Poster ID: 1
A Selective Modular Neural Network Framework

Intisar Chowdhury, Qiangfu Zhao

System Intelligence Lab, The University of Aizu

Objective

State-of-the-art Deep Neural Networks (DNN) now perform with neck-and-neck score or excel human in most of the human perception ability such as, hearing or voice recognition, natural language synthesis, image recognition , policy making and so on. However, these DNNs' practical applicability in real world scenario is limited as we are always equipped with low computational devices in remote areas. To mitigate this issue, we propose a simple yet effective modular neural network framework for multi-class classification. The proposed framework significantly reduces the number of parameters while maintaining the accuracy comparable to more complex DNN such as ResNet family, DenseNet and so on. The framework primarily consists of two major parts, namely a routing module and a set of expert modules. Each of the expert module is a multi-class classifier trained on subsets of data-set. The network leverage the routing module to select only a small set of expert modules for each input datum during the testing phase. The selection of the expert module is carried out based on the routing module's soft-max scores for top-n classes.

Key Idea

The basic idea is, instead of running each and every baseline module in the ensemble, we run only a selected (constant) number of baseline modules. To do so, we propose a framework, where selection of the modules are initiated based on a routing module.

- The routing module is trained to be accurate enough to include the correct answer, say in the top-2 outputs (Here, 2 is called the redundancy rate, which is denoted by R_r. We can use R_r larger than 2 to improve the accuracy, but this maneuver will require us to evaluate more expert networks).
- After we obtain the top-n most likely classes based on the soft-max confidence values of the routing module, the routing module selects 3 or 4 expert modules. These expert modules are then used to make the final decision.
- After evaluating the expert modules, the final prediction is calculated based on the aggregated values of experts and routing module soft-max scores, which we term as the Corrected Activation (CA).

Network Architecture

The proposed SMNN is composed of the following primary parts.

- *Routing module:* The routing module is primarily composed of two parts i) *Feature extractor or convolutional layers* ii) *Fully connected linear classifiers*. The number of convolutional layers of the routing module for all the data-sets were only three. The rest of the parameters such as the kernel size, stride, and max-pooling layer vary for different data-set. The classifier part is a single layer of fully connected neurons.
- *Expert Modules/ Specialists:* Each of expert module has the identical architecture to the routing module. All the expert modules are connected to the routing module through weight vectors.
- *Corrected Activation:* The CA value is the corrected routing module's confidence value, which is calculated based on the expert modules.

$$CA_{i,j} = \frac{exp(z_{m_i c_j})}{\sum_{k \in \{preds\}} exp(z_{m_i c_k})} \tag{1}$$

Where,

$$preds = \begin{cases} \{j-1, j, j+1\} & 1 < j < C \\ \{j-1, j, 1\} & j = C \\ \{C, j, j+1\} & j = 1 \end{cases}$$

$z_{m_i c_j}$ is the CA of expert module i for class j
For final CA,

$$CA_j = \sum_i (CA_{i,j}.q_j)/N_e \tag{2}$$

where, N_e is the number of expert modules evaluated.
Finally, we output the final prediction from the CA.

$$prediction = \arg\max_j (CA_j) \tag{3}$$

Table 1:Accuracy on testing set for Fashion-MNIST data-set.

models	test acc. rate %	train params.	test params.
ALEXNET	91.4±0.27	62.37M	62.37M
RESNET-18	93.50±0.15	11.17M	11.17M
Random forest	84.82	500 DT	500 DT
Distilled CNN	90.92±0.14	0.45M	0.45M
SMNN	91.00±0.30	0.35M	0.16M
Routing module	89.01±0.12	45.35K	45.35K

Table 2:Accuracy on testing set for UCI-HAR data-set.

models	test acc. rate %	train. params.	test. params.
ALEXNET	96.3±0.24	62.37M	62.37M
RESNET-18	93.9±0.36	11.17M	11.17M
Random Forest	85.00±0.01	500 DT	500 DT
Distilled CNN	92.23±0.86	0.45M	0.45M
SMNN	96.00±0.5	2.5M	1.07M
Routing Module	92.23±0.86	0.35M	0.35M

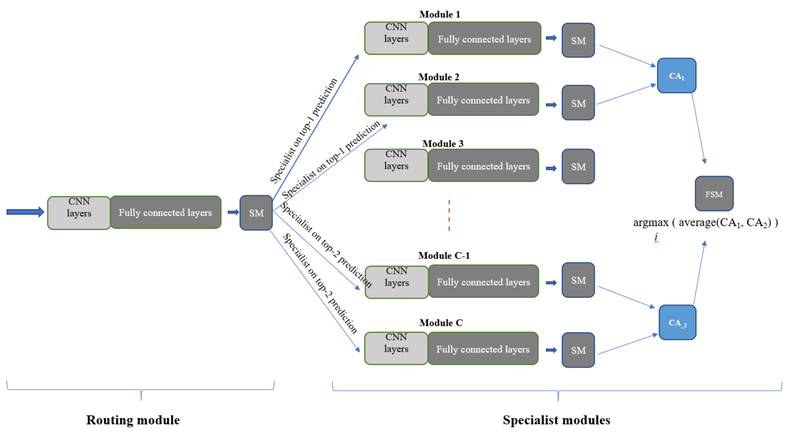

Figure 1:Network Architecture

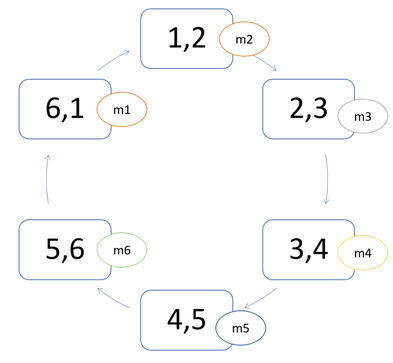

Figure 2:Round Robin based Data partitioning technique

References

[1] C. M. Intisar and Q. Zhao. A selective modular neural network framework. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–6, 2019.