NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

SCSE22-0406 – Query Cost Estimation

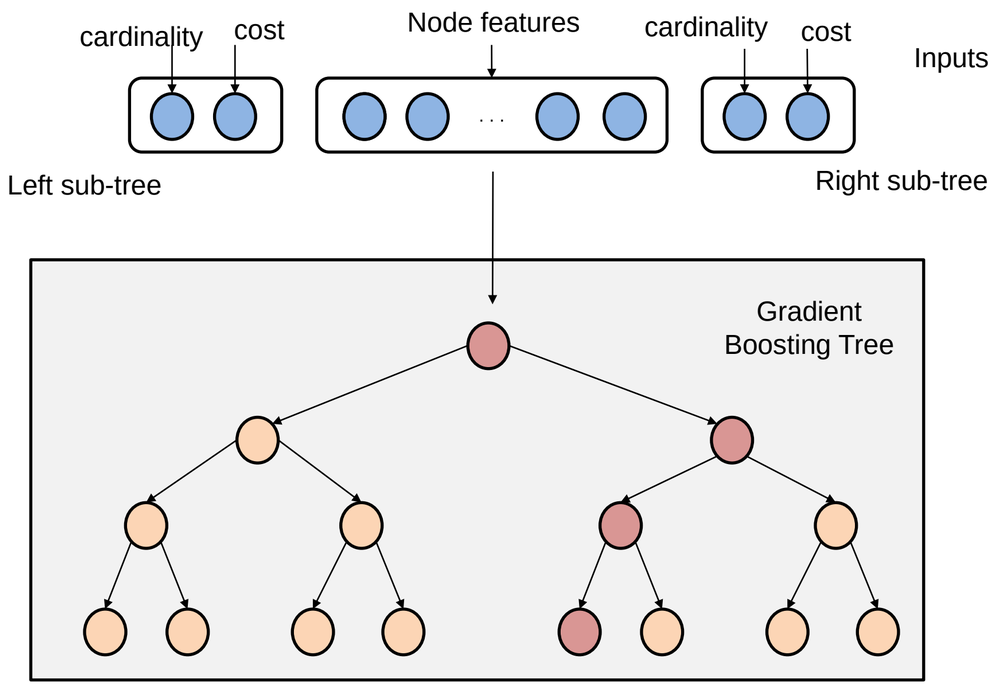
using Deep Learning Techniques

Student: Atul Acharya **Supervisor: Asst Prof Luo Siqiang**

Objectives

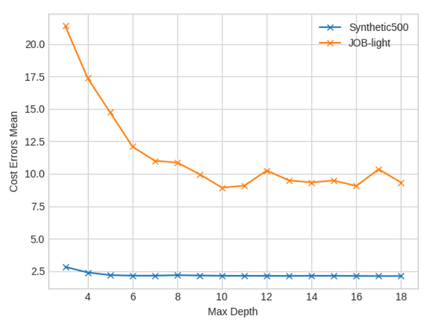
In this project, we explored existing Deep Learning Techniques for cost estimation in the query optimizer component of the DBMS. However, most algorithms cannot match the inference speed of Postgres and require a lot of training data which is very expensive to generate. Hence, the objective of the project was to develop a **novel algorithm** using machine learning and deep learning techniques that can **achieve better accuracies** than the SOTA methods while achieving very **fast inference times** and **utilizing less training data**.

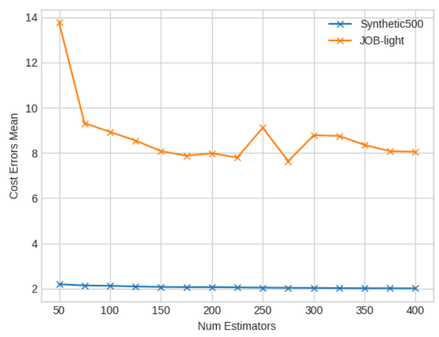
Training Stage

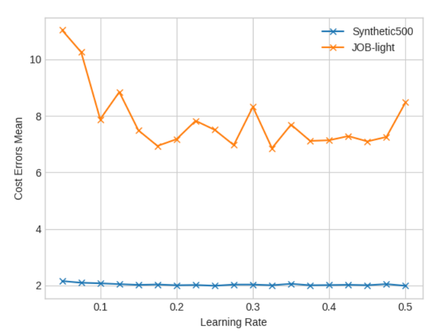


Feature	Encoding Technique
Operation	One-hot encoding
Meta-data	One-hot encoding
Data Distribution	Sample bit map vector
Predicate Tree	Tree Pooling

