Though it did well on simple dataset such as MNIST, the architecture can be improved to do better on complex datasets like CIFAR-10. This brings us to the idea of 3-Level Residual Capsule which not only decreases the number of parameters when compared to the seven-ensemble model, but also performs better on complex datasets when compared to Residual Capsule Network.

**Index Terms**—CNN, Capsule Network, Residual Network, Dynamic Routing, Residual Capsule Network

## I. INTRODUCTION

Convolutional Neural Networks (CNNs) have become an integral part of machine learning. Even with a history of more than 20 years, CNNs have shown that there is always a room for improvement. Since concept of Deep Convolutional Network [1] has been rolled out, we can see a significant improvement in challenging tasks like Image Classification, Image Recognition. Though the deeper networks did improve the performance of the neural network models, stacking of layers brought in a new problem of vanishing gradients. This problem has been alleviated with the introduction of a new network – Residual Network (ResNet) [2]. The network adds Skip connections between the layers i.e., the outputs of the previous layers are added to the outputs of the stacked layers in a feedforward manner. Adding these connections not only decrease the number of parameters, but also helps in concatenating the feature maps for a better gradient flow across deeper networks [14].

Convolutional Neural Networks are our go-to algorithm when it comes to object recognition or object detection. But there are many things that are very unlike the brain that are making the CNNs work not as well as they could. One thing that is missing in Neural Networks is the notion of entity.

Sabour et al. [3] pointed out the drawbacks of the traditional CNNs. Convolutional Neural Networks use multi layers of feature detectors which are replicated across space. These feature extractors with subsampling pooling layers attend only to the active features. In other words, pooling gives only a small amount of translational invariance at each level that is, the exact location of the most active feature extractor is ignored. Pooling also reduces the number of inputs to the next layer of the feature extractor. The downside to pooling is that they fail to use an underlying linear manifold which would deal easily with the effects of viewpoint. We do not want the neural activities to be invariant of the viewpoint instead we want the knowledge of the viewpoint which can be applied to a new viewpoint. Convolutional Neural Networks try to make the neural activities invariant to small changes in viewpoint by combining the activities of the pool. But it is better to aim at equivariance where the changes in viewpoint correspond to neural activities.

Sabor et al. [3] tossed the idea of nesting the layers instead of stacking them. The nested layer is called the capsule, which is a group of neurons. This model is robust to transformations in terms of rotations. The capsule network has two key features: layer based squashing and dynamic routing. It replaces the scalar-output feature detectors of CNNs with vector-output capsules and max-pooling with routing-by-agreement to achieve the state-of-the-art accuracy on the MNIST dataset. The authors have used just a single layer of convolution and capsules. Increasing the complexity and adding depth to the network is a possible improvement. On that basis, ResNets accelerate the speed of training of the deep networks. They reduce the vanishing gradient effect by increasing the depth of network instead of the width, resulting in lesser parameters and obtaining higher accuracy in network performance.

Following the same instinct, we have proposed the use of Residual Network to increase the depth of Capsule Network. The ResNet [2] will be the input to the dynamic routing algorithm [3]. The proposed architecture not just shows a reduction in the model size but also reduces the inference time of the model. The method has been evaluated on MNIST and CIFAR-10 datasets and the results are compared to the Capsule Network keeping the parameters such as learning

---

This is the author's manuscript of the article published in final edited form as:

rate, learning decay, number of capsule parameters same as the model proposed by Sabor et. al [3].

## II. BACKGROUND

The study of neural networks and architectures has been dominant part of machine learning right from the start. But the recent fame of the neural networks has added fuel to the research in this domain. The different CNN architectures proposed have added more layers significantly increasing the parameters and the computational time. The lower layers detected basic features while the higher layers detected more complex features in addition to the lower layers. Though these structures of have boosted the performance, there was a vast increase in the number of parameters.

Highway Networks [4] were the first of architectures to successfully train deep networks with large number of layers. Highway Networks with hundreds of layers can be optimized and trained effortlessly using gating units. By introducing skip connections which are used as bypassing paths, ResNets [2] have achieved a striking performance on Image classification and Image recognition tasks.

