Team 2 Project Proposal

Few-Shot Transfer Learning for Join Order Selection

1. Introduction and Motivation

Database query performance depends critically on join order selection. While traditional optimizers use heuristics, they often produce suboptimal plans. Recent research shows Graph Neural Networks (GNNs) can learn better join orders, but these models require costly retraining for each new database schema. We investigate whether a GNN pretrained on a source workload can adapt to new schemas using only 10-50 example queries through few-shot transfer learning—a crucial step toward practical learned optimizers.

2. Core Research Question

Can a GNN-based model learn schema-agnostic representations that enable fine-tuning to new database schemas using only 10-50 labelled queries, while maintaining performance comparable to models trained from scratch?

3. Proposed Methodology

Phase 1 - Graph Representation & Pretraining: Model queries as graphs (tables as nodes, joins as edges) with schema-invariant features (normalized cardinality, selectivity, join types). Pretrain GNN encoder on TPC-H dataset to learn foundational query patterns.

Phase 2 - Few-Shot Fine-Tuning: Transfer pretrained GNN to IMDB dataset using ~50 labelled queries. Compare three strategies:

- Head-Only: Freeze GNN body, train output layer only.
- Adapters: Insert small trainable modules between frozen layers.
- LoRA: Learn low-rank weight matrix updates

Phase 3 - Evaluation: Compare against database optimizer and from-scratch GNN using:

- Runtime improvement vs. optimal plan.
- Ranking quality (NDCG@k).
- Sample efficiency metrics

4. Anticipated Challenges

- **Feature Engineering:** Creating truly schema-invariant features for effective transfer.
- **Overfitting Risk:** Preventing memorization with minimal training samples.
- **Adaptation Balance:** Optimizing frozen vs. trainable layer configuration

5. Project Timeline

Week 1: Setup TPC-H/IMDB workloads, query generation scripts

Week 2: Implement exhaustive plan enumeration for labeled data

Week 3: Implement/pretrain GNN on TPC-H, establish baselines

Week 4: Few-shot fine-tuning experiments (10/25/50 samples)

Week 5: Ablation studies comparing tuning methods

Week 6: Evaluation, analysis, begin report writing

Week 7: Finalize report and presentation

6. Team Responsibilities

- Task 1 (Mihir, Ruthvik) – Workload setup, query generation, feature pipeline, initial documentation
- Task 2 (Suhas, Aditya)– Encoder architecture design, source workload training, pipeline integration & documentation.
- Task 3 (Aditya, Manasa, Mihir) – Implement adaptation strategies, run experiments, track metrics and experiment logs.
- Task 4 (Manasa, Suhas, Ruthvik) – Develop metrics framework, benchmark against baselines, document evaluation results.
- Task 5 (ALL)– Results analysis, report writing, presentation prep, contribution summaries from all members.

7. Deliverables

- Pretrained and fine-tuned GNN models.
- Complete codebase (data generation, feature extraction, evaluation).
- Comprehensive final report and presentation slides

References:

[1] Li, Z., Li, Y., Luo, Y., Li, G., & Zhang, C. (2025). Graph Neural Networks for Databases: A Survey. arXiv preprint arXiv:2502.12908. https://arxiv.org/abs/2502.12908

[2] "Debunking the Myth of Join Ordering: Toward Robust SQL Execution." (2025). arXiv preprint arXiv:2502.15181. https://arxiv.org/abs/2502.15181

[3] Trummer, I., et al. (2024). JoinGym: An Efficient Join Order Selection Environment. RLJ / UMass. https://rlj.cs.umass.edu/2024/papers/RLJ_RLC_2024_14.pdf

[4] Zhu, J., Cai, S., Shen, Y., Chen, G., Deng, F., & Ooi, B. C. (2025). In-Context Adaptation to Concept Drift for Learned Database Operations. arXiv preprint arXiv:2505.04404. https://arxiv.org/abs/2505.04404

[5] Giannakouris, V., & Trummer, I. (2025). Rethinking Pluggable Federated Query Optimization. Proceedings of VLDB Workshops. https://www.vldb.org/2025/Workshops/VLDB-Workshops-2025/CDMS/CDMS25_07.pdf

[6] Garralda-Barrio, M., et al. (2025). Adaptive Incremental Transfer Learning for Efficient Workload Prediction. Information Processing & Management. https://www.sciencedirect.com/science/article/abs/pii/S0167739X25000251

[7] "Enhancing Text-to-SQL with Dynamic Few-shot and Alignment." (2025). arXiv preprint. https://arxiv.org/html/2502.14913v1