Controlled Hyperparameter Tuning







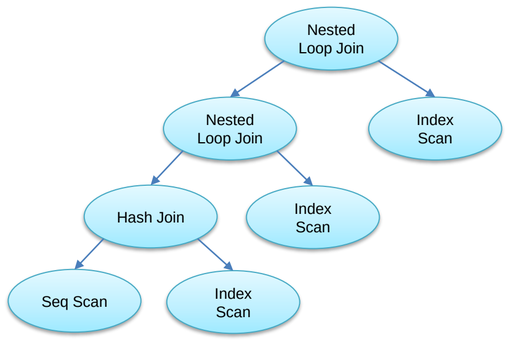
Inference Stage

SQL QUERY

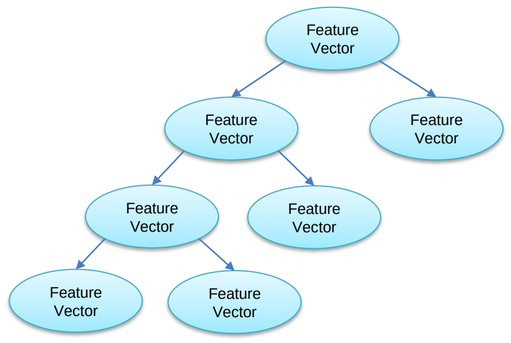
```
SELECT * FROM title t, movie_info mi,
movie_info_idx mi_idx, movie_keyword
mk

WHERE t.id=mi.movie_id AND
t.id=mk.movie_id AND
t.id=mi_idx.movie_id AND
t.production_year>2008 AND
mi.info_type_id=8 AND
mi_idx.info_type_id=101;
```

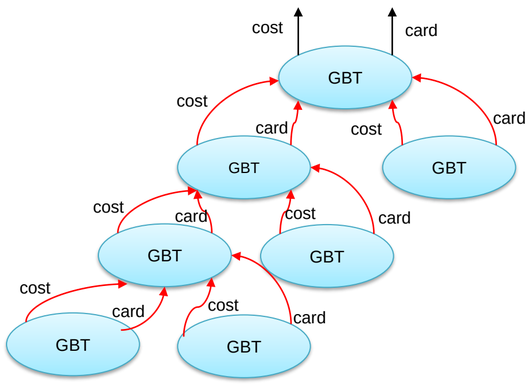
QUERY EXECUTION PLAN



QUERY PLAN VECTORS



MODEL INFERENCE



Results

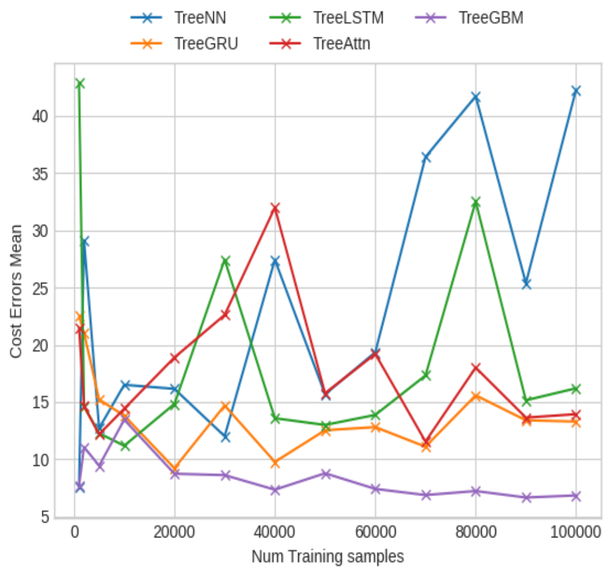
Cost Q-errors

Estimator	mean	median	90th	95th	99th	max	MAE
TreeNN	42.27	6.43	47.25	90.2	886.35	1219.16	60766
TreeGRU	13.3	2.57	20.7	36.32	189.38	341.25	57886
TreeLSTM	16.2	3.12	18.0	59.85	227.47	396.91	59088
TreeAttn	13.94	2.32	34.66	51.68	194.07	271.09	57809
TreeGBM	6.86	2.06	10.6	15.49	88.48	112.71	57634

Estimator	mean	median	90th	95th	99th	max	MAE
TreeNN	3.51	1.88	7.58	14.46	19.54	48.26	2624
TreeGRU	2.33	1.35	5.13	7.26	12.38	37.28	1640
TreeLSTM	2.67	1.41	5.35	11.58	17.16	33.93	1499
TreeAttn	3.32	1.48	5.67	14.37	29.21	36.42	1752
TreeGBM	2.0	1.4	2.97	4.34	13.03	21.02	1795

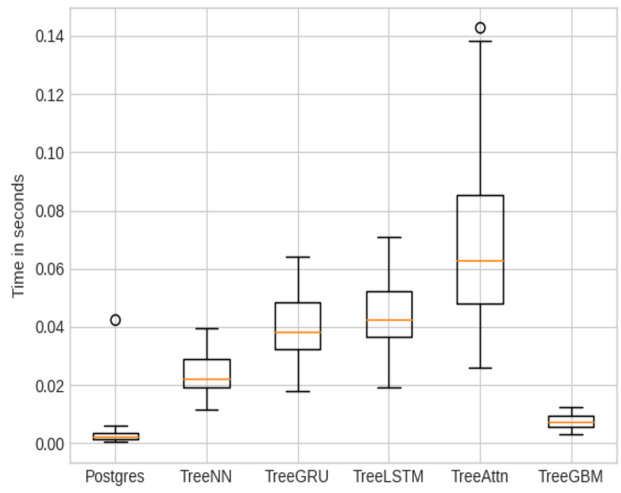
Achieves lower q-errors than SOTA

Sensitivity Analysis



Stable training with less training data

Inference Times



Faster Inference time than SOTA