1 Introduction

2 Methodology

We follow the basic outline process, defining an objective function with the edge mask parameter (z_{ij}) , which is initially defined with a soft distribution bounded in (0,1), but finally converted to a hard-concrete binary distribution to output the result of the edge pruning process.

2.1 Generating Masks

We first generate masks for the edges, their values initially soft distributed between 0 and 1.

Let u be sampled from a uniform distribution :

$$u \sim (\varepsilon, 1-\varepsilon)$$

where $\varepsilon = 10^{-6}$

The sparsity s is calculated as

$$s = \sigma\left(\frac{1}{\beta} \cdot \frac{\mathbf{u}}{1-\mathbf{u}} + \log \alpha\right)$$

$\log \alpha$ is a learnable parameter, β is the temperature set to 2/3 and $\sigma()$ is the sigmoid function

$$\tilde{s} = s \times (r-l) + l$$

$$z = \min\left(1, \max\left(0, \tilde{s}\right)\right)$$

The first line scales the mask value to an appropriate range just greater than [0,1], to allow for flexibility and precise training process, thus introducing an "excess probability". The second line ensures that the final edge mask value z is between 0 and 1.

2.2 Sparsity Enforcement via the Lagrangian Term

$$L_s = \lambda_1 \cdot (t-s) + \lambda_2 \cdot (t-s)^2$$

s : current sparsity
t: target sparsity
λ_1, λ_2 : coefficents optimized via gradient ascent

2.3 Sets of Masks for Edge and Nodes

The effective edge mask is given by

$$\tilde{z}_{n1n2} = z_{n1n2} \times z_{n1}$$

The effective edge mask is the product of the edge mask and the initial node mask , and the edge is pruned based on the effective edge mask value.

2.4 Sets of Masks for Edge and Nodes

$$L = L_{\text{KL}} + L_{\text{edge},s}$$

This is the objective loss that we try to minimize via SGD (Stochastic Gradient Descent), parameter being the effective edge mask value. The loss function has 2 parts, the loss pertaining to the KL divergence (main task loss) and the loss pertaining to the Sparsity Enforcement Target.

2.5 Threshold and Finalized Edge Pruning

Dropout was tested as it reduces over fitting, however, in this experiment, it introduced noise, so it was disabled. The threshold value for conversion of edge masks from soft to hard distribution was not hard coded, it was found by taking permutations of binary mask values, and their average was recorded. Binary search was used to decide on the optimal average based on the target sparsity and hence, the threshold was set and edge pruning achieved.

Note

Optimized Residual Stream :
The activation recieved at a node from the residual stream is now a function of weighted sum of true activations and corrupted activations at previous nodes

$$y_i = f_i\left(z_{0i}y_0 + (1 - z_{0i})\tilde{y}_0 + \sum_{j=1}^{i-1}\left(z_{ji}y_j + (1 - z_{ji})\tilde{y}_j\right)\right),$$

Objective of Circuit Discovery :
We wish to minimize the KL divergence between the output distributions of the pruned circuit and the full model. c here is the sparsity, which is the fraction of retained edges

$$\arg\min_C E_{(x,\tilde{x})\in T}\left[D\left(p_G(y|x) \,\|\, p_C(y|x,\tilde{x})\right)\right]$$