A APPENDIX

A.1 COMPARISON TO META-LEARNING (MAML) + ITERATIVE PRUNING

In Figure 11, we compare MVP with IHT-based Reptile (Tian et al., 2020), which is a meta-pruning method that applies meta-learning to initializing the model weights for pruning. In particular, IHT-based Reptile applies Reptile (Nichol et al., 2018) and iterative pruning to find better weight-initialization for a pruned meta-model. Given a new task, it fine-tunes the pruned meta-model for a limited number of iterations to obtain the final pruned model. The results show that MVP achieves higher accuracy than IHT-based Reptile when their training iterations and the amount of training data are the same, implying that MVP can find better initialization of pruned models for different tasks.

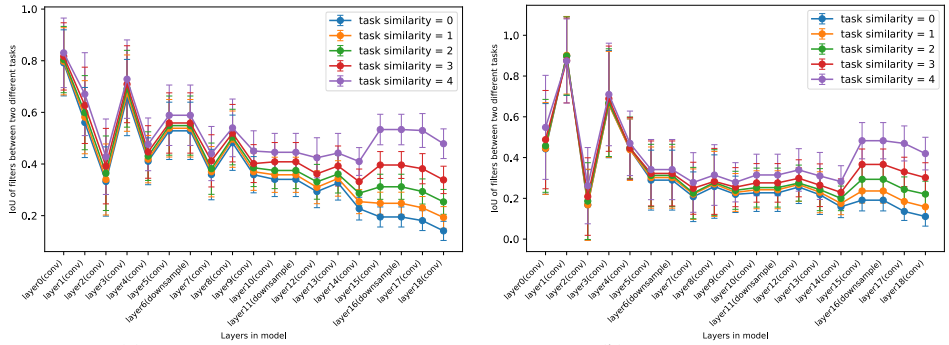
A.2 EFFICIENCY OF HIGHER TASK SIMILARITIES

Much more iterations is required for the accuracy of low similarity tasks to match the accuracy of high similarity tasks. For instance, 200 and 300 more iterations are required for the accuracy of (N=2, task similarity=1) to match the accuracy of (N=2, task similarity=4) for CIFAR-100 pruned models of 90% and 96% pruning ratio. There are two main reasons: (1) more filters from the less similar tasks' models need to be heavily fine-tuned for the target task; and (2) iterative pruning for extremely high pruning ratios is slow because only a few filters can be finally preserved and many competitions or tug-of-wars happen among filters during pruning.

A.3 ADDITIONAL AND COMPLETE RESULTS FOR FIG. 3, FIG. 4, FIG. 5 AND FIG. 6

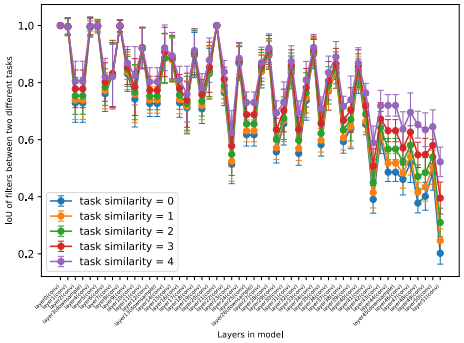
Besides the prototype similarity, we conduct a group of experiments to evaluate MVP in a "weakly-supervised" case where the test task's classes never appear in any training task, though they are allowed to share parent class with some classes in training tasks and the task similarity is changed to the number of shared parent classes between two tasks. In particular, the set of classes over all training tasks is disjoint with that over all test tasks so $|C_i \cap C_j| = 0$, and the task similarity is measured by the number of shared parent-classes (provided in both CIFAR-100 and Tiered-ImageNet) ranging in $\{1, 2, 3, 4, 5\}$, resulting in five sets of training tasks. The results are shown in the third-row plots of Figure 12 and Figure 13. The performance of the weak similarity cases is similar to that of the prototype similarity. When weak similarity increases, the accuracy improves though we seem to lose the advantage of large N due to the noises by involving more weakly-similar tasks.

In the following, Figure 7 and Figure 10 show more experimental results in addition to Figure. 3 and Figure. 5. Figure 8 and Figure 9 report the complete results for the left and right plot of Figure 4, respectively. Figure 12-15 report more experimental results in addition to Figure. 6.

