Selective Neural Network Pruning under Domain Shifts: A Multi-Stage Activation-Based Approach

Pratim Chowdhary
Undergraduate Thesis
Dartmouth College

December 28, 2024

Abstract

Domain shifts pose a significant challenge in deep learning, particularly for classification tasks where only a subset of the originally learned classes remains relevant. This research proposes a novel pruning methodology that selectively removes unimportant neurons or filters in a neural network once the task domain transitions. By employing L1, L2, and Random Forest–based techniques on hidden-layer activations and model outputs—using strategically sampled random inputs—the approach identifies neurons that best explain the variance in the reduced class subset. Subsequent hidden layers are similarly pruned by leveraging the L2 norms of their forward activations as "target" variables. This multi-stage procedure aims to streamline the network's capacity to match the domain's new requirements, thereby reducing model size and enhancing inference efficiency with minimal accuracy loss. Preliminary experiments suggest that this method offers a principled framework for adapting neural networks under domain shifts while maintaining classification quality.

Contents

1 Introduction **1**

2 Background **1**

3 Methodology **1**

1 Introduction

2 Background

3 Methodology