Evolutionary Optimization of Model Merging Recipes

Author: Akiba et al.

created by paper2slides

2024-03-19

https://arxiv.org/abs/2403.13187

Executive Summary

- This work automates the creation of foundation models by leveraging evolutionary algorithms for model merging.
- The approach merges models in parameter space and data flow space to discover effective combinations of diverse models.
- Achieved state-of-the-art results on Japanese LLM benchmarks with models generated through this evolutionary method.
- Open-sourced novel Japanese LLM and VLM models with improved capabilities in math reasoning and cultural context understanding.

Introduction

- Model merging is a novel technique that combines multiple pre-trained models to create a single model with enhanced capabilities.
- Traditional model merging relies heavily on human intuition and domain knowledge, which limits its potential.
- Evolutionary algorithms can systematically discover new model combinations, extending beyond human intuition.
- This work applies evolutionary algorithms to optimize model merging recipes in both parameter space and data flow space.

Model Merging Background

- Model merging combines task-specific models into a comprehensive one without retraining.
- Simple methods like weight averaging have shown promise in image processing and generative models.
- Advanced methods like TIES-Merging and DARE address weight interference in language models.
- Model merging is gaining popularity in the Open LLM community for developing new, cost-effective models.

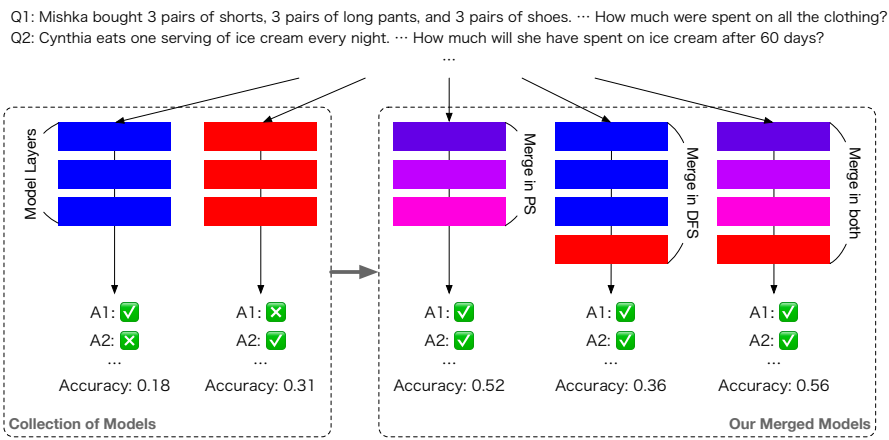

Figure: Overview of Evolutionary Model Merge. The method involves optimizing in both parameter space and data flow space.

Proposed Method: Parameter Space

- Parameter space merging involves integrating weights of multiple models with the same architecture.
- This approach uses techniques like task vector analysis and TIES-Merging with DARE for merging.
- The parameters for sparsification and weight mixing are optimized using evolutionary algorithms like CMA-ES.
- The final model is evaluated on specific tasks, optimizing metrics like accuracy and ROUGE score.

Proposed Method: Data Flow Space

- Data flow space merging preserves weights and optimizes the inference path among layers of different models.
- Large search space necessitates skills to reduce complexity, like managing layer repetitions and input scaling.
- Evolves an indicator array identifying which layers to include or exclude, evolved using CMA-ES.
- Initializing paths with layers from one model and scaling inputs to manage distribution shifts.

Dataset and Evaluation for LLM

- Evaluated on MGSM dataset, a multilingual subset of the GSM8k dataset for mathematics.
- Training set was translated into Japanese to form a disjoint dataset for evolutionary search.
- Evaluation based on the final numerical value and reasoning text in Japanese.
- Accuracy calculated on zero-shot pass@1 performance.

Experimental Results: LLM

- Source models include Shisa Gamma 7B (Japanese LLM), WizardMath-7B, and Abel-7B (Math LLMs).
- Merged models achieved notable improvements: 52.0 (PS merge) and 55.2 (PS+DFS merge) on MGSM-JA.
- Models also excelled in JP-LMEH benchmark, surpassing some state-of-the-art 70B parameter models.
- Analysis confirmed merged models retain foundational knowledge and show emergent capabilities.

Breakdown of LLM Results

Table: Performance Comparison of LLMs.

Id.	Model	Type	MGSM-JA (acc)	JP-LMEH (avg)
1	Shisa Gamma 7B v1	JA general	9.6	66.1
2	WizardMath 7B V1.1	EN math	18.4	60.1
3	Abel 7B 002	EN math	30.0	56.5
4	Ours (PS)	Merged	52.0	70.5
5	Ours (DFS)	Merged	36.4	53.2
6	Ours (PS+DFS)	Merged	55.2	66.2

Proposed Method: VLM

- Extended method to multi-modal VLMs merging Japanese LLMs with LLM components in VLMs.
- Source models: Shisa-Gamma-7b-v1 (Japanese LLM) and `LLaVA-1.6-Mistral-7B` (VLM).
- Created new benchmark datasets: JA-VG-VQA-500 and JA-VLM-Bench-In-the-Wild.
- ROUGE-L metric computed using a Japanese language detector to evaluate model responses.

Experimental Results: VLM

- Evaluated on JA-VG-VQA-500 and JA-VLM-Bench-In-the-Wild datasets.
- Achieved superior performance compared to baselines in both datasets.
- ROUGE-L scores: 19.7 on JA-VG-VQA-500 and 51.2 on JA-VLM-Bench-In-the-Wild.
- Results demonstrate the model's capability to handle culturally specific content while enhancing Japanese VQA performance.

Final VLM Performance

Table: Performance Comparison of the VLMs.

Model	JA-VG-VQA-500 (R-L)	JA-VLM-Bench (R-L)
LLaVA-1.6-Mistral-7B	14.3	41.1
Japanese Stable VLM	0.0	40.5
Ours	19.7	51.2

Conclusions and Future Work

- Evolutionary optimization of model merging effectively automates foundation model creation.
- Novel merged models outperform existing models on Japanese LLM and VLM benchmarks.
- Future work includes expanding to image diffusion models and evolving swarms of diverse models.
- Potential for significant resource savings and quick prototype development in foundation model training.