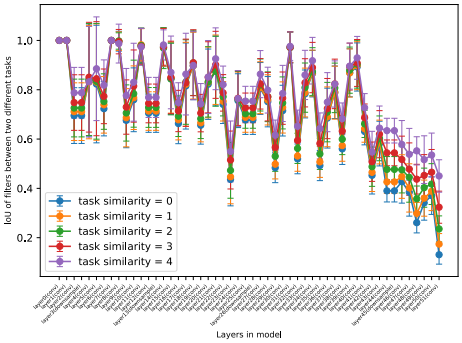


(a) CIFAR-100, pruning ratio: 90%

(b) CIFAR-100, pruning ratio: 96%



(c) Tiered-ImageNet, pruning ratio: 90%



(d) Tiered-ImageNet, pruning ratio: 96%

Figure 7: IoU (mean±std) measuring filter sharing between two tasks of different similarity$\in \{0, 1, 2, 3, 4\}$ in each layer of their pruned models, for all the layers from input to output (left to right).

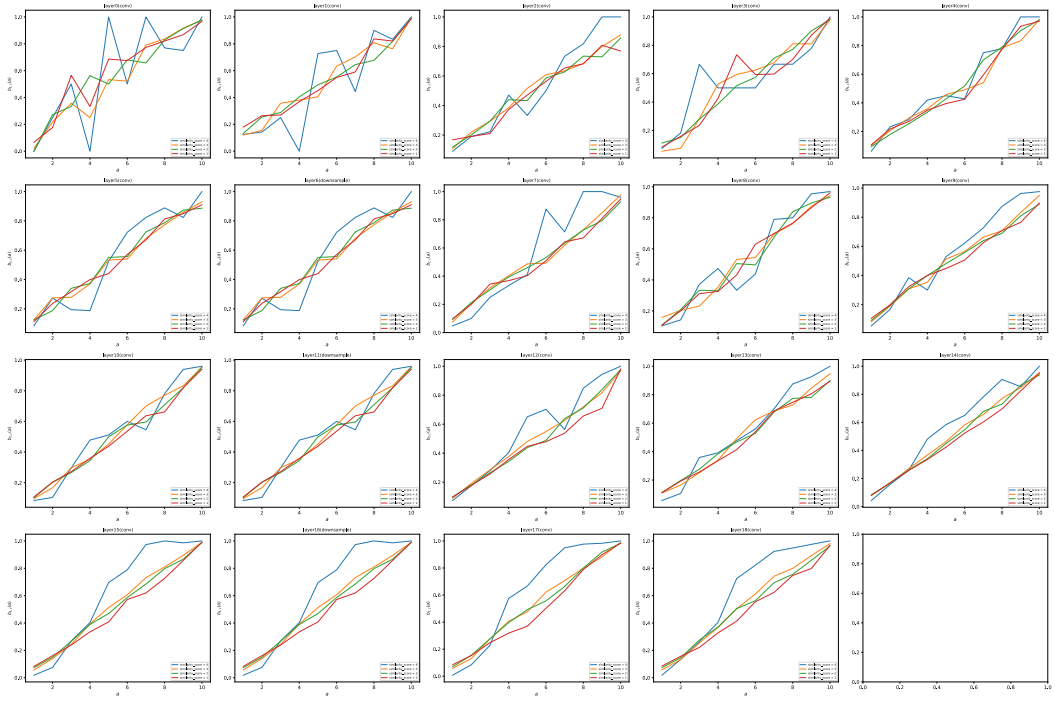


Figure 8: Each plot corresponds to a different layer ℓ and reports the odds $b_{\ell,c}(a)$ of a filter from similar tasks' models being selected in a target task's model if it is selected by a out of 10 similar tasks with certain similarity: a vs. $b_{\ell,c}(a)$, $b_{\ell,c}(a)$ is defined in Sec. 3.3. The left plot in Figure 4 averages $b_{\ell,c}(a)$ over all layers $\ell \in [L-1]$ and similarities $c \in [4]$.

