NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES

Data Science and Information Technologies

SPECIALIZATION

Big Data and Artificial Intelligence

MASTER'S THESIS

QPSeeker - An efficient Neural Planner combining both data and queries through Variational Inference

Christos T. Tsapelas

ATHENS

June 2023

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Επιστήμη Δεδομένων και Τεχνολογίες Πληροφορίας

ΕΙΔΙΚΕΥΣΗ

Μεγάλα Δεδομένα και Τεχνητή Νοημοσύνη

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

QPSeeker - Νευρωνικός Βελτιστοποιητής, βασισμένος στα ερωτήματα και στα δεδομένα, χρησιμοποιώντας συμπερασματολογία μέσω μεταβλητών

Χρήστος Θ. Τσαπέλας

ΑΘΗΝΑ

Ιούνιος 2023

Master's Thesis

QPSeeker - An efficient Neural Planner combining both data and queries through Variational Inference

Christos T. Tsapelas
S.N.: DS1200018

June 2023

Διπλωματική Εργασία

QPSeeker - Νευρωνικός Βελτιστοποιητής, βασισμένος στα ερωτήματα και στα δεδομένα, χρησιμοποιώντας συμπερασματολογία μέσω μεταβλητών

Χρήστος Θ. Τσαπέλας
Α.Μ.: DS1200018

Ιούνιος 2023

ABSTRACT

Query optimization is a well-studied problem in the database community. Recently, deep learning methods have been applied either to assist the query optimizer on measures like cardinality estimation, computational cost prediction and query execution time prediction or implementing neural-based optimizers from scratch. Despite the promising results, very few tackle multiple aspects of the optimizer at the same time or combine both the underlying data and a query workload. QPSeeker takes a step towards a neural database planner, encoding first the information provided from the workload and second the underlying tables, using the power of special-designed language models for tabular data. Next, it applies a special form of attention to combine these two sources to approximate the distributions of cardinalities, costs and execution times of possible query plans. At inference, when a query is submitted to the database, QPSeeker uses its learned cost model and traverses the query plan space using Monte Carlo Tree Search to provide the best execution plan for the query.

SUBJECT AREA: Query Optimization

KEYWORDS: machine learning, multi-modal cross-attention, variational inference, cardinality estimation, cost estimation, query latency prediction

ΠΕΡΙΛΗΨΗ

Η βελτιστοποίηση ερωτημάτων έχει μελετηθεί εκτενώς από την κοινότητα των βάσεων δεδομένων. Πρόσφατα, μέθοδοι βαθειάς μηχανικής μάθησης έχουν εφαρμοστεί με τέτοιο τρόπο ώστε είτε να υποβοηθήσουν το βελτιστοποιητή του συστήματος βάσης δεδομένων στον υπολογισμό ποσοτήτων όπως, εκτίμηση πληθικότητας, πρόβλεψη υπολογιστικού κόστους και πρόβλεψη χρόνου εκτέλεσης ενός ερωτήματος, είτε στην κατασκευή βελτιστοποιητών με τη χρήση νευρωνικών δικτύων από το μηδέν. Παρά τα υποσχόμενα αποτελέσματα, λίγες μέθοδοι αντιμετωπίζουν όλες τις προκλήσεις που αντιμετωπίζει ένας βελτιστοποιητής στην πράξη ή συνδυαζούν και τα δεδομένα του συστήματος βάσης δεδομένων και τα ίδια τα ερωτήματα. Ο QPSeeker κάνει ένα βήμα προς την κατεύθυνση των βελτιστοποιητών βασισμένων σε αρχιτεκτονικές νευρωνικών δικτύων, κωδικοποιώντας αρχικά την πληροφορία που παρέχεται από το σύνολο των ερωτημάτων προς εκτέλεση και για την κωδικοποίηση των πινάκων της υπάρχουσας βάσης, όπως επίσης χρησιμοποιεί τη δύναμη ειδικών προεκπαιδευμένων γλωσσικών μοντέλων, ειδικά σχεδιασμένων για τον χειρισμό πινακοειδών δεδομένων. Στη συνέχεια, εφαρμόζει και υπολογίζει μια ειδική μορφή του μηχανισμού προσοχής, προκειμένου να συνδυάσει τις δύο πηγές εισόδου με σκοπό να κάνει εκτίμηση των κατανομών των πληθικωτήτων, κόστους και χρόνου εκτέλεσης των δυνατών πλάνων εκτέλεσης. Κατά την επεξεργασία ενός νέου ερωτήματος στο σύστημα της βάσης, ο QPSeeker, εξερευνά το χώρο καταστάσεων όλων των πιθανών πλάνων εκτέλεσης με τον αλγόριθμο Monte Carlo Tree Search κάνοντας χρήση του μοντέλου κόστους που διαμορφώθηκε κατά την εκπαίδευσή του, προκειμένου να παρέχει το καλύτερο πλάνο εκτέλεσης για το ερώτημα αυτό.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βελτιστοποίηση Ερωτημάτων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μηχανική μάθηση, πολυτροπική προσοχή, συμπερασματολογια μέσω μεταβλητών, εκτίμηση πληθικότητας, εκτίμηση υπολογιστικού κόστους, εκτίμηση χρόνου εκτέλεσης ερωτήματος

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to Dr. Georgia Koutrika, Research Director at Athena Research Center, for the excellent guidance and persistent support. I would also like to thank my family and friends for the love and support they have provided me throughout this work.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Cost-based query optimization is the process where a database system determines the optimal execution plan for a query. Cardinality estimation, join order selection and computational cost estimation of a (sub)plan highly affect the decisions of the planner during the construction of the execution plan. Even though most databases use hand-crafted heuristics, which encompass many years of research, they do not scale well to modern analytical workloads. Towards this direction, recent efforts have turned their attention to deep neural networks and aim at substituting traditional components of the planner with neural approximators [13, 19, 25]. Despite the promising results of previous efforts, we observe three aspects of the optimization process that most state-of-the-art methods do not tackle in their entirety:

– *Optimization based either only on data or queries*. Most approaches address query optimization from a workload-driven point of view [9,19,36]. There are also efforts focusing on data distribution approximation, mostly for the cardinality/selectivity estimation problem [5, 25], taking into account only the underlying data. We observe that a traditional query optimizer calculates and internally stores statistics regarding the underlying data, which are used for the cost estimation of a plan operator, while it uses the information provided from the workload for query caching or optimization of similar queries, in other words it leverages information from both data and the queries. We believe a ML-based optimizer should follow this approach too.

– *Optimization of a single task during query planning*. When a query is posed to the database system, a traditional query optimizer must estimate the selectivities of the query filters, the cardinalities of join operators, and form an optimal join ordering for the query. Very few ML-based methods tackle the aforementioned set of tasks at once [6], while most focus either only on cardinality estimation [27], query latency estimation [36] or join order selection [34] by trying either to approximate the data distributions or come up with rich query representations. All these tasks are mutually dependent, hence optimizing only for one makes the process inefficient.

– *Training on only one plan of the query plan space per query provided by the DB optimizer.* The proposed methods tackling the join ordering problem or taking into consideration the execution plan of the query for a particular task, rely heavily on the execution plan provided by the database optimizer and do not traverse the query plan space at all. This has as a consequence the biases being present in the optimizer's logic to form a biased dataset and get transferred into the model's weights. As shown in [4], a neural network has the tendency to amplify the biases present in the training set. In query optimization, and especially in production environments, this side effect can have catastrophic results.

Motivated from the above limitations, we propose *QPSeeker* (Query Plan Seeker), a novel *end-to-end neural database planner* that (*a*) simultaneously learns to perform all basic tasks of a traditional optimizer, such as join order selection, cardinality/selectivity estimation and execution time prediction, (*b*) leverages queries and data for training and inference purposes, (*c*) samples the query space of each training query to generate an enriched training set, and (*d*) uses its rich learned model for query planning.

1.1 Overview.

In QPSeeker's core is a model that learns to *approximate the distributions of the cardinality, computational cost and runtime* of the plans in the workload. Our approach assumes that queries with *similar characteristics* (e.g., number of tables, number of joins, filters applied, etc.) and *complexity* in terms of execution time will be close to each other in a latent space, and we *use variational inference* to learn this space. Hence, at the heart of our system lies a Variational Autoencoder (VAE) [12], whose latent space is enforced to follow a Gaussian structure, where each latent dimension represents a latent feature of the data. At the end of training, similar queries and, particularly, similar query execution plans will fall close to each other in the learned latent space.

To jointly learn from both data and queries, information about data and data distributions is used for training, along with features extracted from the query. We address several challenges. One challenge is to capture the data distributions from the table data and provide a rich representation for further processing. Moreover, the query/plan representations inside the system play an important role in order to give the model the ability to capture the correlations between the query, its complexity, and the data. Furthermore, one important question is how we associate the query and the execution plan we are investigating. For this purpose, our approach comprises the following novel components.

First, for the representation of table data in the model, we choose TaBERT [33], a language model for tabular data. TaBERT is trained on millions of tables from the WDC corpus [16] and learns much richer data representations than creating database table embeddings from scratch. Moreover, the pretraining tasks applied to TaBERT, i.e., the datatype and cell value prediction, help TaBERT to learn information about table data distributions.

Second, QPSeeker extracts the sets of relations, joins and predicates from the query (as in [13]), and subsequently *learns the mapping between these three sets and the query plan statistics*. In this way, its model has the ability to capture the distributions of various instances of the above sets (in terms of which elements are present) paired with the particular physical operators in the query plan.

Third, QPSeeker employs a *rich, tree-like, query plan representation* that: (*a*) captures *data distributions using TaBERT*; (*b*) encodes each node in the plan using *information about this node* (including the relations involved, the physical operation, and the contextual representation of the table data) *as well as the impact of the previous operations* in the subplan; and (*c*) computes an embedding vector that contains the prediction of the values for the cardinality, cost, and runtime for each node, as well as a data vector that captures the impact of the node's children.

Fourth, we observe that each plan node does not have the same impact on the final runtime of the query and its computational cost differs from the cost of the other nodes in the plan. For example, an early decision of the planner for an Index Scan over a table, which may seem promising at early stages, may lead to bad paths (join orderings). Therefore, QPSeeker associates a query and a plan by *considering the impact of each plan node on the query estimations through a cross-attention mechanism*. In particular, we apply attention between the query embedding vector and the embedding vector of each node in the plan, in order to score which nodes have the most impact on the final estimations.

