Interpretability in Gated Modular Neural Networks

Yamuna Krishnamurthy, Chris Watkins

Computer Science Department, Royal Holloway University of London

Motivation

- Modular neural networks (MNN), in which different sub-networks carry out different well-defined tasks, offer potential advantages for interpretability and transferability over monolithic deep networks.
- Research in MNN architectures has concentrated on their performance and not on their interpretability.
- We attempt to address this gap in research in MNN architectures, specifically in the simplest gated modular neural network architecture, Mixture of Experts (MoE) [1].

What is interpretability?

- Gating network learns a meaningful modular decomposition of the input space into regions with natural 'rules'.
- Facilitates
  - Attributing errors to the gate or
  - Attributing errors to the modules
  - Model debugging

Is an interpretable task decomposition guaranteed?

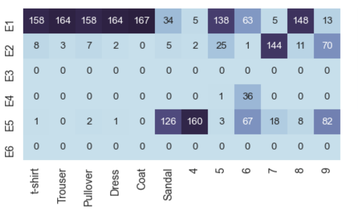

Figure 1 Gate allocation of tasks to modules for the combined FMNIST and MNIST datasets shows that the same module E1 is used for FMNIST and MNIST data . So, an interpretable task decomposition is not always guaranteed even for simple datasets with clear differences in data.

Can a gate learn a good decomposition?

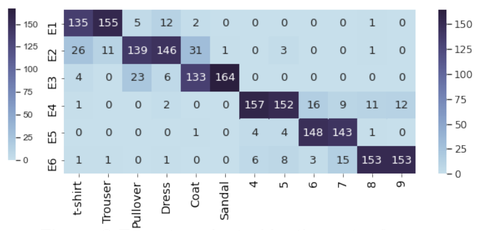

Figure 3 Experts selected by the gate, for test data classification of combined FMNIST and MNIST data, by a model trained with experts pre-trained on custom split of the class: {[t-shirt,Trouser], [Pullover,Dress], [Coat,Sandal], [4,5],[6,7],[8,9]}. The gate does learn a good task decomposition with pre-trained experts.

Our Contribution [2]

Empirical analysis of the state of interpretability in Mixture of Experts (MoE) and our 2 key findings based on this analysis:

1. MoE can be interpretable
2. Current training methods of MoE from random initialisation typically does not produce an intuitively reasonable modular decomposition of the input space, even in very simple cases.

Why does the gate not always learn a good decomposition?

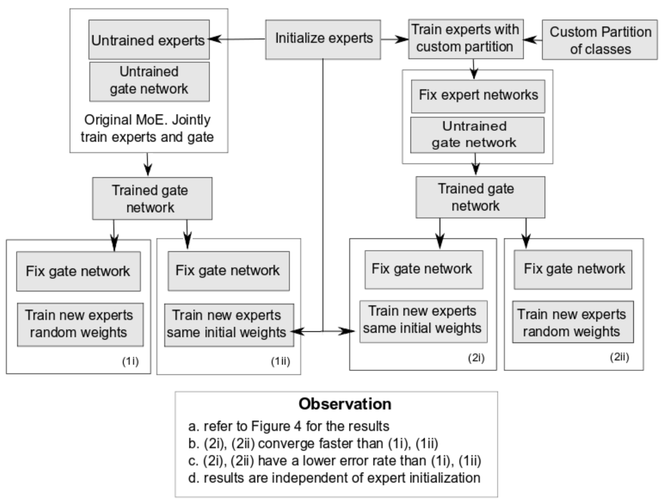

Figure 2 Experiment designed to analyse if the gate can learn good task decompositions (Figure 3) and if there is a training or error advantage for bad decompositions (Figure 4).

Is there a training or error advantage for bad decompositions?

Figure 4 shows that there is neither a training nor an error advantage for a bad decomposition.

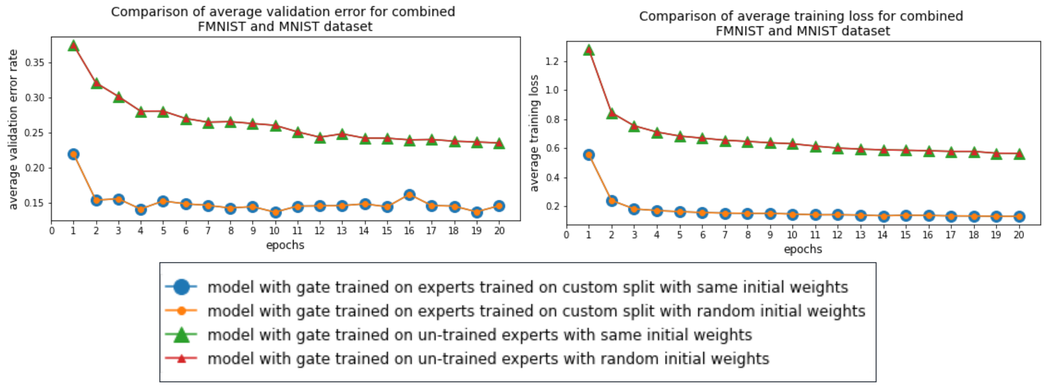

Figure 4 Comparison of training loss and validation error for pre-trained gate, trained with experts pre-trained on custom partition of classes and un-trained experts, and then training new experts with default parameter initialization and experts with the same initial parameter initialization as the experts trained from for MNIST and combined FMNIST and MNIST datasets.

Does the heuristic of jointly training the gate and modules cause bad decompositions?

Figure 5 shows that joint module and gate training could lead to poor allocation of data samples and hence bad decompositions. We are investigating generic training methods of using additional information for better gating decisions and consequently good decompositions..

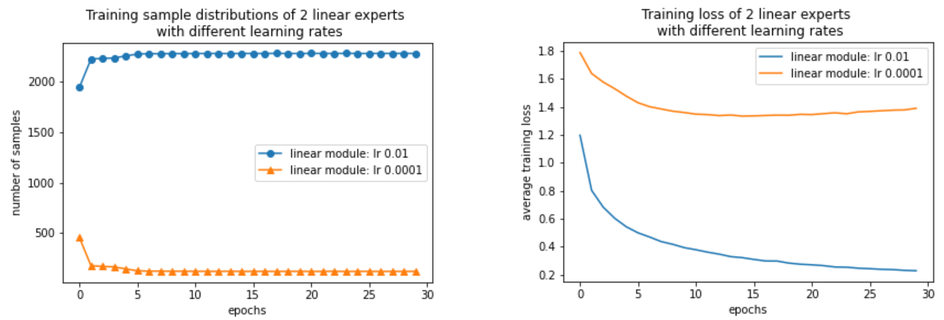

Figure 5 Sample distribution and training loss of GMNN trained with 2 modules with different learning rates. Module with higher learning rate captures most of the data samples and learns faster.

Conclusion

- MoE are indeed inherently interpretable
- Existing architectures and methods of training them do not guarantee an interpretable task decomposition among the modules
- There is no learning or error disadvantage to learning an interpretable task decomposition
- Heuristic of jointly optimising the gate and experts leads to uninterpretable task decompositions

References

[1] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. Adaptive mixture of local expert. *Neural Computation*, 3:78–88, 02 1991.

[2] Krishnamurthy, Yamuna and Watkins, Chris. Interpretability in Gated Modular Neural Networks. In Explainable AI approaches for debugging and diagnosis Workshop at NeurIPS, 2021