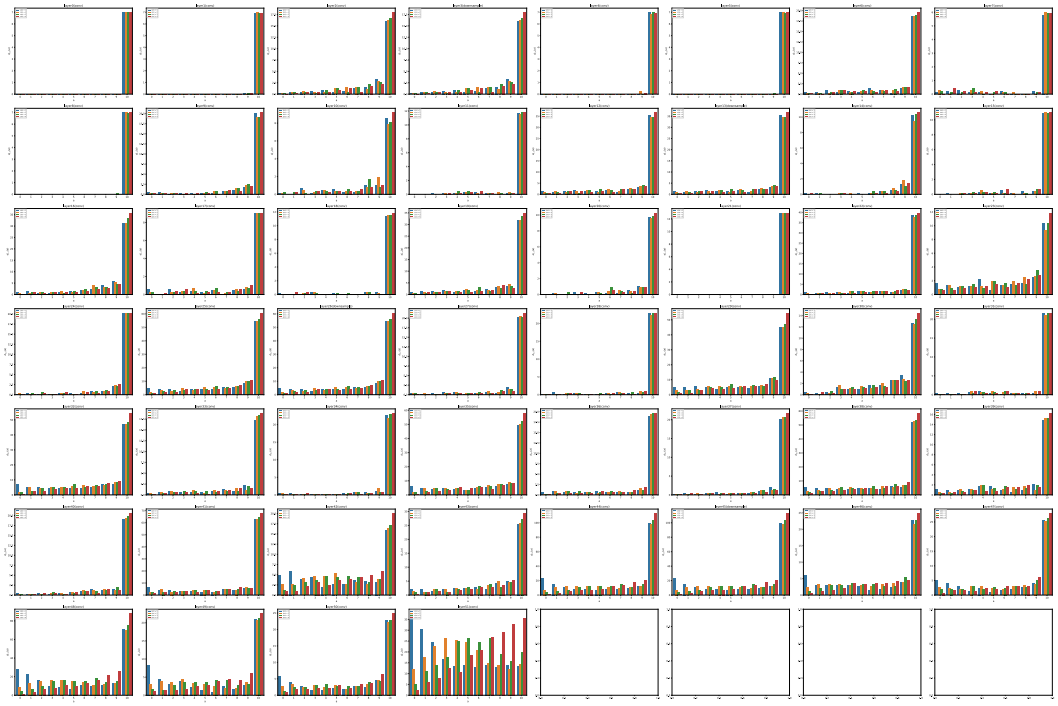


Figure 9: Each plot corresponds to a different layer ℓ of the target task's model and shows the histogram of how many filters are selected by a out of 10 similar tasks with certain similarity: a vs. $d_{\ell,c}(a)$. $d_{\ell,c}(a)$ is defined in Sec. 3.3. The right plot in Figure 4 averages $d_{\ell,c}(a)$ over all layers $\ell \in [L-1]$.