For training, we generate sample plans for each query instead of relying on a single, 'best', plan provided by the DB optimizer. In this way, we "mimick" a traditional optimizer that traverses the plan space for a given query and estimates the query execution cost and runtime along with the cardinalities of the intermediate results. For each query, we create samples from the plan space, by considering different join orderings and different methods for the operators of the query. The rationale for sampling the plan space is that the DB optimizer relies on internal statistics and formulas to make its estimations and come up with the best plan. However, if we just rely on this plan to train our optimizer, we will not be able to acquire such broad knowledge. Furthermore, we have the choice either to user the internal cost model of the DBMS or to use a user-defined one to generate the training data. Using sampling coupled with variational inference allows us to train our model not to directly learn a mapping of the workload to the target values, but to approximate the distributions of the cardinality, computational cost and runtime of the execution plans per query in the workload.

At inference, we use Monte Carlo Tree Search [14] along with our learned cost model, which combines information from both the data and the query, to traverse the query plan space.

1.2 Contributions.

The contributions of this thesis are:

- We introduce QPSeeker, a novel neural planner that simultaneously learns to perform all basic tasks of a traditional optimizer, such as join ordering, cardinality/selectivity estimation and execution time prediction.
- We cast our learning problem to a variational inference problem. We train our VAE-based model to approximate the distributions of the cardinality, computational cost and runtimes of the execution plans. Our model can capture hidden commonalities between the queries and the data into a latent space.
- We leverage both data and queries. We employ a rich query plan representation that captures the correlations between the query, its complexity, and the data. For the representation of table data, we choose TaBERT that captures the data distributions and provides a rich data representation.
- We calculate the impact each plan node has on the query's estimations through an attention mechanism.
- For training, we generate sample plans from the query space of each training query to generate an enriched training set. In this way, we train our model not to directly learn a mapping of a workload to the target values, but to approximate the distributions of the cardinality, computational cost and runtime of the execution plans per query in the workload.
- At inference, we use the learnt model and Monte Carlo Tree Search for query planning.
- We present our detailed experimental results. Our experiments show that QPSeeker achieves being an all-in-one planner that performs all tasks of a query optimizer in

an effective way outperforming competitors. Especially, for complex queries, it outperforms PostgreSQL. Furthermore, it learns better using complex workloads, and it shows excellent adaptability to different workloads, where competitors cannot cope.

2. RELATED WORK

In the last years, there has been significant efforts into the integration of machine learning models into query optimizers.

2.1 Learned Cardinality Estimation

For the regression problems of cardinality and selectivity estimation, many (un)supervised methods have been proposed. MSCN [13] is a supervised method that uses set extraction of the basic elements (relations, joins and predicates) from each query. Following the same rationale, a new heuristic metric called *Plan-Error* is proposed for cost-guided cardinality estimation [26]. *These approaches neglect the presence of the underlying data and their effect on query performance.*

There are approaches that try to capture the underlying data. Flow-Loss [25] defines another metric, where the query plan is formulated as an electric circuit and the model estimates the cheapest path. DQM [5] faces the cardinality estimation task as both (un-)supervised problem, by estimating distribution densities. Naru [32] and its predecessor, UAE [29], use autoregressive models to approximate joint distributions over the database tables. NeuroCard [31] estimated the cardinalities over extracted samples from full outer joins of the database tables. DeepDB [8] introduced relational sum product networks, which are tree-structured to capture the data distributions using several local PDFs, and as we go up the tree, each node stores cumulative PDFs. *These approaches can approximate quite accurately the table distributions for a small number of joins, but they do not scale well for many joins.* FLAT [38] uses another type of network, called factorize-sum-split-product network (FSPN), to capture the underlying data calculating the level of dependency of column via conditional factorization. Finally, Fauce [18] introduced a model incorporating uncertainty in its predictions.

2.2 Learning Join Orders

Another line of research has used reinforcement learning (RL) to find plans with low cost (e.g., [15], ReJOIN [21]). Neo [20] proposes a learnable query optimizer which incrementally searches and builds the physical query plan. Despite the significant training time, the produced query plans were very competitive compared to the plans of a commercial optimizer. Neo also introduced the formulation of the query plan as a Tree-Convolution, also used in Bao [19]. Bao uses RL to learn hints at the plan operator level to advise the Postgres query optimizer. Bao's approach was adopted to shrink the very large search space of the SCOPE optimizer [2] to make it work in a cloud environment [24]. RTOS [35] introduced a Tree-LSTM structure used with Q-Learning to tackle the join order selection task, but also needed large number of episodes to achieve comparable results. Balsa [30] used RL to produce query plans by applying query plans to Postgres and evaluating its choices by trial and error.

2.3 Learning Cost Estimation

Plan cost estimation is a critical task of query optimization. E2E-Cost [27] and QPPNet [22] featurize the physical query plan as a tree and propose to train a regression model to predict the cost of a physical plan. E2E-Cost mimics the tree-structure of the query plan into a Tree-LSTM model. Moreover, it introduced a new approach to create a neural representation keeping the logical semantics of a predicate as well as a new method to create embeddings for string values. QPPNet also follows the tree structure of the plan, associating each plan operator to a small MLP. It proposed a plan-structured deep neural network, i.e., a neural network model specifically designed to predict the latency of query execution plans by dynamically assembling neural units in a network isomorphic to a given query plan. Zero-Shot [6] aims at generalizing learned cost estimation to unseen databases. In contrast to workload-driven approaches, zero-shot cost models suggests a new learning paradigm based on pre-trained cost models.

2.4 Comparison

QPSeeker is an end-to-end neural-based database planner that can *perform all tasks of a traditional optimizer*, i.e., join order selection, cardinality/selectivity estimation, cost and execution time prediction (while most approaches focus on a single task). Furthermore, existing approaches are haunted by complex designs and significant training times. For example, Neo [20] reported 24h for training, while QPPNet [22] used a network of 8 layers, each additional hidden layer adding on the order of 2^{14} additional weights, that did not converge until epoch 1000 (28 hours). Thanks to leveraging Variarional Inference [1] boosted by a language model (TaBERT), our QPSeeker is *considerably leaner, just 14.4M parameters in total*, and *can be trained in a short time*, (less than 1h) as we will see in the experiments, making it the first viable solution that brings deep learning inside the query optimizer.

Furthermore, most approaches do not leverage both the queries and the data. E2ECost [27] includes a small sample from each table in the encoding similar to [13]. QPSeeker *combines information from the queries and the data*. It employs a *rich query plan representation* approach that (a) *captures data distributions using TaBERT* [33], a language model suited for tabular data, and (b) uses *attention to weigh in the impact of each query plan node* on the query estimations. Note that recent works have focused on query plan representations. For example, QueryFormer [37] proposes a different scheme based on Transformers that integrates histograms obtained from database systems into query plan encoding.

3. THE PROPOSED FRAMEWORK

3.1 Problem Statement

Given a query, the goal of the planner is to come up with a good execution plan. An execution plan is represented as a tree whose internal nodes are operators and its leaves correspond to the input tables of the query. During query planning, the accurate estimations of cardinalities, costs, and runtimes of the execution plan nodes give the ability to the optimizer to build good plans.

We consider a workload W of query and execution plan pairs $(QEPs)$, where each QEP is characterized by its cardinality, computational cost, and runtime. We aim to construct a model, which associates the table data with the physical operations in the plan to predict the resulting cardinalities, computational costs and runtimes for W, and further to approximate the data distributions and the cardinality, computational cost and runtime distributions from the available workload, and make predictions for unseen queries. During inference, when a query is posed to the database system, QPSeeker uses the trained model and traverses the query plan space to evaluate candidate plans and suggest the one to be executed.

3.2 QPSeeker Pipeline

We provide an overview of QPSeeker's pipeline (Figure 1). A query, execution plan $QEPs$, and the data are the input. The *Parser* parses the query and extracts the relations, the joins and the predicates in the WHERE clause. The relations and joins are one-hot encoded based on the database schema and these encodings are passed to the *Query Encoder* to build the query embedding vector.

Based on the relations, joins and filters of the query, the *Plan Encoder* encodes each physical operator in the query plan and computes the values of the plan in a bottom-up fashion. Each plan node takes as input (*a*) the sum of one-hot encodings of the relations being present at each level of the subplan, (*b*) the physical operation applied also in one-hot encoding, and (*c*) the contextual representation of the table data extracted from TaBERT. The output of each node is an embedding vector, where the last dimensions are the estimations about the cardinality, cost and latency of the plan at this level. The root