The Capsule Networks [3] are a great leap into the neural networks. Capsules are a group of neurons that represent various properties of entities in an image. These properties may include parameters like colour, positions, size, orientation, hue, etc. Capsules output a vector, which implies that the lower level capsules selectively agree on the parent capsule. The connection strength between the lower level capsule and parent capsule is increased when the prediction for parent capsule matches with the actual output of the parent capsule. During training, all activity vectors are masked but the correct activity vector which is then used to reconstruct the input image. The output is used to compute the loss. This way, the network is stimulated to learn more representations of the image.

## III. MATERIALS AND METHODS

### A. Residual Capsule Network

The Residual Capsule Network [15] is one such architecture which is a combination of both Residual networks and Capsule Network to make the Capsule Network deeper yet efficient. The first Convolutional layer of baseline Capsule Network learns very basic features Capsule Network is modified to form a deeper architecture where an 8-layered deeper convolutional subnetwork based on skip connections is created. Every layer is concatenated to the next layer in the feed-forward manner adding up to make the final convolution layer. This leads to better gradient flow as compared to the directly stacked convolution layers, thus decreasing the number of parameters.

TABLE I  
COMPARISON OF RESIDUAL CAPSULE NETWORK ON CIFAR-10 DATASET

| Model                    | Parameters       | Test Accuracy             |
|--------------------------|------------------|---------------------------|
| Capsule Network          | 7x14.5M = 101.5M | 89.40% (7 model ensemble) |
| Residual Capsule Network | 11.86M           | 84.16%                    |

Table 1 shows that though Residual Capsule Network did reduce the number of parameters, it did not outmatch the performance of seven-ensemble Capsule Network on complex datasets like CIFAR-10.

### B. Proposed 3-Level Residual Capsule Network

A simple layer of Residual Network [2] did not solve the problem for complex dataset like CIFAR-10. This maybe because of the simple primary capsules which may not be enough to compute all the features of the image and the relation between part—whole of each entity. To overcome this, we propose an architecture: 3-Level Residual Capsule Network for complex datasets. This architecture has been inspired by the paper published on YOLOv3 by Joseph Redmon et. al [9]. The primary capsules are created to carry information at 3 different scales of the image. Each level of the 3-Level Residual Capsule Network is similar to Residual Capsule Network.

The most prominent feature of this network is that it makes detections at three different scales. This network is capable of detecting objects of different sizes. As the network goes deeper, the feature map keeps getting smaller making it difficult to detect. In order to detect all the entities of the image, the feature maps need to be detected at different scales. Since YOLOv3 has similar structure to ResNets it is a better way to implement the idea into Residual Capsule Network and improvise it to grasp features at different scales. Inspired by the idea of detecting images at three different scales, we have implemented our proposed Residual Capsule Network to carry out information of various scales of images, hence diversifying the capsules. If we cone one level of the proposed 3-Level Residual Capsule Network, the image activations will be same as the earlier Residual Capsule Network and baseline Capsule Network.

The ensemble baseline Capsule Network model by Sabor et. al [3] gives an accuracy of 89.4%. Though the ensemble model is achieving a high accuracy, it leads to a large increase in the number of parameters. Our architecture aims at reducing the number of parameters. Figure 1 shows the queuing of the multiple Residual Capsule Network into layers forming 3-Level Residual Capsule Network. Each of the single layered Residual CapsNet comprises of 12 capsules. The stride at each level is defined to downsample the input image before feeding into the next level of the network. In addition to the three DigitCaps layers, we created one more DigitCaps output layer by routing the concatenation of the three Primary capsule layers [8]. Adding the extra layer would help the model learn better from the merged features assembled from various levels. While testing, the four DigitCaps layers are concatenated to avoid any imbalance learning by the model. DigitCaps generated from various Primary capsule layers played a major role in affecting the reconstruction, which is a combined effect of the different layers of capsules.