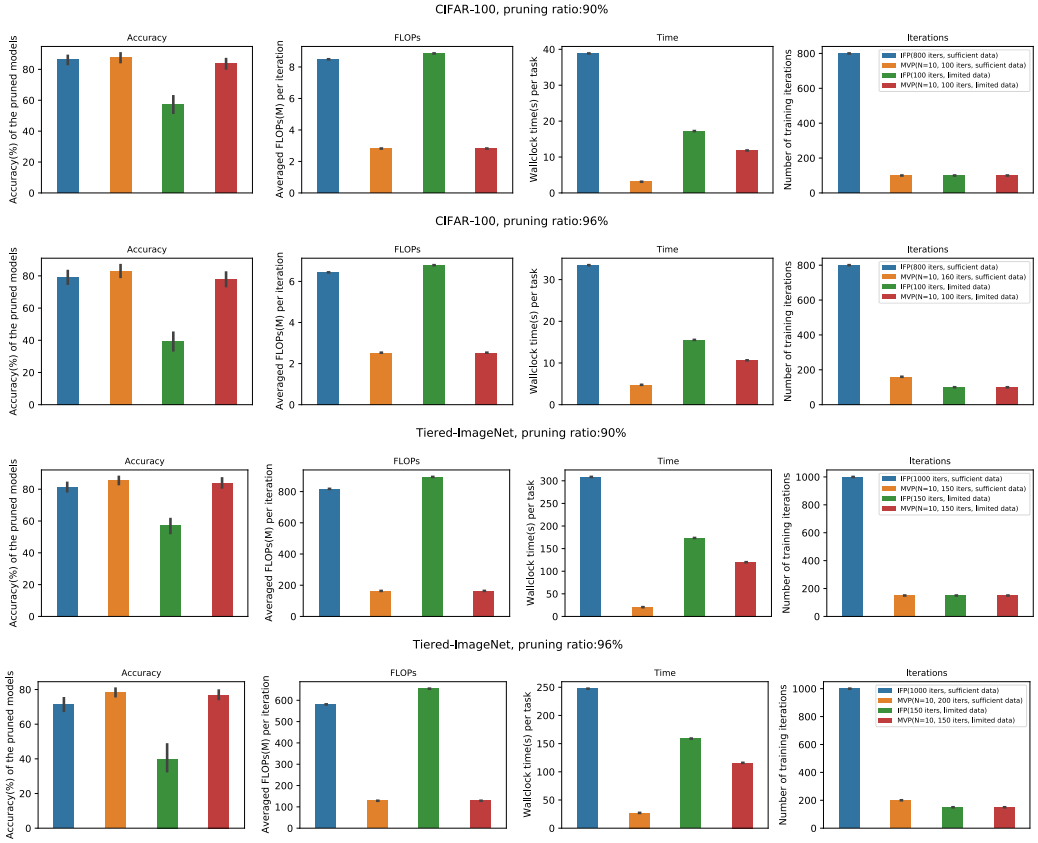


Figure 10: Test accuracy, memory and computational costs of MVP and IFP in the experiments.

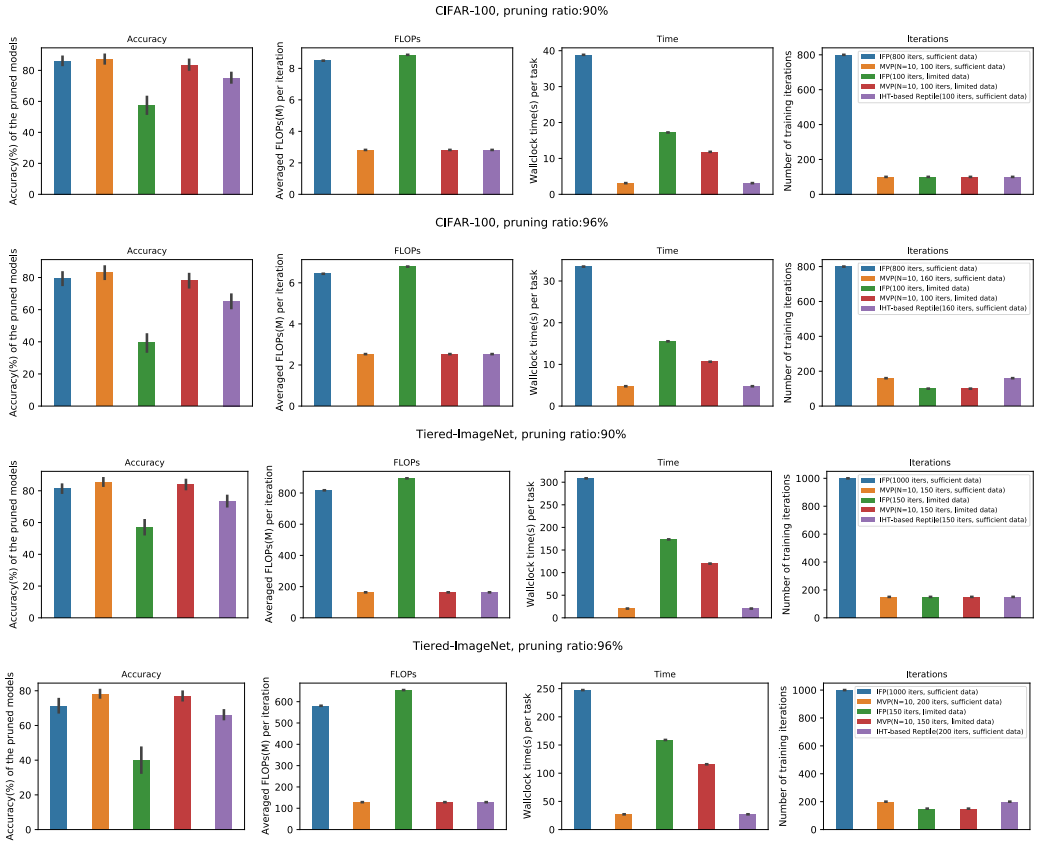


Figure 11: Test accuracy, memory and computational costs of MVP, IFP and IHT-based Reptile (Tian et al., 2020).