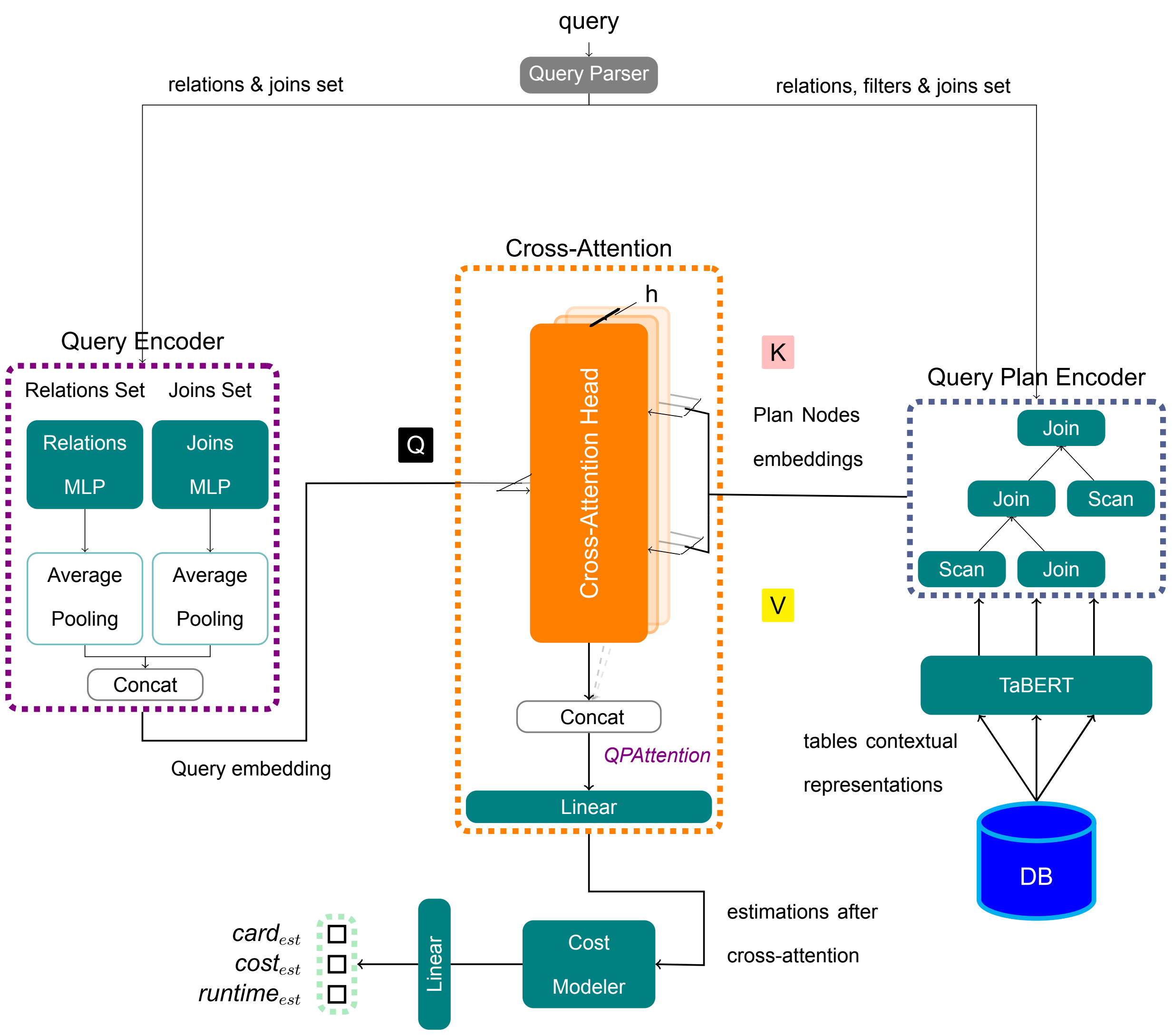

Figure 1: QPSeeker's architecture. With teal are coloured all the neural-based modules of QPSeeker. The relations and joins present in the query are passed through their corresponding MLPs forming the query embedding. TaBERT provides the encodings for base relations, serving as input to Plan Encoder's leaves. Each node in the plan is encoded by an LSTM cell. The output of each plan node is stacked and *QPAttention* is calculated scoring which nodes have the most impact on the query. Finally, the Cost Modeler (VAE) estimates the cadinality, cost and runtime of the (query, plan) pair.

node holds these values for the entire plan.

Next, we combine the outputs of the Query Encoder and Plan Encoder, i.e., the query and the plan embedding vectors, using attention (*QPAttention*) to score which nodes of the plan affect the most the given query, followed by a dense layer, with output size equal to the sum of the query and plan embedding vectors. Finally, the result of the attention mechanism is fed to the cost modeler, a Variational Auto-Encoder that makes the predictions for the given query as an output vector, similar to the Plan Encoder's output, with the last three dimensions being the estimates for the cardinality, cost and latency of the query plan for the given query.

During the training phase, we follow the encoding process discussed above for the table data, queries and execution plans, and we train QPSeeker using the $QEPs$ in a workload W. We use two different schemes regarding the derivation process of the execution plans: (*a*) one plan per query, suggested by the DB optimizer, and (*b*) sampling plans from the plan space of a query. (More details for each method in Section 5.)

On inference, when a query is posed to the system, QPSeeker uses its learned model to traverse the query plan space, using Monte Carlo Tree Search (MCTS). Initially, we construct the join graph from the query and then, the algorithm constructs candidate execution plans starting from base relations and until all joins are applied. At each step, a plan operation is chosen to be applied and based on the decision, MCTS randomly generates the next plan operators, until the plan is completed. We evaluate the simulated plan with QPSeeker's model and update the rewards of each node in the plan, if is the best plan found so far. More details in Section 5.3.

4. QEPS ENCODING

In this section, we describe the encoding for a QEP. We begin with the query encoding followed by the encoding of the execution plan. Given the two encodings, we describe how we associate them through *QPAttention*. Finally, we provide an illustration of our variational inference model, serving as the basis of our cost modeler, to approximate the statistics for the set of *QEPs* in the workload W and its utilization through a variational autoencoder.

Query Parsing. From each query, we extract three sets: (*a*) the set T_{q} of query relations, (*b*) the set J_{q} of joins, and (*c*) the set P_{q} of conditions over the database relations. The sets T_{q} and J_{q} are passed to the query encoder, while T_{q}, J_{q} and P_{q} are passed to the query plan encoder for further processing.

4.1 Query Encoding

The query encoder is responsible for providing a rich representation of the query. Its output will be used to compute the association between the query and the execution plan.

We follow the feature extraction process described in [13]. Each relation in T_{q} is mapped to a one-hot vector of size N, where N is the number of database relations. Similarly, each join in J_{q} is transformed into a one-hot vector with length M equal to the number of all possible joins in the database. Subsequently, we transform each set of vectors into a fixed-size input for further processing. We map T_{q} and J_{q} to two fixed size arrays, NxN and MxM, respectively. The NxN (MxM resp.) array contains all the one-hot vectors for T_{q} (J_{q}, resp.) at the first rows and the rest are all zeros.

We feed each matrix, along with a column vector that serves as mask to filter the non-zero rows, into a feed forward network with five hidden layers (i.e., a Multi-layer Perceptron, MLP). Finally, the encoder applies mean pooling among the elements of each set to derive one representation for each set and concatenates the two representations to form the query embedding vector.

4.2 Query Plan Encoding

The goal of the plan encoder is to capture the result of the interactions of the physical operators over the database tables learning from the operators and the structure of each query plan. In this way, the model can learn, for example, the cost of applying a Hash Join over two tables, where the outer table is accessed via an index.

Generally, the performance of each plan operator is highly correlated with that of its children in the execution plan. During the flow of computation performed inside the plan encoder, we wish to capture this interaction between the nodes at the operator level. Hence, at each node, we compute an embedding vector that contains the prediction of the values for the cardinality, cost, and runtime for this node, as well as the interaction between the nodes. To capture the correlations between the nodes, we need somehow to inform the parent about the output of its children. Hence, apart from the current state of the subplan, which will be discussed shortly in node input, each node in the plan encoder passes its output to its parent.

Figure 2 depicts our plan tree encoding. *Plan Encoder* assembles the plan operators in a tree, having the same tree structure as the execution plan provided by the optimizer.

Node Encoding. A query plan consists of two types of nodes: (*a*) the leaf nodes that correspond to the scan operations over the base tables of the database, and (*b*) the intermediate nodes that correspond to the join operations. In our configuration, each plan node is modeled as an LSTM cell [10]. Similar to the query plans produced by a database system, where each node in a plan is affected only by its children, the input of each LSTM cell can come only from its children. Additionally, the architecture of the LSTM cell suits very well the query plan encoding process, as it can capture useful information over long sequences (its inputs) and decide which information from its ancestors is useful and which not.

Node Output. Each node of the plan outputs a vector (of size 1500) that contains useful information about the interactions of the operators in the query plan. An example vector is seen at the output of the root node in Figure 2. In our formulation, we force from the Variational AutoEncoder output the last three dimensions of this vector to be the estimations of the cardinality, cost, latency of the node in the plan. The remaining dimensions comprise a data vector that captures the interactions between the nodes of the (sub-)plan under this node. As estimates for the whole query plan, we consider the output of the root

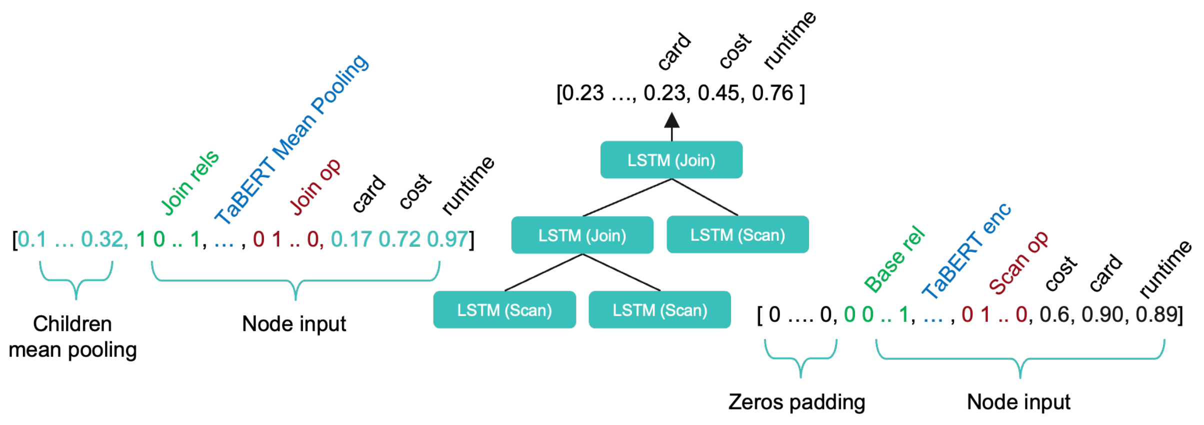

Figure 2: Plan Tree Encoding

node of the plan.

Node Input. The input of a node is a fixed size (2048) vector that combines different types of information. On the right side of Figure 2, we see the input of a leaf node, and on the left side, we see the input of an intermediate node. They essentially share a similar structure but they also have some differences as we explain below.

The input of a leaf is the concatenation of the following vectors (looking at the example vector from right to left):

a. Estimations for the cardinality, cost and runtime for the operation of this node. For a given query plan, we use EXPLAIN to get this information from the DB optimizer.

b. The physical operator applied to this node in one-hot encoding.

c. The representation of the data processed. If there is a filter in the set P_q of predicates over a column of the table, we take the representation of this column filtered based on the query predicate, otherwise the table representation. In both cases, the representation is provided by TaBERT, as we explain below.

d. The table accessed in this leaf node in one-hot encoding.

e. Zeros for padding. The input of each node consists of two parts. One part comes from its children nodes and one concerns the operation of the node per se. Since leaf nodes do not have children, they only encode information regarding the node and the first part is padded with zeros to tell the plan encoder that there are no children for this operator of the plan, hence there is no information from a predecessor node to affect the node.

The input of a non-leaf node is the concatenation of the:

a. Estimations for the cardinality, cost and runtime for the operation of this node. These are computed by mean pooling the last three dimensions of the output vectors of the node's children.

b. The physical operator applied to the node in one-hot encoding.

c. The representation of the data processed, which comes from the result of mean pooling over the output from the [CLS] token of each joined relation. This token has a special functionality as it holds information over the entire table. More details about the [CLS] token are provided below.

d. The relation encoding is the sum of one-hot vectors of all relations joined up to this level of the plan. Providing this encoding to the LSTM cell, we inform the plan which relationships are present in the subplan and which are not.

e. The information about the interaction between the children and parent. Instead of zeros in leaf nodes indicating the absence of an ancestor, we desire that features from children nodes are passed up the tree. Hence, we provide also the result of mean pooling from the data vectors of the node's children.

