Deep Mixture of Experts via Shallow Embedding

1 INTRODUCTION

In this paper, we explore the design of deep mixture of experts models (DeepMoEs) that compose hundreds of mixture of experts layers.

Our contributions can be summarized as: (1) We first propose a novel DeepMoE design which allows

the network to dynamically select and execute part of the network at inference. (2) We theoretically

analyze that the proposed DeepMoE design preserves the expressive power of a standard convolutional network with reduced computational cost. (3) We further introduce two DeepMoE variants that are more accurate and efficient than the prior methods on different benchmarks.

2 RELATED WORK

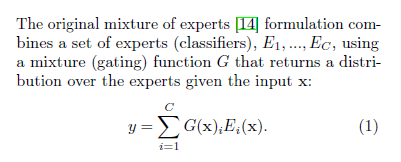
Mixture of experts

Conditional computation

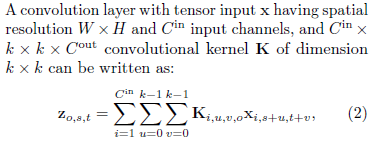
Dynamic channel pruning

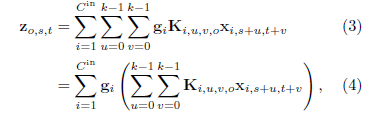
3 DEEP MIXTURE OF EXPERTS

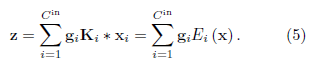
3.1 MIXTURE OF EXPERTS



3.2 DEEPMOE FORMULATION

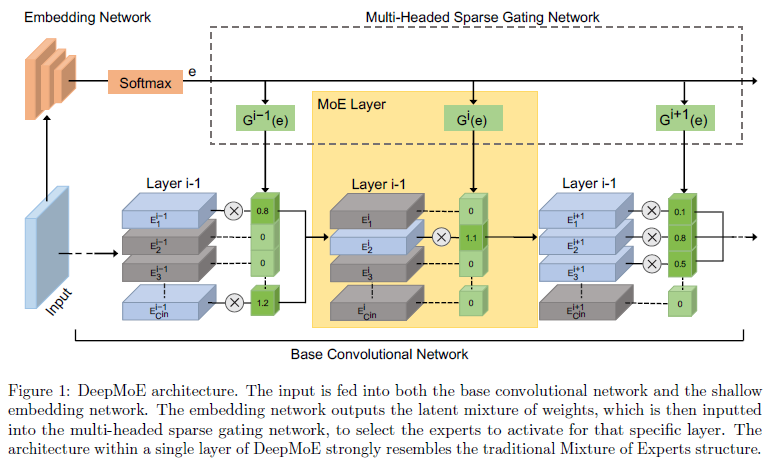






3.3 DEEPMOE ARCHITECTURE

DeepMoE is composed of three components: a base convolutional network, a shallow embedding network, and a multi-headed sparse gating network.

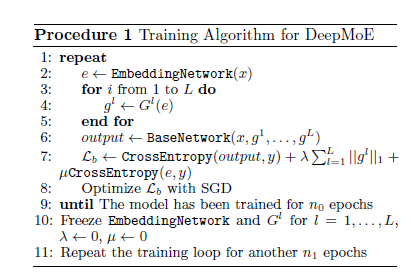


3.4 DEEPMOE TRAINING

As with standard convolutional neural networks, DeepMoE models can be trained end-to-end using gradient based methods.

The overall goals of the DeepMoE are threefold: (1) achieve high prediction accuracy, (2) lower computation costs, and (3) keep the network highly expressive.





4 EXPRESSIVE POWER

The expressive power of deep neural networks is associated with both the width and the depth of the

network. We define the expressive power of a convolutional neural network as the ability to construct labeling to differentiate input values.

5 EXPERIMENTS

5.1 WIDE-DEEPMOE

5.1.1 Improved Accuracy with Reduced Computation

5.1.2 Memory Usage

Another aspect to consider about DeepMoE is its memory footprint (proportional to the number of parameters).

5.2 NARROW-DEEPMOE

In this section we compare DeepMoE to current static and dynamic channel pruning techniques. We show that DeepMoE is able to out-preform both dynamic and static channel pruning techniques in prediction accuracy while maintaining or reducing computational costs.

5.2.1 Narrow-DeepMoE vs Dynamic Channel Pruning

5.2.2 Narrow-DeepMoE vs Static Channel Pruning

5.3 ANALYSIS

5.3.1 Gating Behavior Analysis

5.3.2 Regularization Effect of DeepMoE

5.3.3 DeepMoE vs Single-Layer MoE

5.4 SEMANTIC SEGMENTATION

The hyper-parameter lambda can adjust the trade-offs between computer efficiency and prediction accuracy.

6 CONCLUSION

In this work we introduced our design of deep mixture of experts models, which produces a more accurate and computationally inexpensive model for computer vision applications. Our DeepMoE architecture leverages a shallow embedding network to construct latent mixture weights, which is then used by sparse multi-headed gating networks to select and re-weight individual channels at each layer in the deep convolutional network. This design in conjunction with a novel sparsifying and diversifying loss enabled joint differentiable training, addressing the key limitations of existing mixture of experts approaches in deep learning. We provided theoretical analysis on the expressive power of DeepMoE and proposed two design variants. The extensive experimental evaluation indicated that DeepMoE can reduce computation and surpass accuracy over baseline convolutional networks, as well as improving upon the residual network result on the challenging ImageNet benchmark by a full 1%. Through our analysis we were also able to prove that our embedding and gating network is able to resolve coarse grain class structure in the underlying problem. This work shows promising results when applied to semantic segmentation tasks, and could be incredibly useful for various other problems.