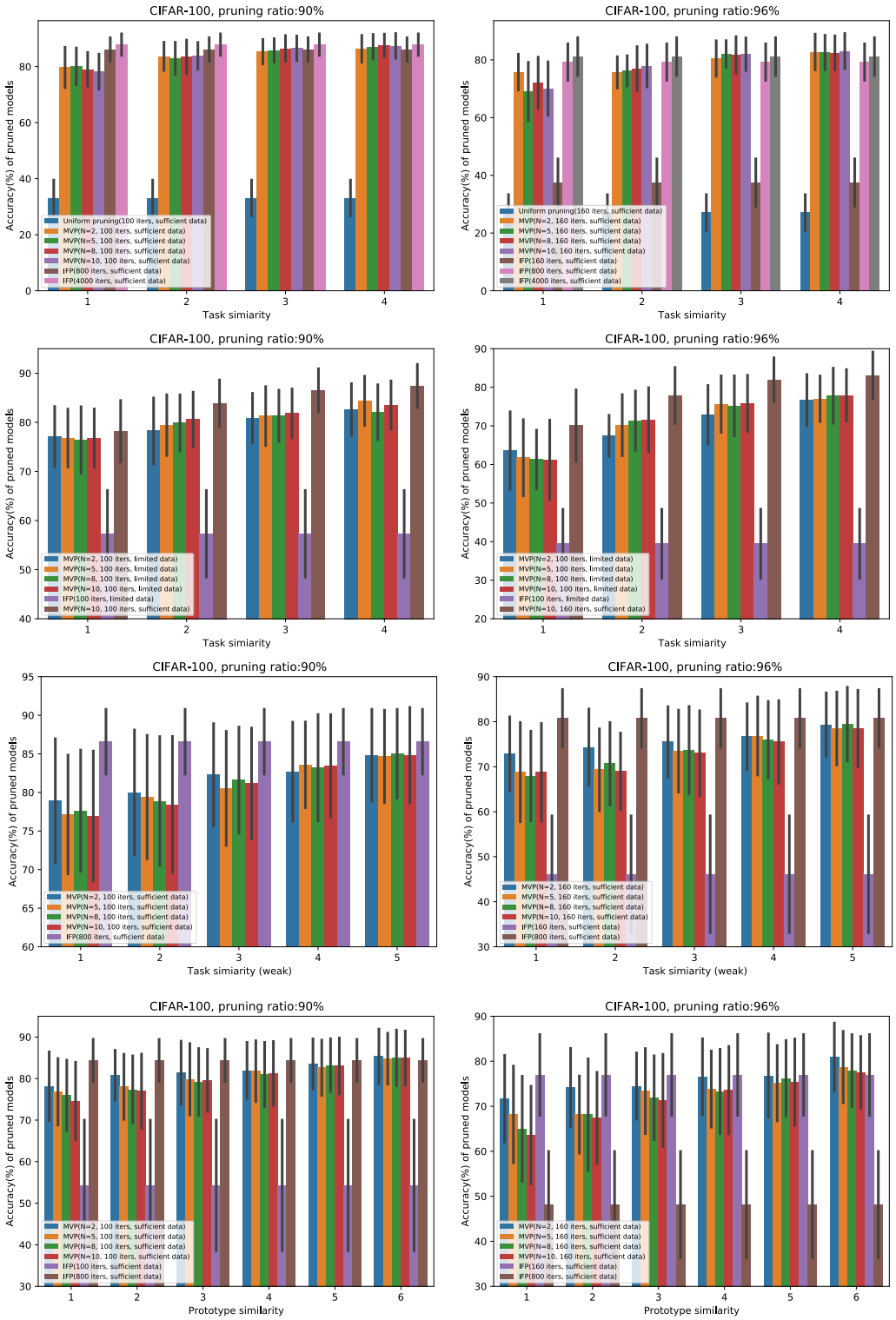


Figure 12: Test-set accuracy of pruned models produced by different methods (MVP, IFP, uniform pruning) using different number of iterations (**Top-row plots**), different amount of training data (**Second-row plots**), training tasks with weaker similarity (**Third-row plots**) and training tasks with prototype similarity (**Bottom-row plots**) for CIFAR-100.