4.2.1 TaBERT - Table Data Representation

While the query and the plan representations are crucial, the representation of the table data and their distributions are also very important. One approach would be to create embedding vectors from scratch for each database like Neo [20] and TLSTM [27], but such a strategy has limitations if the table data changes, because the model has to be retrained again. To override the above restrictions, we reap the benefits of transfer learning properties found in large pretrained language models. TaBERT [33] is a special case of BERT [3] for tabular data, and provides much richer and robust tabular data representation, unbounded from the strict assumptions regarding their datatypes and their prior distributions as in a RDBMS.

We use TaBERT as follows: for each QEP in the workload, it tokenizes the columns of the table, and for each column, it creates (name, datatype, value) triplets separated by a special symbol. Each value is extracted from the top-*K* rows of the table with the biggest n-gram overlap with the query. Then, these triplets are concatenated with the query and served as input to a BERT model. After the initial BERT encoding, TaBERT needs to gather the information from all row-level encodings into one vector containing an output for each column. Consequently, it calculates cell-wise attention over all rows, called vertical attention. Each cell contains the output of TaBERT for each column in the table. Finally, as in language modeling, where the [CLS] special token at the start of each sentence holds information about the whole sentence, similarly this token in the output of TaBERT holds information for the whole table.

TaBERT is trained on *Masked Column Prediction (MCP)* and *Cell Value Recovery (CVR)* objectives. The former encourages the model to recover the name and the datatype of the masked column from its contexts, hence learning, in this way, the correlation between the masked column and the other columns in the table. With this task, we can pass to our plan encoder information about the datatype of a column. The latter objective encourages the model to predict the values of the masked columns. More precisely, after column masking and the extraction of top-*K* rows from the table, TaBERT is tasked to predict the values of the masked cells. In this way, TaBERT captures information about the column distribution, along with its context within the rows. The developers provide three different models for $K = [1, 2, 3]$. With the use of TaBERT, this information is also inferred in QPSeeker's plan encoder.

Hence, for each condition in P_{q} and relation present in the query, we use the latent representation extracted from TaBERT by passing the query and the corresponding table where the condition applies to. We extract the representation of the respective column in the condition and the table representation to be used in the inputs of the plan encoder as described earlier. The table representation is extracted for all tables in the query, through the [CLS] token.

4.3 Attending the queries to the query plans

When the plan encoding phase is finished, QPSeeker combines the query embedding vector and the plan embedding vector into one embedding vector, as shown in Figure 3. However, the simple approach of concatenating these two vectors into one common vector does not have any semantic value, as they represent two different sources, i.e., the query and the query plan along with table data. Instead, we apply cross-attention inspired by the Perceiver architecture [11]. Furthermore, we observe that each node does not have the same impact on the plan in terms of the execution time and computational cost for the complete plan. For example, the selection of Sequential Scan instead of the use

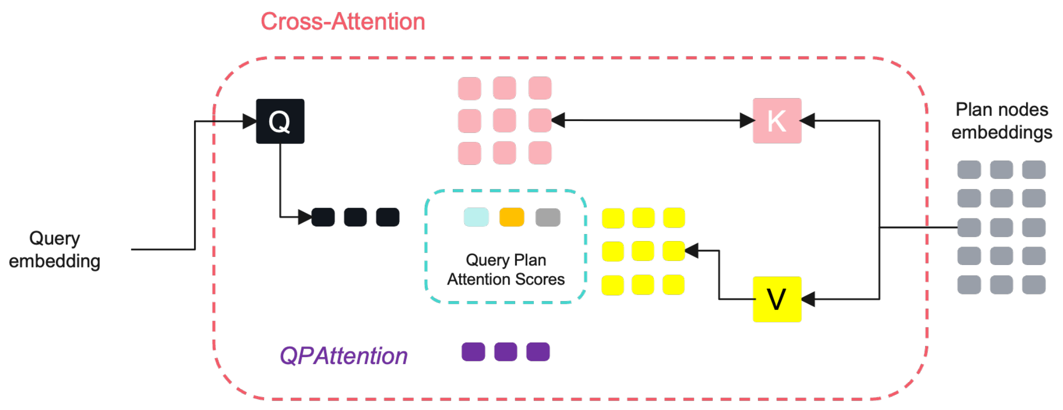

Figure 3: Attention between the query embedding and the plan nodes' embeddings.

of an index over a large table with a high selective filter affects more the final execution time of the plan. Or the selection of an operator requiring more memory and hence more computational cost, like a Hash Join, will have a higher value for its cost, than an Index Scan. Therefore, we desire to give a score to each plan node and measure which nodes in the plan have the higher impact on the final estimations. To this direction, we make use of cross-attention between the query and the output of each node in the plan.

4.4 Cost Modeler

So far, we have encoded the query, the table data associated with the query plan, and for each QEP, we have calculated how the plan is associated with the query by weighing in the impact of each plan node on the estimations for the query through an attention mechanism. However, for each query, the space of possible plans is huge, and each plan has different execution time and computational cost. Our goal is to capture the distributions of the cardinalities, costs and execution times for the plans in the space of a query, and to be able to generalise for the entire workload.

For this purpose, at the heart of QPSeeker lies a variational autoencoder, acting as the cost modeler. The objective of the cost modeler is not only to approximate the target distributions but also to be able to generalise on unseen queries, by providing accurate estimates for each plan node statistics, and consequently, for a whole execution plan suggested by QPSeeker, through variational inference [1]. Our belief is that execution plans with analogous complexity, in terms of runtime and execution cost, or with similar characteristics, such as relations and filters applied, if projected to a structured latent space, will have representations close to each other, depicting these similarities. The use

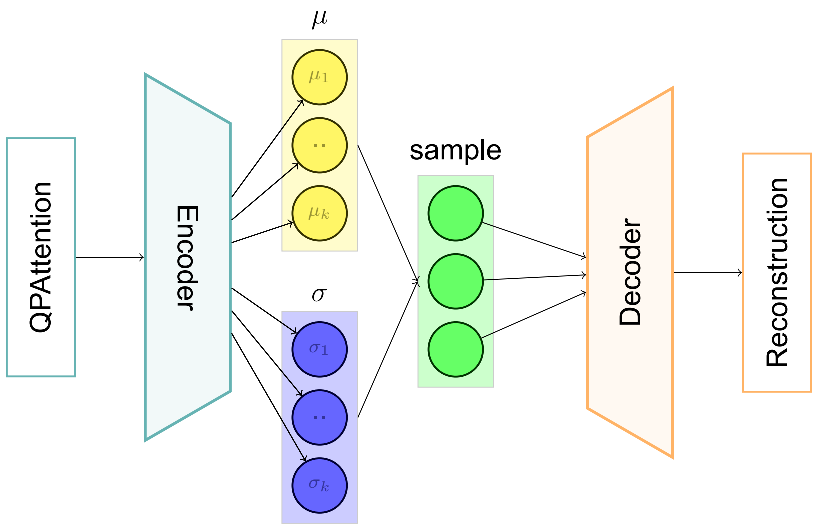

Figure 4: VAE's architecture. The latent space z formulates a mixture of unit-variance Gaussians

of the VAE aims at the formulation of such a latent space, through its functionality.

More precisely, our approach for the *Cost Modeler* is based on the following framework: Given a set W of observed variables, i.e., in our case, a workload W of $QEPs$, where each QEP is characterized by its cardinality, cost, and runtime, infer a latent variable z, which generates the initial observations. The described conditional probabilty can be written as:

$$p(z|W) = \frac{p(W|z)p(z)}{p(W)}$$

where the density of workload W can be computed as:

$$p(W) = \int p(z|W)p(z)\,dz$$

As we observe, the calculation of the density function for our workload W demands the computation of $p(z)$, which we do not have access to, as it is a latent variable, hence the above integral is intractable. Despite the intractabilty of the integral, we can approximate the above value by applying variational inference [1].

In variational inference, we specify a family of densities over latent variable z, with the purpose to find the best candidate approximation to the exact conditional. Hence, let $q(z|W)$ be the approximation of the latent variable generating the values for cardinalities, costs and runtimes for the $QEPs$ in our given workload W. Since we want to find the best candidate to approximate the latent variable, the optimization problem is to minimize the error between the latent and our approximation. Since both values are density functions, our goal is to minimize the Kullback-Leibler (KL) diveregence, which can be written as

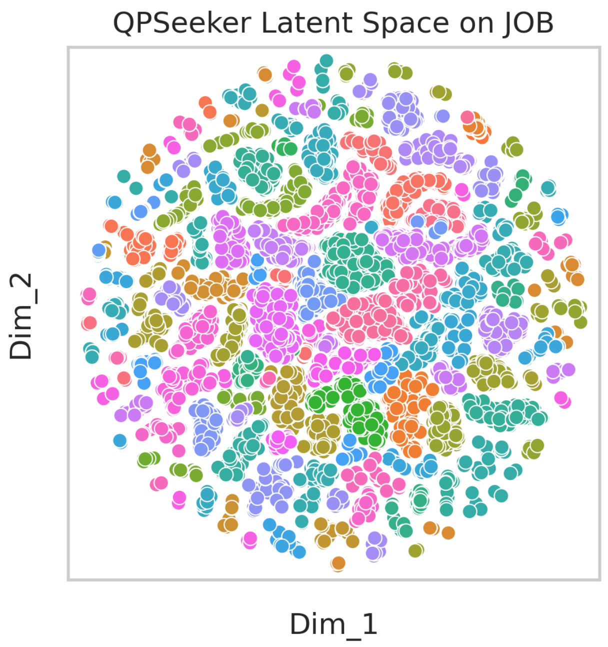

Figure 5: t-SNE projection of Cost Modeler's (VAE) latent space on JOB. The colour codes indicate query plan samples produced from the same query tempalate.

follows:

$$KL_{min}(q(z|W), p(z|W)) \tag{4.4.1}$$

All we need is to specify the form of the latent variable.

4.4.1 Forcing structure into the latent space.

VAE implements the previously described model. It consists of three parts: (*a*) the encoder, encoding the input into the latent space, (*b*) the sampler, sampling from the latent distribution, and (*c*) the decoder, which receives the sample from the latent and decodes it to the initial input.

Initially, the encoder receives the result of *QPAttention* and encodes it into a latent space, serving as the $p(W|z)$ in our framework. Then, the sampler samples a data point from this latent distribution, and finally, the decoder receives this vector from this distribution and outputs the reconstructed vector, serving as the approximation $q(z|W)$ in our framework. Finally, the reconstructed vector is passed into a linear layer, to get the estimates for a particular QEP.

