Discovering Knowledge-Critical Subnetworks in Pretrained Language Models

Deniz Bayazit, Negar Foroutan, Zeming Chen, Gail Weiss, Antoine Bosselut

Do LMs have specialized subnetworks that encode specific pieces of knowledge?

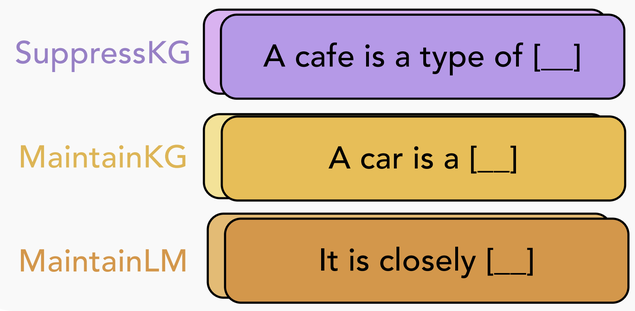
Searching for the Drop in the Ocean

To identify subnetworks essential for the model's understanding of a concept within the context of the full model, we define knowledge-critical subnetworks as those whose removal impairs the model's ability to express specific knowledge.

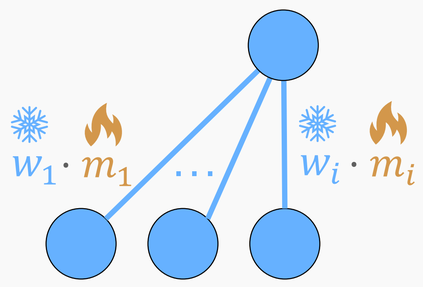
Specifically, a subnetwork is critical to a target knowledge, s.t., when the subnetwork is removed:

1. the target knowledge is removed
2. the model maintains original behavior
3. the subnetwork does not contain parameters irrelevant to modeling the target knowledge

Relational Knowledge & LModeling

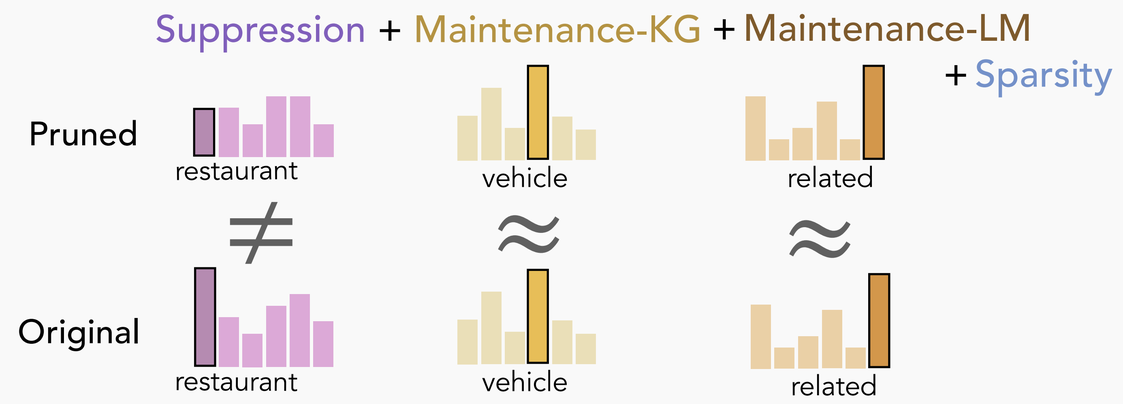


Learning to Mask



The Secret to Finding the Drop

Model Behavior Objectives



Translating to a Loss Objective for Weight-Masking

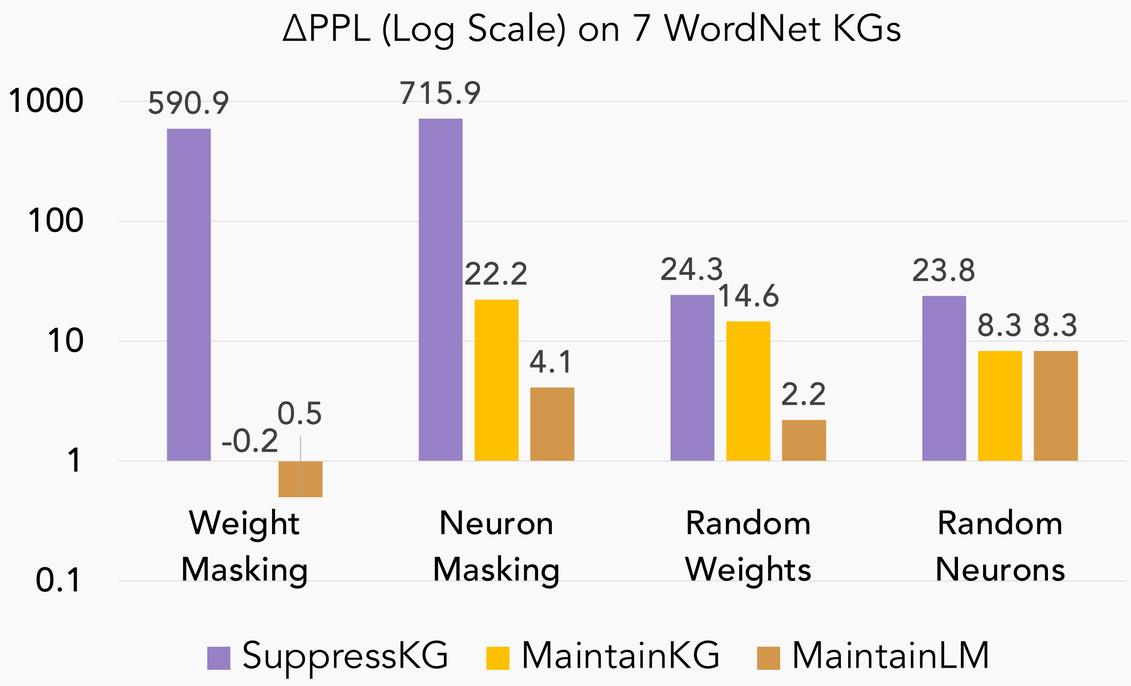
1. **Suppression** KL Div (Uniform Dist. (reference) || Remaining Model Dist. (predicted))
2. **Maintenance** KL Div (Original Model Dist. || Remaining Model Dist.)
3. **Sparsity** Average of softmaxed masking logits