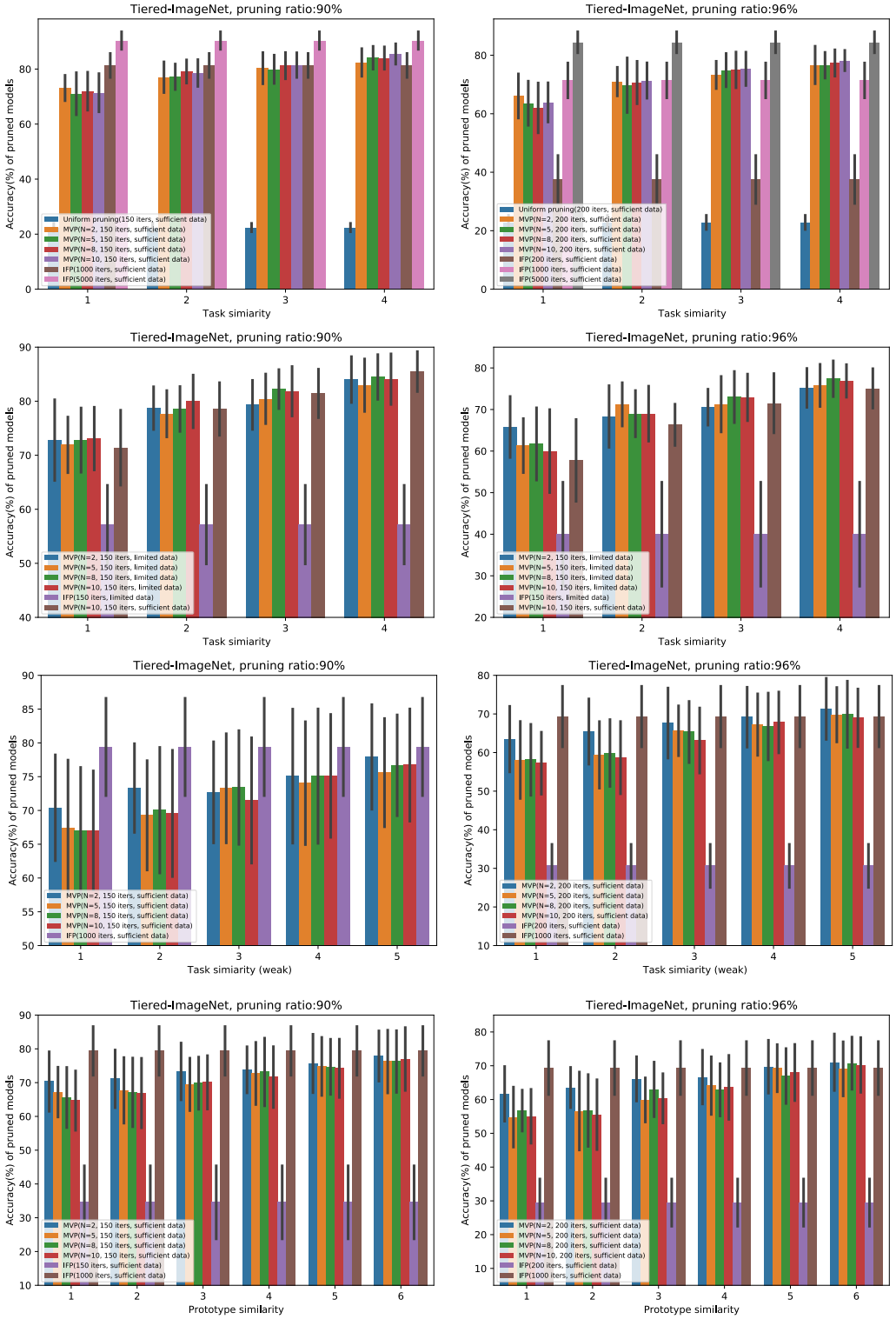


Figure 13: Test-set accuracy of pruned models produced by different methods (MVP, IFP, uniform pruning) using different number of iterations (**Top-row plots**), different amount of training data (**Second-row plots**), training tasks with weaker similarity (**Third-row plots**) and training tasks with prototype similarity (**Bottom-row plots**) for Tiered-ImageNet.