As described above, in order for VAE to conform with our described framework, its latent space must describe a distribution, thus it is forced to have a structure. QPSeeker forces this structure to be a mixture of unit-variance univariate Gaussians, and the latent space represents the parameters of the distributions mixture. The first half represents the mean and the other half the variance of the latent distributions, as shown in Figure 4. Finally, during training, QPSeeker minimizes: **a)** the reconstruction loss of *QPAttetion* and the KL divergence described in equation 4.4 and **b)** the mean squared error (MSE) between the true values of QEPs and QPSeeker estimates. The loss can be written as:

$$QPSeeker_{\text{loss}} = ||x - \hat{x}||^2 + \beta * KL[N(\mu_i, \sigma_i), N(0, 1)] \tag{4.4.2}$$

where x are the estimates of QPSeeker and $\hat{x}$ the true values of a particular QEP. For $\beta > 1$, we emphasize on the KL, encouraging QPSeeker to learn broader distributions. More on the effect of β in the experiments section.

Figure 5 shows how QPSeeker has organized its latent space for QEPS produced by sampling from JOB. We used t-SNE [28] method to project the 32 latent features into 2-d plane. The color codes indicate that these QEPs have been produced from the same query template. QPSeeker has organized its latent space, not only in a way that query plan samples from the same query template are close to each other, but also samples from different queries.

5. TRAINING LOOP & INFERENCE

In this section, we describe the training loop of QPSeeker. We train under two settings: (*a*) for each query, we use the query plan provided by the DB optimizer, and (*b*) we enumerate the query plan space and extract a sample for training (Section 5.1). In all cases, the first step is the extraction of the query representation as described in Section 4.1. Then, we encode each execution plan as described in Section 4.2, and we apply the cross-attention mechanism showed in Section 4.3 to create the input for the cost modeler. For the reconstruction loss, we inject the last three dimensions of the output vector of the variatonal autoencoder with the true values extracted from the execution of the query plans in Postgres. In this way, we force the cost modeler to learn the true values of the distributions for the cardinalities, costs and latencies of the query plans.

5.1 Query Plan set selection

In this section, we describe our approach for generating training data from samples of the query plan space. In order to learn the distributions of the cardinalities, costs and execution times of the query plans set, the naive approach is to enumerate the query plan space per query and construct all possible query plans per query. As the number of relations and joins increases, the plan space is growing exponentially and the time to enumerate and execute all these plans is prohibitive. Hence, from the query graph, we enumerate all the possible join orderings. We transform each join order into the corresponding binary left-depth query plan tree, and we randomly select an operation from Postgres for each node of the plan. All leaf nodes refer to table scans and all intermediate nodes are the joins operations between two tables. For each plan we construct, we calculate their corresponding measures, by a simple yet effective user-defined cost model, then we sort them based on the cost and pick the first 15% as the query plan set for a particular query. Our cost model is defined below:

```
1.  Seq Scan = tbl_blocks / block_size +
           tbl_rows * cpu_tuple_cost

2.  Index Scan = index_height +
           index_leaf_pages / 2 * cpu_tuple_cost
```

```
3.  BitmapIndexScan = index_height *
            random_page_cost + log(tbl_blocks / block_size)

4.  Merge Join = |relA| * log(|relA|) +
          |relB| * log(|relB|) + |relA| + |relB|

5.  HashJoin = |relA| + 2 * |relB|

6.  NestedLoops = |relA| * relB_blocks + relB_blocks
```

Finally, all produced query plans are submitted to Postgres for execution. In order to inject our plan in the optimizer, we use the PgCuckoo [9] extension with some modifications, which forces the optimizer to use our hand-crafted query plan at runtime. Moreover, we use the EXPLAIN ANALYZE functionality of the database system to get the statistics from the execution of our plan. In order to reduce the range of values among all the plans in the query workload, making the predictions for QPSeeker easier, we apply Min-Max scaling on the cardinalities of the queries, their execution times and the cost per physical node in the plans. We also apply the same process for the intermediate cardinalities of all subplans.

5.2 PgCuckoo - Forcing Execution Plans to Postgres

For both training and inference, QPSeeker needs to specify the query plan to be executed from Postgres' optimizer. To do so, we initially used the *pg_hint_plan* extension of Postgres, used by many similar systems like QPseeker, which need to take control of planner decisions for query execution. By specifying "hints" to the optimizer, provided as comment in the query, the user can tweak the optimizer specifying the scan and join methods, as well as the join ordering of the tables within the query.

Despite the extend use of pg_hint_plan extension by the research community, we decided that this tool does not fit our needs, as the final plan executed by Postgres does not always corresponds with the hints provided, especially during the creation of training samples from the query plan space. We observed that for the queries where the cost estimation exceeds a specific limit (2000000000 computational units) the planner ignores the given hints and continues the planning of the query using its internal functionality.

To overcome the above behavior and make sure that the execution plan will be executed

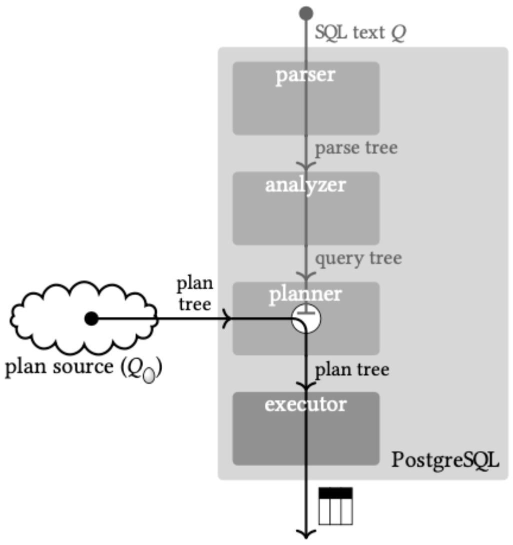

Figure 6: PgCuckoo workflow. PgCuckoo translates the query plan provided into algebraic code ready for immediate execution from the executor.

as expected, we make use of PgCuckoo [9] a Haskell based library which communicates directly with Postgres executor module for immediate execution. PgCuckoo works as an external code generator, where the external application makes use of the provided library and then its translated into algebraic code in the form of plan trees. This code is executed from the *Executor*.

In our framework, we have created a plan rewriter, where the produced plan from QPSeeker is rewritten, follwing the PgCuckoo's Haskell library format. Then, our plan is compiled into algebraic code used by Postgres executor. In this way, we are able to control the plan generation granularity at plan-operator level. This process is used for both the generation of plans through sampling using our user-defined cost model for training, as well as the plans produced during inference.

5.3 Inference - Monte Carlo Tree Search

After training the cost model of QPSeeker, the planner can be used for planning new queries. As mentioned before, as the number of relations increases, the number of possible execution plans grows exponentially making the plan space intractable. In order to traverse the search space fast and find a good execution plan, we use the Monte Carlo Tree Search (MCTS) [14] algorithm. In its basis, MCTS uses randomness to select the next plan operator using sampling, thus it can estimate a near optimal action in current state with low computation effort. Moreover, the fact that it chooses the best action based on long-term rewards, makes it very appealing for the query plan decision. We use vanilla MCTS to traverse the query plan space in a bottom-up fashion. We start from base relations and apply one join at a time until all relations are present in the final plan. As a reward function, we use the UCT formula:

$$\frac{r_i}{n_i} + C\sqrt{\frac{lnt}{n_i}} \tag{5.3.1}$$

where for the i-th node, r_i is its reward and calculates how many times the node is present in the best plan so far during the simulations. Next, n_i is the number of rollouts, t is the number of rollouts of the parent node, and *C* is the exploration coefficient parameter, ranging between [0, 1]. For the evaluation of each plan node, we use QPSeeker's internal cost model, and finally, the best plan is considered the execution plan with the least estimated execution time. The reward for each node being present in the best plan discovered so far in the simulation is one unit. For each query, we set a planning time cut-off of *200ms* and if the agent has not finished traversing the space in this time budget, we select the best plan found so far. MCTS consists of four steps:

1. *Selection.* Based on the current state of the selected subplan, the agent chooses the new plan operator in the plan with highest value, based on our policy, forming the new state of the plan.
2. *Expansion.* The agent generates all possible nodes of the query plan, based on the previously selected action in the plan.
3. *Rollout.* We start a simulation from the current state of the plan by randomly selecting the next operator to be applied, until the plan is complete.
4. *Backpropagation.* Based on the played simulation, we estimate the execution time of the simulated plan using QPSeeker's cost model and update the rewards.

6. EXPERIMENTS

In this section, we perform a brief analysis of available workloads and next, we describe our experimental setup. Initially, we evaluate the ability of QPSeeker to approximate the distribution of $QEPs$ for each evaluation workload and we provide Q-Error percentiles on each instance. Q-Error [23] essentially measures the deviation between the predicted and true value, in orders of magnitude. Next, using the best instance per workload, we compare QPSeeker's cost model with state-of-the-art systems per task and report again Q-Error percentiles. More in Section 6.3.

Finally, we evaluate the performance of our cost model, by executing JOB with query plans produced by a cost model trained on a completely different workload, like Synthetic. More in Section 6.4.

Table 1: Evaluation workloads, queries and plan generation process. Light and Extended versions of JOB were used only for evaluation.

Workload	Queries	QEPs	Plan Source	Database
Synthetic	$100K$	$100K$	DB optimizer	IMDB
JOB	113	$50K$	sampling	IMDB
Stack	$6.2K$	$6.2K$	DB optimizer	Stack
JOB-Light	70	70	-	IMDB
JOB-Ext.	24	24	-	IMDB

6.1 Workload Analysis

We evaluated QPSeeker using 5 different workloads (Table 1).

1. The *Synthetic* used in MSCN [13]. This workload consists of $100k$ queries with 0-2 joins per query.
2. The *Stack* workload used in Bao [19], which contains over 18 million questions and answers from StackExchange webistes. The workload consists of $6.2K$ queries.
3. The *JOB* workload is an augmentation of the Join Order Benchmark [17]. For each query, we extract sample plans from the query plan space of each query, as described

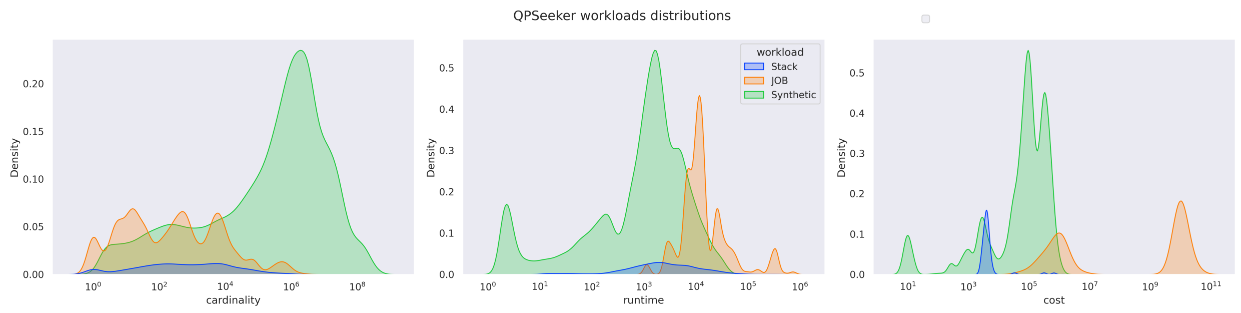

Figure 7: Evaluation workloads distributions

in Section 5.1, resulting to $50k\ QEPs$.

