# Exploring the Efficacy of Meta-Learning: Unveiling Superior Data Diversity Utilization of MAML Over Pre-training

**Reviewed on OpenReview:** `https://openreview.net/forum?id=RDpDFzijX4`

## Abstract

Currently, data and model size dominates the narrative in the training of super-large, powerful models. However, there has been a lack of exploration on the effect of the *quality* of the training dataset on performance. In this work, we show positive correlations between accuracy and data diversity, providing an argument for the research of data "quality" beyond size. In our analysis of pre-training and model-agnostic meta-learning methods on twelve popular visual datasets (e.g., Omniglot, CIFAR-FS, Aircraft) and five model configurations, including MAML variants with different numbers of inner gradient steps and supervised learning, we show moderate to strong positive correlations (R-squared: 0.15-0.42) between accuracy and data diversity, and weaker but significant correlations (R-squared: 0.2) between loss and diversity. These findings support our hypothesis and pave the way for deeper exploration of how data quality, captured by diversity, influences model performance. This initial study highlights the potential of Task2Vec diversity as a valuable measure in the rapidly evolving field of large-scale learning, where understanding data quality is key to build more powerful and generalizable models.

## 1 Introduction

Current trends in building large, robust vision models emphasize scaling model size and complexity, but other characteristics of datasets, like quality, are vague and overlooked. Our study departs from the prevailing focus on data size and explores this gap by proposing a paradigm shift: embracing dataset diversity as a key ingredient for building more generalizable and performant vision models. By analyzing the impact of diverse datasets on popular vision models, we aim to shed light on this under-explored facet of AGI research.

## 2 Methodology and Results

Our study measures training data diversity using the Task2Vec metric, which quantifies intrinsic datset variability in a few-shot learning setting. We analyze pre-training and model-agnostic meta-learning methods on twelve popular visual datasets (e.g., Omniglot, CIFAR-FS, Aircraft). We explored supervised learning and meta-learning settings through three model configurations: a supervised learning model (SL) and first-order (FO) and higher-order (HO) meta-learning models with Model-Agnostic Meta-Learning (MAML). Performance was measured by test accuracy across all configurations.

Figure 1 depicts the observed relationship for each model. Notably, FO and HO MAML models with 5 and 10 inner gradient steps demonstrate a significantly stronger positive

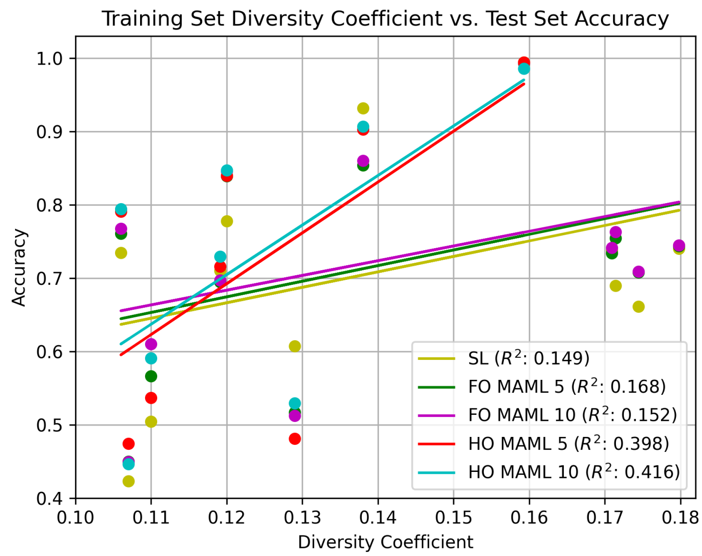


Figure 1: Higher-Order Meta-Learning Leverages Data Diversity More Effectively than Supervised Learning

correlation between dataset diversity and relative model performance compared to the SL model and FO MAML models with fewer inner gradient steps.

This observation is reinforced by the R-squared correlation coefficient, quantifying the proportion of variance in test accuracy explained by dataset diversity. The statistically positive correlation observed between the diversity coefficient and test accuracy, suggesting that data diversity generally improves model performance. Significantly higher $R^2$ values were observed for HO MAML 5 and HO MAML 10 models. The HO MAML 10 model exhibits a remarkable 40% and 20% explained variance in both accuracy and cross-entropy loss, respectively, attributable to dataset diversity. Similar trends were observed for HO MAML 5. We acknowledge a potential confounding factor: uncontrolled amount of data points across training checkpoints. To address this, we propose selecting checkpoints with similar average Task2Vec complexity, ensuring a fairer comparison based on intrinsic data variability.

## 3 Conclusion

Our research unveils the potential of data diversity as a crucial factor in boosting the performance of vision models, beyond the current emphasis on scaling model size and complexity. Our analysis revealed a positive association between data diversity and model performance, indicated by a consistent upward trend across all investigated configurations. Notably, meta-learning models demonstrated a significantly stronger positive correlation with diversity compared to supervised learning approaches. Our findings suggest a paradigm shift towards "quality-aware" data selection could address the growing computational costs associated with scaling models and contribute to a more efficient and impactful research trajectory in the exploration of more powerful AI.

## Broader Impact Statement

By fostering rigorous data quality analysis, this research empowers both researchers and practitioners to make informed decisions in LLM development. Societal consequences of this work include reduced costs through the improvement of data selection, which subsequently minimizes required data quantities, lowering development and inference costs and making LLMs more accessible. Another important consequence is the potential environmental gains from reduced energy consumption achieved by more efficient training. Finally, research on data quality through diversity naturally encourages data fairness. Further exploration is warranted to uncover additional societal and ethical implications.

## Acknowledgments and Disclosure of Funding

No funding was involved with this work. We would also like to acknowledge Xiaolei Shi and Daniel Kolano for their contributions to this work.

## References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning, 2019.

Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *IEEE Access*, 7:64323–64350, 2019. doi: 10.1109/ACCESS.2019.2917620.

Alycia Lee, Brando Miranda, Sudharsan Sundar, and Sanmi Koyejo. Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data, 2023.

Brando Miranda, Patrick Yu, Yu-Xiong Wang, and Sanmi Koyejo. The curse of low task diversity: On the failure of transfer learning to outperform maml and their empirical equivalence, 2022.

Brando Miranda, Patrick Yu, Saumya Goyal, Yu-Xiong Wang, and Sanmi Koyejo. Is pre-training truly better than meta-learning? *arXiv preprint arXiv:2306.13841*, 2023.

Vivek Ramanujan, Thao Nguyen, Sewoong Oh, Ludwig Schmidt, and Ali Farhadi. On the connection between pre-training data diversity and fine-tuning robustness. *arXiv preprint arXiv:2307.12532*, 2023.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023.