Name: Mah Yuen Yee
Student ID: 30213673
Final Project Title: Investigating the Performance of a Capsule Network in Digit Classification Task

1. A brief background of the problem

Convolution neural network (CNNs), a class of artificial neural networks that is mostly applied for visual imagery analysis, has been widely used for various applications such as image and video recognition, image classification, image segmentation, medical image analysis, etc. Featured with 3D volumes of neurons, local connectivity, shared weights, and pooling, CNNs have made numerous breakthroughs in processing image, video, speech, and text in recent years. Despite its success, CNN contains several fundamental drawbacks.

First, CNNs are designed to cope with translations, but they cannot handle the effects of changing viewpoints such as rotations and scaling. As CNNs have no built-in understanding of 3D space like the human brain does, the current common method of gathering images displaying objects in various positions to feed the model is not efficient. Second, as the pooling feature in CNNs helps the feature generalization and recognition of feature independent of its location in the image, the pooling layers inevitably lead to a loss of information and ignore the relative spatial relationships between the simple and complex features. For example, a CNN will classify an image as a face if, say, a mouth, two eyes, and a nose are present, whether these features are placed at the right location. In short, CNNs recognize objects very differently from humans so they can be bizarrely weak at dealing with some visual tasks such as generalizing across viewpoints and adversarial examples that a human can easily do. One way to resolve the fundamental limitations is to bring in any approach that is closer to replicating the human vision.

2. A brief overview of the proposed method

The proposed method is to employ the capsule network proposed by Hinton et al. [1] and Sabour et al. [2]. There are 2 main novelties in the network – capsules and dynamic routing. Capsules are designed to achieve viewpoint invariance in the activities of neurons without losing any information yet encoding relative spatial relationship between features. In short, a capsule is effectively a neuron that recognizes an implicitly defined visual entity over its receptive field and outputs both the probability of the presence of the entity and a set of “instantiation parameters” that may include the precise pose (translation and rotation), lighting and deformation of the visual entity relative to the implicitly defined canonical version of that entity.

On the other hand, to recognize an object, human brain deconstructs the hierarchical representation of anything from visual information received and matches it with learned patterns and relationships in the brain. To mimic this process of so-called inverse graphics, dynamic routing between capsules was proposed. To replace max-pooling and to employ the idea of parse trees, a “prediction vector” that compares the similarity between the capsule output to each capsule input is used for the routing coefficient to be updated correspondingly. The routing coefficient determines the weights of each capsule input onto the capsule output, i.e., effectively determines the parent capsule of any capsule from the layer below.

3. Research conducted

The idea of capsule was first introduced by Hilton et al. [1] in 2011. Nevertheless, it was not until 2017 that there is finally a proposition of an algorithm by Sabour et al. [2] that is able to implement the capsule network. Even though the capsule network can achieve state-of-the-art results in MNIST digit classification [2], it has not managed to perform up-to-par to the CNNs for other tasks that involve more complex data. Xi et al. [3] have managed to improve the capsule network performance slightly by adding a convolution layer and ensemble averaging. Hinton et al. [4] has proposed and implemented Expectation-Maximization (EM) algorithm for routing that greatly reduced the test error and showed more resistance to white box adversarial attack than the baseline CNN. Afshar et al. [5] has proposed the capsule-network-based COVID-CAPS that is capable of handling small datasets for COVID-19 diagnosis from X-ray images. The pre-trained COVID-CAPS performance was superior to the pre-trained CNN-based model and the reference model that is based on deep features and support vector machine in terms of accuracy, specificity, and number of trainable parameters.

4. Proposed dataset

To investigate the performance of a capsule network in digit classification task, the torchvision built-in MNIST dataset will be used. Provided there are sufficient time and resources, the performance of the capsule network will be investigated in other classification tasks such as the image classification task using CIFAR-10 dataset.

5. Reference

[1] G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming Auto-Encoders,” in International conference on artificial neural network, 2011, pp. 44-51, doi: 10.1007/978-3-642-21735-7_6.

[2] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in Advances in neural information processing systems 30, 2017.

[3] E. Xi, S. Bing, and Y. Jin, “Capsule network performance on complex data,” 2017, arXiv:1712.03480.

[4] G. E. Hinton, S. Sabour, and N. Frosst, “Matrix capsules with EM routing,” in International conference on learning representations, 2018.

[5] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images," Pattern Recognition Letters, vol. 138, pp. 638-643, Oct 2020.