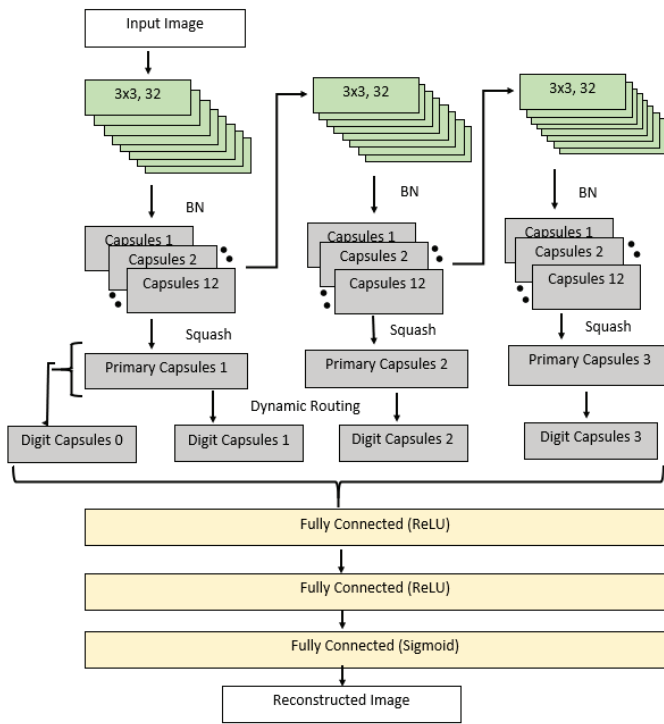


Fig. 1. 3-Level Residual Capsule Network

#### IV. EVALUATION

Although the paper results do not show much of performance improvement in accuracy over state-of-the-art image classification, the proposed architecture is well motivated to decrease the number of parameters, i.e., in decreasing the complexity of the model and of interest and we believe it contributes to advancing state-of-the-art. We have evaluated our proposed models on the basic complex dataset: CIFAR-10 and compared the results with the Baseline Capsule Network by Sabor et. al [3], DCNET and DCNET++ by Phaye [8]. We ran all our evaluations on Aorus GeForce RTX 2080Ti GPU. We have run our models for 120 epochs each. Our implementation is in Keras and we use Adam optimizer with parameters set to 0.001 as initial learning rate and a decay rate of 0.9. We have used publicly available code [10] and made changes to suit our requirements for the proposed network.

CIFAR-10 [12] dataset consists of 60,000 images of 32x32 size each in 10 classes, with 6000 images per class. The images are divided into 50,000 training images and 10,000 test images. We compare our proposed 3-Level Residual Capsule Network with seven ensemble Capsule Network [3] and DC++Net [8].

When the number of parameters is compared with the seven-ensemble Capsule Network, the proposed network has fewer parameters by 90.7M. And when compared with the DC++Net, the proposed network has reduced 2.6M parameters as shown in Table 2.

TABLE II  
PERFORMANCE OF VARIOUS CAPSULE NETWORK MODELS ON CIFAR-10 DATASET

| Model                                     | Parameters       | Test Accuracy             |
|---|------------------|---------------------------|
| Baseline Capsule Network                  | 7x14.5M = 101.5M | 89.40% (7 model ensemble) |
| DCNet                                     | 11.88M           | 82.63%                    |
| Residual Capsule Network                  | 11.86M           | 84.16%                    |
| DC++Net                                   | 13.4M            | 89.71%                    |
| Proposed 3-Level Residual Capsule Network | 10.8M            | 86.42%                    |