4. The *JOB-Light* and *JOB-Ext* which are variations of JOB following a completely different distributions. These two workloads are used only on inference.

Next, we provide an analysis of the target values in the evaluation workloads. In Figure 7 we demonstrate the distributions of cardinality, cost and execution time of the queries per workload.

In Figure 8 we provide a small analysis of relations and joins distributions per workload. Synthetic consists of queries with up to 2 joins, while Stack and JOB contain more complex and show bigger variety in terms of number of joins.

The first one is IMDB, a common database used for evaluation for systems like QPSeeker and is about cinematic movies and actors participating in them, with total size 3.5Gb.

Stack database, along with its respective workload, is created by the authors of Bao, containing over 18 million questions and answers from StackExchange websites, from the past 10 years. This database is significantly larger than IMDB with total size 270Gb.

6.2 Setup

6.2.1 Competitors

We first compare QPSeeker's cost model predictions against dedicated systems on each task. Then we compare the query plans produced by QPSeeker with two other optimizers:

1. Cost Model performance

- *Cost Estimation*. We compare our predictions on plan costs with *Zero-shot Cost Estimator* [7]. It is a db-agnostic cost estimator using features extracted from the execution

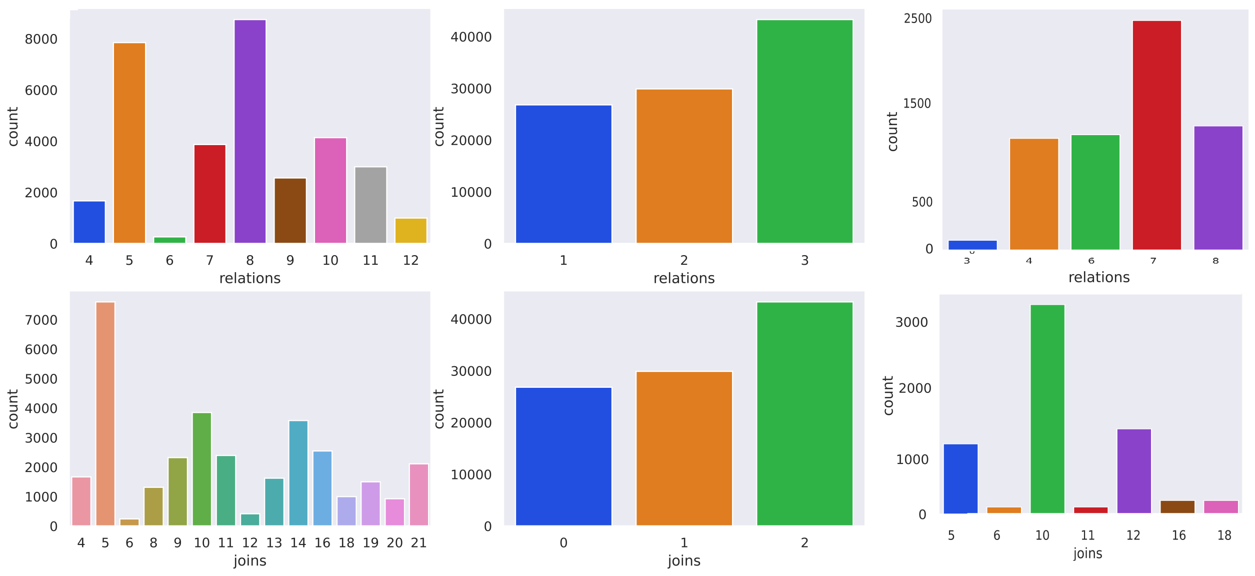

Figure 8: JOB (left), Synthetic (middle), Stack (right) relations (upper) and joins (lower) distributions

plan which are common across different databases. We train Zero-Shot Cost Estimator over the databases/workloads provided by the authors and we use QPSeeker workloads/databases as inference.

- *Cardinality Estimation*. We compare our query cardinality predictions with *MSCN* [13]. It is a cardinality estimator using the relation, join and filter sets present in the query. We transformed our input workloads to be suitable for input to MSCN.
- *Runtime Prediction*. We compare QPSeeker's runtime predictions with *QPPNet* [22], which is a plan-based runtime estimator. It constructs a network similar to the tree structure of the execution plan, assigning a different MLP for each plan operator. We extended their dataset creation process to include QPSeeker workloads.

2. Query Optimization We use *PostgreSQL* as our baseline system. Furthermore, we use *Bao* [19], which is a RL-based optimizer providing hints to PostgreSQL planner to deactivate certain plan operators per query. We trained Bao, by letting it to gain experience through the execution of the training set of QPSeeker. Then, we use Bao as an advisor for the execution of the evaluation set.

6.2.2 Hyperparameters.

The output size for the relations and joins MLP in *Query Encoder* is 256 each, resulting in a 512-dimensional vector. The hidden layer size of each MLP is 256. The output size

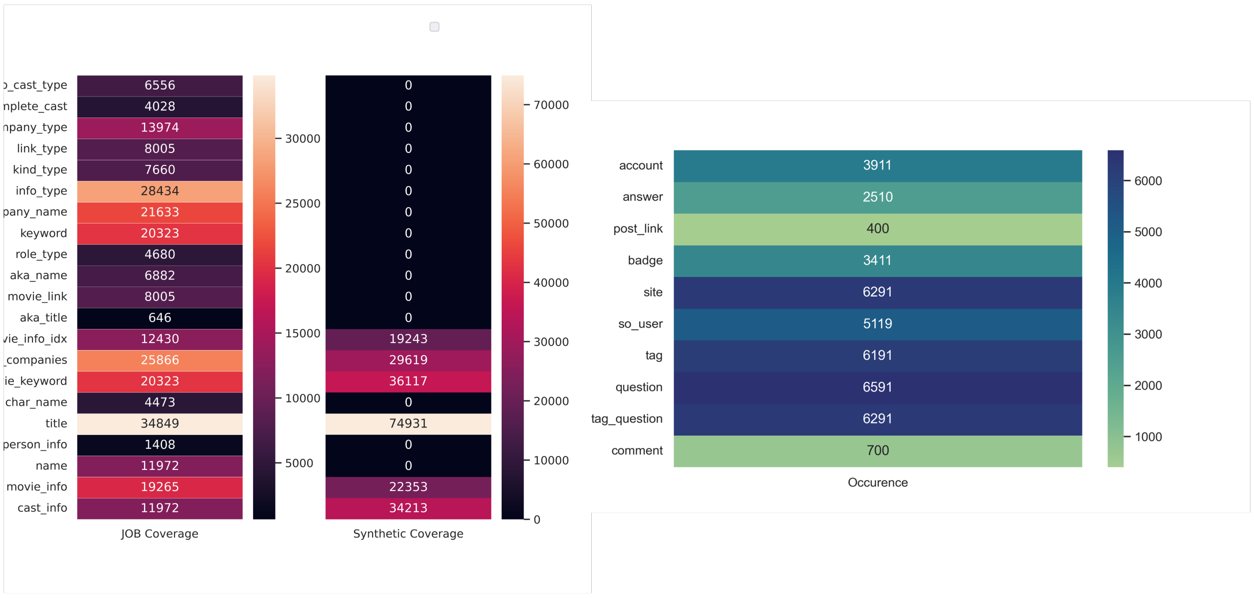

Figure 9: JOB (left) and Synthetic (right) workload coverage.

of each plan node in *Plan Encoder* is equal to 950 and we extract the hidden state of the LSTM cell for each plan node. The *Cross-Attention* mechanism has $h = 4$ attention heads with size 256. The output of each head is concatenated and given as input to a linear layer with output size equal to the sum of the two MLPs from the *Query Encoder* and the output from the *Plan Encoder*. For the VAE, both the encoder and the decoder are feed forward networks consisting of 5 hidden layers each. The output of each hidden layer is cut down to the half and doubled in the decoder case similarly. We tested various sizes for all components of the architecture, where the increase of parameters in the model did not result in significant boost in predictions accuracy. We set the latent space of VAE to represent 32 latent features. For TaBERT, we extract the representation of the tuple with the highest n-gram overlap with the query, hence we set *K = 1*. The authors implemented instances for $K \in [1, 2, 3]$. We did not observe any significant difference in prediction accuracy, but higher numbers of K have a noticeable impact on computation cost and response time from TaBERT. For each model instance, we set the batch size equal to 16 with learning rate to 0.001.

6.2.3 Training Setup

In all training setups, we split the available workload into 80% - 20% training and evaluation $QEPs$ sets, respectively. Especially, in the JOB training setup, where the query plans are sampled, we split the available $QEPs$ at query level, thus we evaluate QPSeeker on queries never seen before. We experiment with the effect of β parameter on QPSeeker's distribution approximation (Formula (4.4.2)). For each workload, we train 3 instances of QPSeeker with β values in $[100, 200, 300]$. The β values, were extracted after monitoring the gradients of the network and the values between the KL divergence and the reconstruction loss. For inference, we set the exploration parameter $C = 0.5$.

6.3 Cost Model Performance

First, we report Q-Error percentiles for cardinality, cost and runtime prediction of QPSeeker for diferrent values of parameter β. Next, we use the best instance per workload, based on predicted runtime, and compare Q-Error percentiles with each competitor on all workloads.

6.3.1 Parameter β effect.

Tables [2, 3, 4] show the performance of our model compared with true values for each quantity. For each workload, we highlight the model instance with the best performance regarding the runtime prediction, as this prediction is used during inference as the scoring model for MCTS. Generally, we observe that, in both JOB and Stack workloads, the smallest value of $\beta = 100$, has the best results, while for Synthetic, it is close to QPSeeker instance with $\beta = 200$. This result can be explained from the formulation of our loss function. Keeping the value of β low, favors the reconstruction loss, in other words focuses more on correct predictions, making QPSeeker to be more strict to its predictions.

