Structured Pruning of a Model under the Constraint that we can only have enough memory for inference.

Q: How do we prune?

A: SparseGPT, Wanda, Magnitude → Yes
We can use the saliency and do a one shot pruning. But can we do better?

* We can use reasonable amount of forward passes.

* Idea 1: Sample some $\hat{M}_i$ (pruned Models) and test it on downstream tasks.

Q1: How to Sample

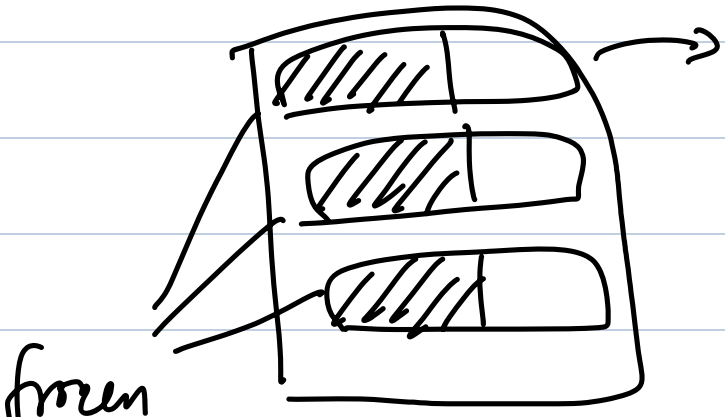
Q2: What do I do with the data $\{\hat{M}_i, s_i\}$ $\hat{M}_i$:

Sampling.

* Space is too large, there are some obvious wrong samples. for ex. if you don't sample any thing in a layer that's a broken model so let us keep some module in each layer

* How do you choose the components to keep? Lets use "saliency" from existing methods

Infact if my target sparsity is $p \in (0,1)$ let us keep $(1-2p)$ fraction of components fixed.



Remaining space of components from which we need to choose half.

frozen (always chosen)

* How to Sample? Uniform?

→ Use saliency to sample with some prior.

② What do I do with $\{\hat{M}_i, s_i\}$

→ Just pick best (one naive way)

Q Can we do better?

possible modules (neurons, heads, etc) are much larger than "Reasonable # of forward passes" which means that we won't be able to get information on most configurations / modules if done naively.

Q. What do we do?

Idea: Create a Model of utility

$$s_i = \hat{u}(\hat{M}_i)$$

↓

approx utility

$$\hat{s}_i = \hat{u}(\{m_1, m_2, \ldots m_n\})$$

What is the Simplest Model?

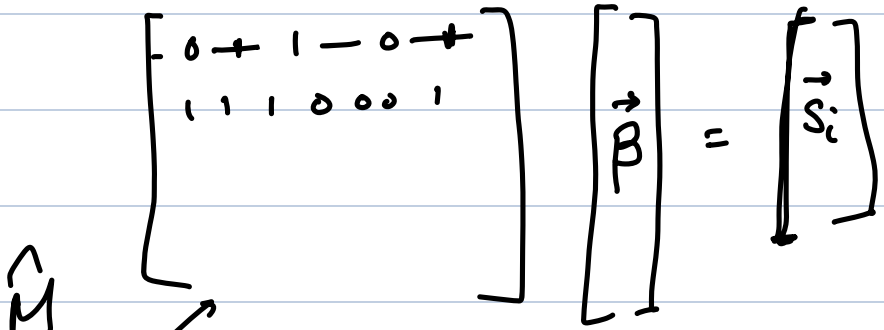
linear model

$$\hat{S}_i = \sum_{m_j \in \hat{M}_i} \beta_j(m_j)$$

N: # models

n: # data samples or # forward passes

Linear Regression problem



$n \times N$ $\quad N \times 1 \quad n \times 1 \quad N >> n$

Sparse Matrix X

We want to estimate β.

Once we get β and if the model Really holds the all we need to do is choose ~ half (i.e p fraction of entire model) modules with highest "β" scores.

$\hat{M}\beta = S \leftarrow$ underspecified

$$\beta^* = \arg\min \| \hat{M}\beta - s \|_2 + \gamma \|\beta\|_2$$

Thus add regularization

Q: How do you sample matrix ($\hat{M}$)

Q → prior based yes. but how many non-zeros per row?

A: ↳ try to maintain fraction "p"

Q: Even with prior it is likely that we miss some module altogether.

⟹ β_i for that module would be a BAD estimate

Trick: If you sample $\hat{M}_i$ also sample its complement. i.e $\hat{M}_i'$

$$\hat{M}_i' = \{ m_i \mid m_i \notin \hat{M}_i \text{ and it comes from set of modules under consideration} \}$$

Exercise: Prove that having the complement of a sample in the data helps reduce variance of β estimation.