What happens when knowledge-critical subnetworks are removed?

Suppresses the Target, Maintains the Rest

Measure model confidence with perplexity
↑ on Suppression
≅ 0 on Maintenance

$$\Delta\text{PPL} = \text{PPL}(\text{Remaining Model Dist.}) - \text{PPL}(\text{Original Model Dist.})$$

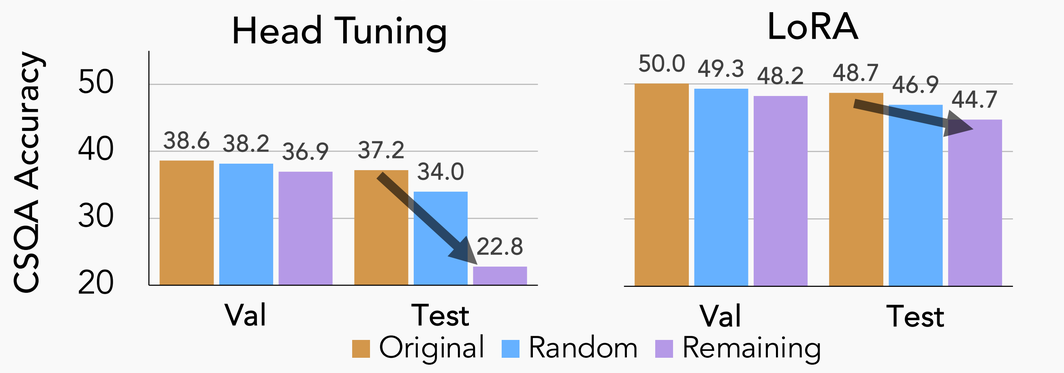


Fine-grained masking is necessary for identifying knowledge-critical subnetworks

Prevents transfer to Downstream Task

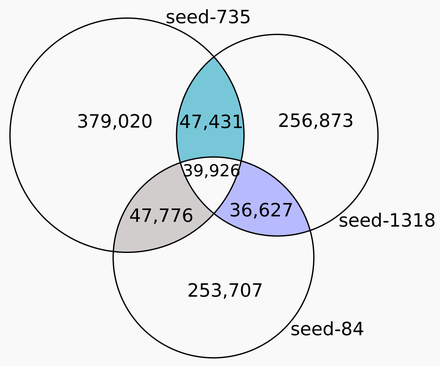
1. Train Knowledge: MaintainKG
   Test Knowledge: SuppressKG
   Find a knowledge-critical subnetwork!
2. Set subnetwork weights to 0
   Finetune on Train Set

New Val. Acc. ≅ OG Val. Acc.
New Test Acc. ≪ OG Test Acc.

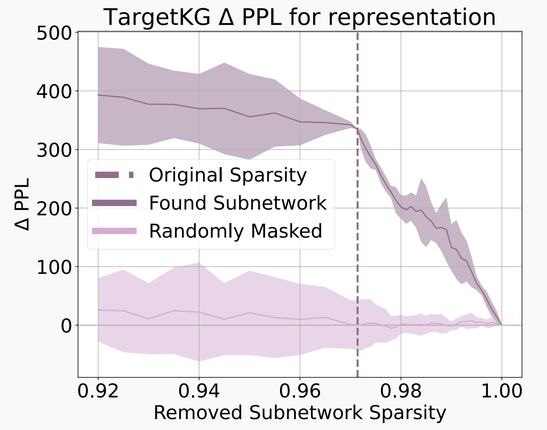


Structural Findings

Subnetwork parameter overlap for different seeds and same KG



Randomly adding or removing parameters from knowledge-critical subnetworks



Subnetworks are distinct across random seeds but composable and robust to perturbations

We uncover ultra-sparse subnetworks (98%+) in LMs crucial for expressing specific knowledge.

Removing them leaves the main network intact but weakens its ability to represent that knowledge.

Paper Link