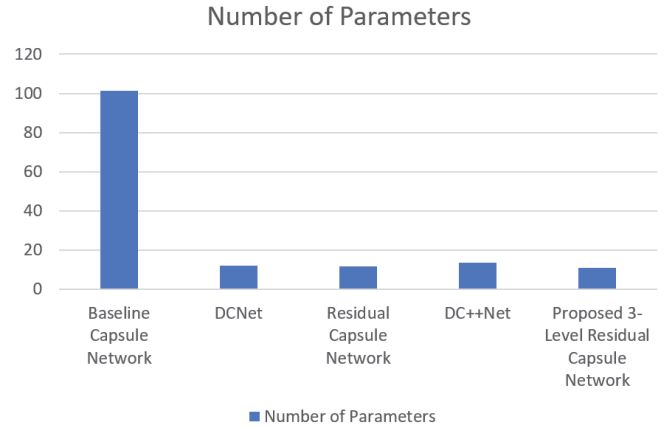


Fig. 2. Comparison of Number of Parameters

#### DISCUSSION

The proposed architecture is a combination of the best features of Residual Capsule Network and a motivation from YOLOv3 model [9] to make the Residual Capsule Network deeper yet efficient. The Residual Capsule Network [15] is replicated across 3 stages to form a 3-Level Residual Capsule Network to grasp features at different levels. The proposed 3-Level Residual Capsule Network model is capable of reducing the number of parameters by 89.36% by compromising on the accuracy by 2.98% when compared to the seven-ensemble Capsule Network, also, there has been a significant decrease in the number of parameters by 9.09%, 8.93% and 19.4% when compared to the DCNet, Residual Capsule Network and DC++Net respectively. The proposed model has especially overcome the concerned raised by Residual Capsule Network by increasing the accuracy by 3.79% when trained on complex dataset as CIFAR-10.

#### REFERENCES

- [1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." 2012.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE Conference on Computer Vision and Patter Recognition. arXiv preprint arXiv: 1512.03385, 2015. [online] [https://arxiv.org/pdf/1512.03385.pdf] [Accessed on July 20, 2019].

- [3] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. “*Dynamic routing between capsules.*” In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.
- [4] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. “*Highway networks.*” CoRR, abs/1505.00387, 2015.
- [5] Geoffrey E Hinton, Sara Sabour, Nicholas Frosst. “*Matrix Capsules with EM Routing.*” 2018. [online] [<https://openreview.net/pdf?id=HJWLFGRb>] [Accessed on July 20, 2019].
- [6] Edgar Xi, Selina Bing, Yang Jin. “*Capsule Network Performance on Complex Data.*” arXiv preprint arXiv:1712.03480, 2017. [online] [<https://arxiv.org/pdf/1712.03480.pdf>] [Accessed on July 20, 2019].
- [7] Geoffrey E Hinton, Alex Krizhevsky, Sida D Wang. “*Transforming Auto-Encoders.*” In *International Conference on Artificial Neural Networks*. Springer, 2011.
- [8] Sai Samarth R Phaye, Apoorva Sikka, Abhinav Dhall, Deepti Bathula. “*DCNET and DCNET++: Making the Capsules Learn Better.*” arXiv:1805.04001, 2018. [online] [<https://arxiv.org/pdf/1805.04001.pdf>] [Accessed on July 20, 2019].
- [9] Joseph Redmon and Ali Farhadi. “*YOLO v3: An incremental improvement.*” arXiv preprint: 1804.02767, 2018. [online] [<https://pjreddie.com/media/files/papers/YOLOv3.pdf>].
- [10] SSRP, Multi-level-DCNet. [online] [<https://github.com/ssrp/Multi-level-DCNet>] [Accessed on July 20, 2019].
- [11] Yann LeCun, Corinna Cortes, Christopher JC Burges. “*The MNIST Database of Handwritten Digits.*” 1998. [online] [<http://yann.lecun.com/exdb/mnist/>] [Accessed on July 20, 2019].
- [12] Tom Hope, Yehezkel S. Reshe, Itay Lieder (2017-08-09). “*Learning TensorFlow: A Guide to Building Deep Learning Systems.*” “O’Reilly Media, Inc. pp. 64. ISBN 9781491978481.
- [13] Jonathan Hui. “*Understanding Dynamic Routing between Capsules (Capsule Networks).*” [online] [<https://jhui.github.io/2017/11/03/Dynamic-Routing-Between-Capsules/>] [Accessed on July 20, 2019].
- [14] Sree Bala Shruthi Bhamidi, “*Residual Capsule Network*”. [online] [<https://scholarworks.iupui.edu/handle/1805/19902>] [Accessed on July 20, 2019].
- [15] Sree Bala Shurthi Bhamidi and Mohamed El-Sharkawy, “*Residual Capsule Network*”, IEEE UEMCON 2019, Columbia University, New York, October 10-12, 2019.