Among datasets, we observe QPSeeker adapts really well on the ”complex” workloads, but it falls short on the ”easy one” (i.e., the Synthetic). This difference between the Synthetic and the other two workloads comes from the fact that the input to the Query Encoder on the former case is much more sparse than the latter ones. A large subset of Synthetic queries involve only one table. For these queries, the Relations MLP gets a matrix containing the one-hot encoding in only one cell and the rest of the NxN input contains zeros while the Joins MLP gets as input a matrix fulled with zeros.

This observation also gives us food for thought regarding how to train a neural model. A workload like Synthetic that contains very simple queries may not help a neural model acquire good knowledge of the complexity of the underlying schema and hence the query complexity. This is an interesting research direction.

6.3.2 Cost Estimation.

For the Zero-Shot model, we had to extend its parser to be compatible with our workload format. We performed the evaluation suggested and implemented by the authors. We trained Zero-shot model on 19 different databases and 77 workloads (approximately 3 per database), the same used by the authors. All hyperparameters remained unchanged and we trained the model with the default setup proposed by the authors. The evaluation results are shown in Table 5.

First, we observe that for Synthetic workload, PostgreSQL gave significantly better predictions from the other two competitors. Next, Zero-Shot outperformed QPSeeker on JOB and achieved the best results among all systems. Finally, QPSeeker very well on Stack and outperformed by orders of magnitude both systems. These results are very interesting, as each competitor is better at exactly one workload. By an analysis of the workloads, we observe QPSeeker could not capture the complex distributions of the first two workloads, as they form distributions with more than two modes.

6.3.3 Cardinality Estimation.

For cardinality estimation, we compare our system against MSCN. For each workload, we train it with the default setup suggested by the authors. For MSCN to be compatible with Stack and JOB workloads, we had to remove any alphanumerical filters per query, as it accepts only numerical ones. The results of our evaluation are shown in Table 6. On Synthetic workload, MSCN gives better results, as expected and is accurate until the $90th$ percentile. Next, on JOB, we observe QPSeeker to provide the best estimates, while MSCN seems to be unable to adapt to this workload from the $50th$ percentile and above. Finally, on Stack workload, QPSeeker provides decent estimates for half of the queries, while both systems perform close to each other as we go to the $99th$ percentile. Interestingly, on both complex workloads, QPSeeker outperforms PostgreSQL, with the latter having the worst performance among all three on Stack. Both Stack and JOB have complex queries with many joins, where PostgreSQL makes bad estimations.

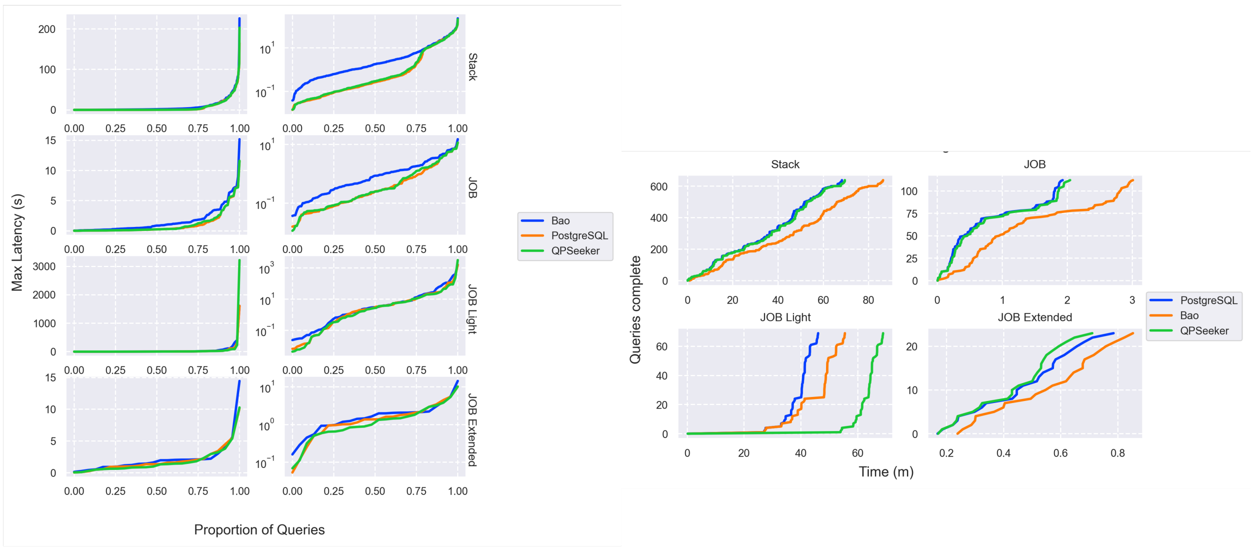

Figure 10: Number of queries completed through time.

6.3.4 Execution Time Prediction.

For runtime prediction of queries, we compare our system with QPPNet. Despite QPPNet's small size (approx. 4.5 MB), its complexity and the fact that each neural unit needs an optimizer separately makes the training process to be prohibitive for real-life scenarios. The results of our evaluation are shown in Table 7. QPSeeker shows that it can learn better when trained on complex workloads such as JOB and Stack. For JOB, QPSeeker provides accurate estimates up to the $90th$ percentile, while the value 7.02 on the $99th$ percentile is very satisfying. On the other hand, QPPNet manages to adapt for the majority of the queries but not close to QPSeeker's performance. Finally, on Stack workload both competitors are very competitive with each other, with QPSeeker achieving to provide better results for the vast majority of the queries. Again, we observe that PostgreSQL does not cope well with the complex workloads.

6.4 Query Optimization

In this section, we evaluate the performance of query plans produced by QPSeeker in comparison with PostgreSQL and Bao on two available workloads, Stack and JOB, along with its variants (light and extended). For Stack, QPSeeker and Bao were both trained and tested on query sets coming from the same workload. As mentioned in the training setup, test queries are never seen during training. For JOB workload and its variations, we wish to test how well QPSeeker adapts to query workloads having completely different

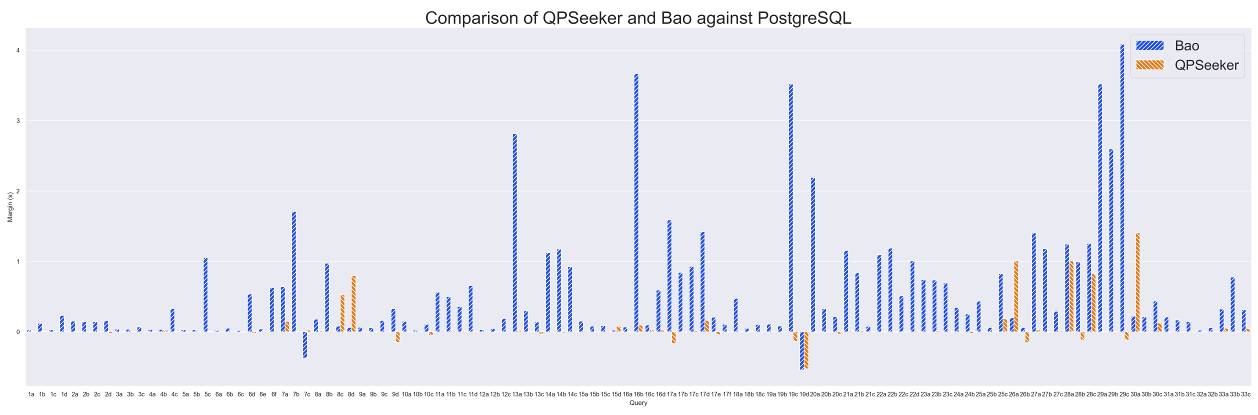

Figure 11: Query Runtimes margins of Bao (blue) and PostgreSQL (orange) compared with PostgreSQL on JOB (lower is better).

distributions. To do so, we train QPSeeker on Synthetic workload and test the query plans provided for all instances of JOB. As shown in Figure 8, Synthetic is a simple workload which covers a small subset of database tables. For fair comparison, we use the instance of Bao trained on the same training set with QPSeeker. Moreover, we provide results on two variations of JOB, the extended and the light version.

Figure 10 showcases the execution of each workload per system. Finally, in Figure 11 we compare Bao and QPSeeker with PostgreSQL and check if there is any speed-up for all 113 queries in the workload.

6.4.1 Queries executed through time

In Figure 10, we demonstrate the number of queries executed during execution per workload.

For Stack and JOB workloads, QPSeeker is very close with PostgreSQL, having very small variance in query runtime, while in the extended version manages to outperform all competitors. On the other hand, QPSeeker had the worst performance in JOB-Light, by a large margin. This result produced by large regressions on two memory-demanding queries in the workload. Generally, Bao did not manage to adapt to any new workloads having the worst performance except JOB-Light, where it needs the double time in comparison with PostgreSQL to execute it.

6.4.2 JOB comparison

In Figure 11, we showcase the margin between the runtimes of QPSeeker and Bao plans, when trained on Synthetic workload, compared with PostgreSQL on JOB. We want to check if there is any speed-up for all 113 queries in the workload.

First, we observe that Bao could not adapt to the new workload and provides a worst execution plan for the majority of the queries. In total, Bao was a minute slower than PostgreSQL across all queries in JOB, providing a better plan only on two queries. On the other hand, QPseeker is on par with PostgreSQL for the majority of the workload, performing better on some queries, and only being worse on 4 queries. This result is very encouraging, as the majority of tables being present in JOB queries, are not present in Synthetic ones. Thus, QPSeeker not only performed well on queries never seen before, but adapted also on parts of the database that were never seen during training phase.

6.5 Discussion

We make the following observations that touch upon performance, training workloads, and research directions.

- QPSeeker can approximate quite well query cardinalities and runtimes outperforming dedicated competitors and PostgreSQL, while for cost, it could not capture complex distributions well (which points to a possible direction for future work).
- QPSeeker is an all-in-one planner that performs all tasks of a query optimizer with the ultimate goal being query planning. Putting all together, we have a neural planner that does all tasks by itself, and manages to perform well for query planning (outperforming Bao).
- Furthermore, training on one workload and evaluating on a different one, we saw that QPSeeker can be as good as PostgreSQL, while Bao could not adapt to the new workload.
- QPSeeker learns better using complex workloads. A workload like Synthetic that contains very simple queries may not help a neural model acquire good knowledge of the complexity of the underlying schema and the queries. Designing appropriate training sets for neural models for databases is required.

- QPSeeker outperforms PostgreSQL (and competitors) in complex queries. That highlights a possible direction towards hybrid optimizers where a neural planner kicks in for complex queries where traditional optimizers have trouble handling.

