Diversity-Driven Generalization in Mathematical Reasoning Ensembles

#diversity

Author:
Nicholas Jiang

0. Abstract

We propose a framework for studying collaborative generalization in mathematical reasoning systems. Inspired by work from Lightman et al., which showed that stepwise supervision improves reasoning [1], we investigate whether solvers can learn from one another's complete solutions in a fully self-supervised setting. Using the miniF2F benchmark [2], we plan to construct multiple ensembles of solvers with varying degrees of diversity, measured by a novel Task2Vec-based Ensemble Diversity Coefficient (EDC) inspired by Miranda et al. [3]. We will fine-tune ensemble members on peer-generated proofs and evaluate generalization on held-out problem sets. We hypothesize that EDC will significantly predict ensemble improvement and that diverse ensembles will demonstrate stronger generalization from peer learning.

1. Central Thesis

Ensemble diversity enables collaborative generalization: Self-supervised learning from peer successes is hypothesized to be more effective when solvers in the ensemble are diverse.

We formalize this idea using a measure of diversity based on Task2Vec embeddings [4]. We then propose to quantify how much ensemble diversity (via EDC) predicts downstream generalization gain after collaborative fine-tuning.

2. Experimental Design

Dataset

- miniF2F benchmark [2]: A suite of formal and informal math problems.
- K=20 repeated random 50/50 splits of the dataset into:
  - Train A: for collaborative fine-tuning.
  - Test B: for evaluation of generalization.

Solvers

- A pool of LLM-based solvers (e.g., different seeds, architectures, or fine-tuning paths)
- Solvers output formal proofs in a language such as Lean 4, allowing automatic verification of correctness.
  - An example of such a model is the DeepSeek-Prover-V1.5 [5] that has shown >50% baseline accuracy on miniF2F
- We will construct multiple distinct ensembles of equal size by sampling from this pool of solvers.

Diversity Metric

Inspired by Miranda et al. [3], we extend the idea of diversity coefficient for datasets to solvers:

$$\mathrm{cmdiv}(m_1, m_2; D) \;=\; \mathbb{E}_{B\sim D}\Big[\, d\big(\vec{f}_{m_1,B}, \vec{f}_{m_2,B}\big)\Big] \tag{1}$$

$$\mathrm{EDC}(\mathcal{M}; D) \;=\; \mathbb{E}_{\text{two distinct } m\sim\mathcal{M}}\big[\mathrm{cmdiv}(m_1, m_2; D)\big] = \frac{1}{\binom{M}{2}} \sum_{1\le i<j\le M} \mathrm{cmdiv}(m_i, m_j; D) \tag{2}$$

where $\mathcal{M} = \{m_1, m_2, \ldots, m_M\}$ is the ensemble and each m_i is a solver, $\vec{f}_{m_i,B}$ is a Task2Vec embedding [4] on batch B from dataset D from solver m_i, d is a vector distance measure (e.g. cosine distance)

Compare Model Diversity Coefficient (cmdiv) is the average distance in the two solvers' embedding of batches from a dataset D. It is meant to quantify the difference in the two models' internal representations of a dataset.

Ensemble Diversity Coefficient (EDC) is the average pairwise diversity across solvers in the ensemble. Note that it is averaged over unordered pairs because of the symmetry of cmdiv.

Learning Protocol

For each constructed ensemble i=1…N:

- Compute EDC_i from miniF2F problems
- We will apply the following protocol across K random splits of the dataset:
  - Baseline:
    - Measure individual and ensemble accuracy on Test B for split s: $S_{i,0,s}$.
  - Collaborative Training:
    - For problems in Train A of split s where only one model in ensemble i succeeds, fine-tune the other ensemble members on that solution.
  - Evaluation:
    - Re-evaluate ensemble i on Test B of split s after peer-learning: $S_{i,1,s}$.
  - Compute boost for ensemble i on split s: $\Delta S_{i,s} = S_{i,1,s} - S_{i,0,s}$.

3. Analysis

We will assess whether ensemble diversity (EDC) predicts ensemble improvement across the constructed ensembles using:

Regression Model

We regress the per-split increase in overall test-set accuracy on ensemble diversity and the ensemble's own baseline accuracy over all ensembles i=1,...,N and all K=20 random splits

$$\Delta S_{i,s} = \beta_0 + \beta_1\, EDC_i + \beta_2\, S_{i,0,s} \; + \; u_i + v_s + \varepsilon_{i,s}$$

with

$$u_i \sim \mathcal{N}(0, \sigma_u^2), \quad \varepsilon_{i,s} \sim \mathcal{N}(0, \sigma^2), \quad v_s \sim \mathcal{N}(0, \sigma_v^2)$$

- EDC_i: The Ensemble Diversity Coefficient for ensemble i.
- $\Delta S_{i,s}$: Accuracy gain of ensemble i on split s' Test B after fine-tuning
- $S_{i,0,s}$: The baseline Test B accuracy of ensemble i on split s
- β_1: Effect of diversity after adjusting for baseline accuracy
- u_i: Ensemble random intercept (unmodelled persistent traits)
- v_s: Split random intercept (correlation induced by re-used problems)
- $\varepsilon_{i,s}$: residual error for the i,s observation
- Statistical significance tested via t-test ($H_0 : \beta_1 = 0$).

Model will be fitted computationally (e.g. via Python's statsmodels.MixedLM) to estimate σ_u^2, σ_v^2 and returns BLUPs for u_i and v_s

Effect Size Measures

- Standardized β1$*$: Measures impact of EDC in standard deviation units.
- Cohen's f2: Evaluates overall variance explained by the model.

---

4. References

[1] H. Lightman *et al.*, "Let's Verify Step by Step," May 31, 2023, *arXiv*: arXiv:2305.20050. doi: 10.48550/arXiv.2305.20050.

[2] K. Zheng, J. M. Han, and S. Polu, "MiniF2F: a cross-system benchmark for formal Olympiad-level mathematics," Feb. 28, 2022, *arXiv*: arXiv:2109.00110. doi: 10.48550/arXiv.2109.00110.

[3] B. Miranda, A. Lee, S. Sundar, A. Casasola, and S. Koyejo, “Beyond Scale: The Diversity Coefficient as a Data Quality Metric for Variability in Natural Language Data,” Aug. 26, 2024, *arXiv*: arXiv:2306.13840. doi: 10.48550/arXiv.2306.13840.

[4] A. Achille *et al.*, “Task2Vec: Task Embedding for Meta-Learning,” Feb. 10, 2019, *arXiv*: arXiv:1902.03545. doi: 10.48550/arXiv.1902.03545.

[5] H. Xin *et al.*, “DeepSeek-Prover-V1.5: Harnessing Proof Assistant Feedback for Reinforcement Learning and Monte-Carlo Tree Search,” Aug. 15, 2024, *arXiv*: arXiv:2408.08152. doi: 10.48550/arXiv.2408.08152.