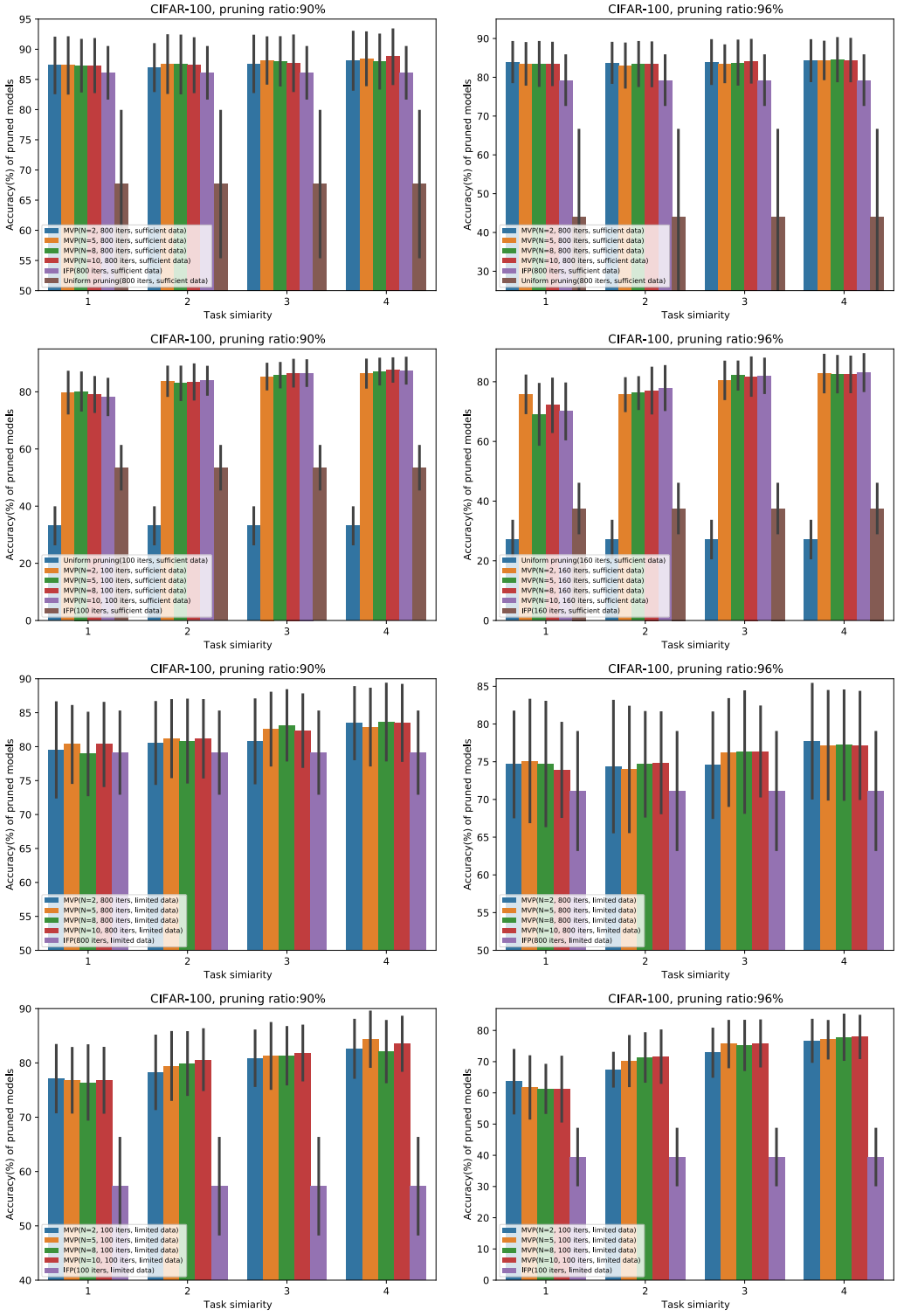


Figure 14: Test-set accuracy of pruned models produced by different methods (MVP, IFP, uniform pruning) using large number of iterations and sufficient training data (**Top-row plots**), small number of iterations and sufficient training data (**Second-row plots**)), large number of iterations and limited training data (**Third-row plots**)), small number of iterations and limited training data (**Bottom-row plots**)) for CIFAR-100.

