Task-Specific Skill Localization in Fine-tuned Language Models
Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, Sanjeev Arora
1Princeton University, 2Google NYC (work done at Princeton University), 3Correspondence: ap34@princeton.edu,nsaunshi@google.com

Fine-tuning (FT) ≡ Acquiring new skills

0.01% Sparse Grafted model recovers > 95% accuracy. Improves calibration by 40 – 90% No re-training of Graft involved; different from lottery tickets and parameter-efficient methods.

We perform grafting on Roberta-base and GPT-2 fine-tuned on GLUE tasks and measure the accuracy and calibration error of the grafted models.

Skill localization via Grafting

- For a language model fine-tuned (with SGD) on a downstream task, a very sparse region within the model can localize the acquired skills.
- Region Sparsity → Good few-shot generalization

Grafting learns a binary mask γ using the fine-tuned (θf) and pre-trained (θpt) models, and creates a grafted model θγf(γ).

\[ \gamma = \gamma \odot \theta_{pt} + (1 - \gamma) \odot \theta_{ft} \]

γ is learned to minimize \( L_{food}(\theta_{ft}(\gamma)) \).

Compared to original FT model and many parameter-efficient methods, grafting shows better skill localization since it has
- better calibration while retaining original performance,
- better performance in OOD tasks,
- better skill composition via unification of grafts in multi-task learning (and continual learning) ¹.

Example

- 0.01% Grafting
- SST-2: WiSE-Graft applies Grafted model and the pre-trained model.
- SST-2: WiSE-Graft interpolates smoothly between Grafted model and the pre-trained model.

Grafting in multi-task setting

On a multi-task trained model, we learn task-specific grafts by learning binary masks γ on individual tasks.

- Overlap between grafts is a measure of task similarity ¹.
- Grafted model for a task transfers across similar tasks (Fig. a).
- Union of grafts for a subset of tasks perform well on only the subset and related tasks (Fig. b).

Comparison with BitFit[1] and LoRA[2]

- Graft regions form meaningful sub-nets.
- Re-training on 0.01% graft parameters recovers FT performance → graft regions form meaningful sub-nets.
- Comparison with BitFit[1] and LoRA[2] → sparsity of updates isn’t the only factor behind good calibration results.

AdamW requires explicit ℓ1 regularization for sparse grafts

Graft parameters ≠ parameters with the biggest movement.

Layer-wise distribution: Concentrated in the middle layers

Component-wise distribution: Mostly first MLP layer and Value weights