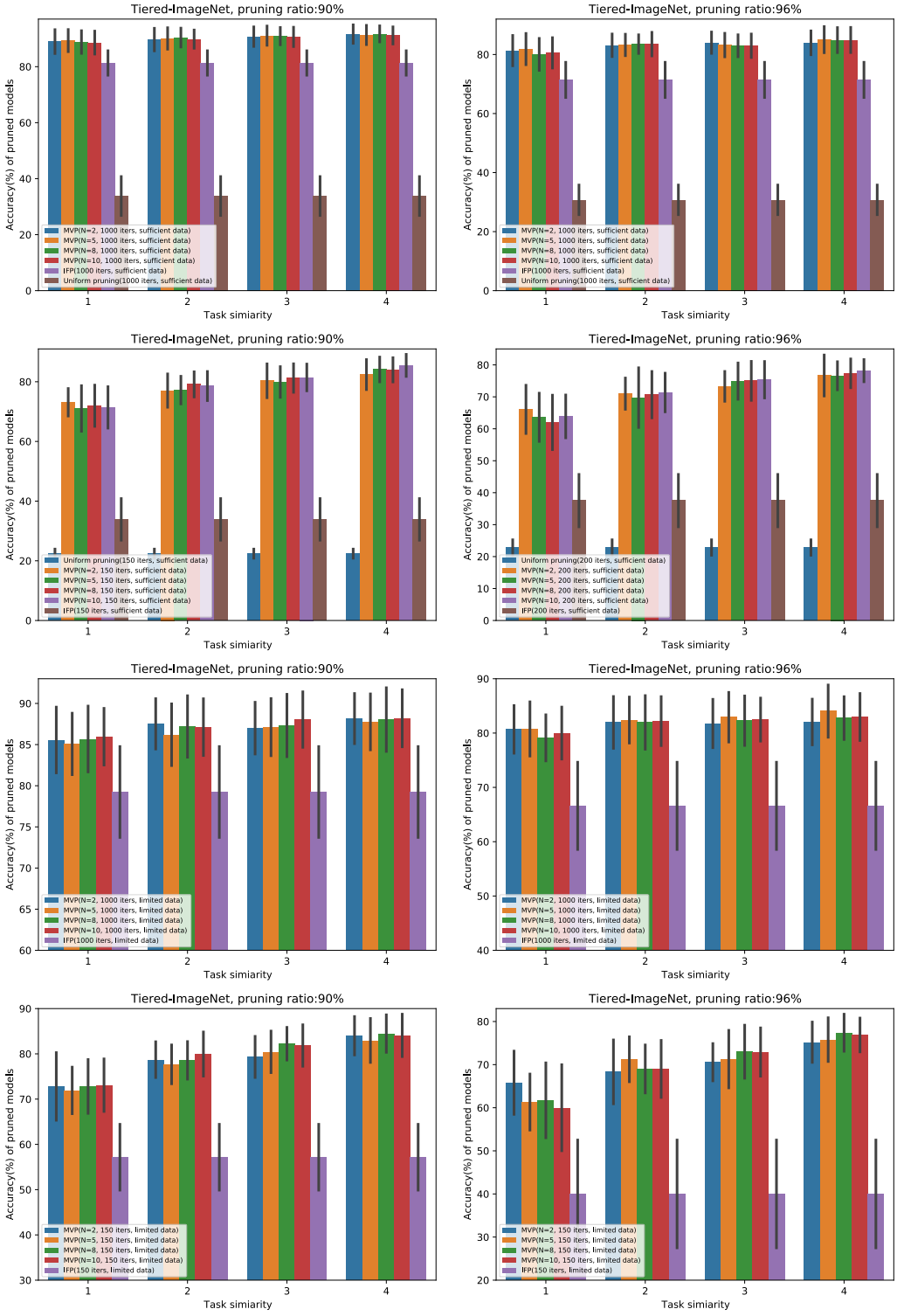


Figure 15: Test-set accuracy of pruned models produced by different methods (MVP, IFP, uniform pruning) using large number of iterations and sufficient training data (**Top-row plots**), small number of iterations and sufficient training data (**Second-row plots**)), large number of iterations and limited training data (**Third-row plots**)), small number of iterations and limited training data (**Bottom-row plots**)) for Tiered-ImageNet.