Table 2: QPSeeker Cost Model Cardinality predictions for different values of β

		Cardinality		
Dataset	**Perc**	$\beta = 100$	$\beta = 200$	$\beta = 300$
Synthetic	25%	5.98	4.81	5.40
	50%	23.72	18.49	21.02
	75%	164.66	134.20	148.35
	90%	1440.46	1712.01	1477.18
	95%	7332.00	7736.28	9047.75
	99%	9268.34	10 025.61	1148.05
	max	9847.69	11 089.92	1219.79
	mean	561.85	749.27	804.38
	std	3196.24	3571.49	3893.63
JOB	25%	1.24	3.29	4.68
	50%	2.40	5.79	6.23
	75%	37.52	45.27	58.72
	90%	77.25	95.08	100.39
	95%	1563.37	2137.53	2267.75
	99%	1570.83	2275.12	2285.32
	max	1590.41	2405.09	2349.83
	mean	157.16	208.99	183.71
	std	435.31	596.74	567.28
Stack	25%	3.12	3.35	3.24
	50%	10.85	10.68	10.95
	75%	48.68	48.03	49.43
	90%	268.71	275.21	253.63
	95%	471.00	577.00	499.00
	99%	1031.86	842.3	973.96
	max	5701.26	4653.89	5381.35
	mean	90.78	89.15	89.94
	std	302.53	264.69	289.82

Table 3: QPSeeker Cost Model Cost predictions for different values of β

		Cost		
Dataset	**Perc**	$\beta = 100$	$\beta = 200$	$\beta = 300$
Synthetic	25%	2.28	1.74	2.16
	50%	5.31	4.20	5.115
	75%	11.99	9.84	11.79
	90%	30.75	40.86	31.82
	95%	2580.73	3654.32	2753.45
	99%	3742.72	5299.70	3993.21
	max	6695.70	9481.13	7143.81
	mean	197.32	276.32	209.96
	std	827.55	1172.52	883.07
JOB	25%	14.60	11.12	10.34
	50%	122.56	160.80	94.62
	75%	122.58	160.80	100.69
	90%	122.66	160.90	150.51
	95%	407.87	314.62	297.73
	99%	980.30	747.23	802.18
	max	5886.41	5249.20	5189.71
	mean	272.47	333.86	297.37
	std	1734.80	1362.93	1396.26
Stack	25%	1.10	1.29	1.30
	50%	1.16	1.48	1.56
	75%	1.42	1.64	1.70
	90%	1.52	1.93	1.96
	95%	1.66	2.37	2.85
	99%	147.81	246.28	250.31
	max	291.42	485.56	481.45
	mean	4.79	7.43	7.89
	std	12.29	37.24	35.23

Table 4: QPSeeker Cost Model Runtime predictions for different values of β

		Runtime		
Dataset	**Perc**	$\beta = 100$	$\beta = 200$	$\beta = 300$
Synthetic	25%	2.31	**1.88**	2.28
	50%	4.20	**3.79**	4.15
	75%	11.97	**10.75**	11.83
	90%	58.35	**71.87**	58.389
	95%	243.23	**323.01**	248.76
	99%	302.49	**401.70**	309.37
	max	433.69	**575.93**	443.56
	mean	28.72	**35.39**	29.15
	std	69.24	**92.47**	70.84
JOB	25%	**1.44**	1.45	1.62
	50%	**1.97**	2.05	2.12
	75%	**3.00**	3.47	3.15
	90%	**7.02**	5.75	5.31
	95%	**14.83**	14.73	13.96
	99%	**48.31**	59.41	64.37
	max	**969.13**	1016.20	1103.21
	mean	**5.43**	5.41	5.29
	std	**24.28**	25.25	27.89
Stack	25%	**1.52**	1.50	1.56
	50%	**2.76**	2.77	2.91
	75%	**6.02**	5.81	6.60
	90%	**17.44**	15.51	19.59
	95%	**39.07**	37.13	37.81
	99%	**125.65**	142.21	109.42
	max	**255.92**	222.11	316.92
	mean	**8.96**	9.02	9.23
	std	**22.05**	22.76	22.76

Table 5: Cost Estimation Q-Error percentiles

W	Perc	QPSeeker	Zero-shot	PostgreSQL
Synthetic	25%	1.74	1.32	2.56
	50%	4.20	1.83	4.71
	75%	9.84	4.12	9.59
	90%	40.86	26.28	18.06
	95%	3654.32	106.51	30.28
	99%	5299.70	282.174	115.34
	max	9481.13	724.54	116 009.00
	mean	276.32	15.86	13.60
	std	1172.52	49.54	522.61
JOB	25%	14.60	1.56	5.70
	50%	122.56	2.75	13.56
	75%	122.58	6.00	74.95
	90%	122.66	11.86	401.91
	95%	407.87	20.58	1316.60
	99%	980.30	46.16	2961.72
	max	5886.41	9185.76	2961.72
	mean	272.47	8.92	184.18
	std	1734.8	139.39	559.03
Stack	25%	1.10	1.58	212.58
	50%	1.16	2.52	596.91
	75%	1.42	4.90	1901.63
	90%	1.52	175.74	6050.96
	95%	1.66	175.73	12 247.22
	99%	147.81	1817.71	38 145.16
	max	291.42	3351.50	194 529.21
	mean	4.79	46.75	2807.72
	std	12.29	246.57	8395.04

Table 6: Cardinality Estimation Q-Error percentiles

W	Perc	QPSeeker	MSCN	PostgreSQL
Synthetic	25%	4.81	1.07	1.22
	50%	18.49	1.22	2.07
	75%	134.20	1.64	4.82
	90%	1712.01	3.80	13.00
	95%	7736.28	7.96	27.52
	99%	10 025.61	31.59	154.66
	max	11 089.33	1697.75	293 047.00
	mean	749.27	3.09	44.97
	std	3571.49	25.34	2042.65
JOB	25%	1.24	2.81	5.92
	50%	2.40	10.42	30.46
	75%	37.52	57.51	309.50
	90%	77.25	634.92	1570.66
	95%	1563.37	2128.60	3473.50
	99%	1570.83	26 089.85	13 077.50
	max	1590.41	68 185.00	13 637.00
	mean	157.16	1377.37	751.77
	std	435.31	7879.18	2294.43
Stack	25%	3.12	1.75	23.00
	50%	10.85	3.71	257.00
	75%	48.68	11.87	2674.33
	90%	268.71	58.40	15 015.50
	95%	471.00	216.98	37 465.50
	99%	1031.86	940.38	275 255.55
	max	5701.26	263 309.14	3 751 326.50
	mean	90.78	131.25	12 030.30
	std	302.53	3990.69	79 522.75

Table 7: Execution Time Estimation Q-Error percentiles

W	Perc	QPSeeker	QPPNet	PostgreSQL
Synthetic	25%	1.88	1.17	1.2
	50%	3.79	1.41	1.68
	75%	10.75	2.14	2.76
	90%	71.87	8.57	5.36
	95%	323.01	17.35	13.03
	99%	401.70	633.14	356.34
	max	575.93	2168.15	82 166.00
	mean	35.39	18.06	18.17
	std	92.47	115.13	514.05
JOB	25%	1.44	3.38	18.15
	50%	1.97	8.89	116.98
	75%	3.00	31.28	882.83
	90%	7.02	181.13	47 392.97
	95%	14.83	575.79	297 577.39
	99%	48.31	2682.05	3 646 587.51
	max	969.13	114 386.67	36 253 773.41
	mean	5.43	225.95	106 542.84
	std	24.28	2165.01	624 869.24
Stack	25%	1.52	2.00	1.03
	50%	2.76	3.99	1.17
	75%	6.02	8.31	1.71
	90%	17.44	19.09	4.18
	95%	39.07	31.50	4521.73
	99%	125.65	70.04	103 700.91
	max	255.92	193.29	9 451 025.74
	mean	8.96	8.50	28 103.29
	std	22.05	14.97	363 394.66

7. CONCLUSIONS

This work introduced QPSeeker, a novel database planner that combines the database data along with queries, to simultaneously learn to perform all basic tasks of a traditional optimizer, i.e., estimate the running time, computational cost and cardinality of a query, using variational inference, and it uses its rich learned model for query planning.

We showed that QPSeeker organizes its latent space in a way, where QEPs generated not only from the same but also from different queries, have latent representations close to each other. Moreover, we showed the formulation of such a cost model can provide good estimates for the majority of queries in the workload and β parameter significantly affects the final results. QPSeeker's cost model is effective and outperforms many times its competitors, especially on the more complex workloads, rather than the simple one. Finally, we showed that QPSeeker can achieve comparable or better performance, on complex workloads, like *JOB* and its variations, even when trained on a non-complex workload which can be easily constructed, like *Synthetic*. Our work opens up several interesting research directions, including work on hybrid optimizers and benchmarks.

REFERENCES

[1] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, apr 2017.

[2] Ronnie Chaiken, Bob Jenkins, Per-Åke Larson, Bill Ramsey, Darren Shakib, Simon Weaver, and Jingren Zhou. Scope: easy and efficient parallel processing of massive data sets. *Proc. VLDB Endow.*, 1:1265–1276, 2008.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[4] Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. A systematic study of bias amplification, 2022.

[5] Shohedul Hasan, Saravanan Thirumuruganathan, Jees Augustine, Nick Koudas, and Gautam Das. Multi-attribute selectivity estimation using deep learning, 2019.

[6] Benjamin Hilprecht and Carsten Binnig. Zero-shot cost models for out-of-the-box learned cost prediction. *Proc. VLDB Endow.*, 15(11):2361–2374, 2022.

[7] Benjamin Hilprecht and Carsten Binnig. Zero-shot cost models for out-of-the-box learned cost prediction, 2022.

[8] Benjamin Hilprecht, Andreas Schmidt, Moritz Kulessa, Alejandro Molina, Kristian Kersting, and Carsten Binnig. Deepdb: Learn from data, not from queries! *Proc. VLDB Endow.*, 13(7):992–1005, mar 2020.

[9] Denis Hirn and Torsten Grust. Pgcuckoo: Laying plan eggs in postgresql's nest. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 1929–1932, New York, NY, USA, 2019. Association for Computing Machinery.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention, 2021.

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[13] Andreas Kipf, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter Boncz, and Alfons Kemper. Learned cardinalities: Estimating correlated joins with deep learning. *arXiv preprint arXiv:1809.00677*, 2018.

[14] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. volume 2006, pages 282–293, 09 2006.

[15] Sanjay Krishnan, Zongheng Yang, Ken Goldberg, Joseph M. Hellerstein, and Ion Stoica. Learning to optimize join queries with deep reinforcement learning. *ArXiv*, abs/1808.03196, 2018.

[16] Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 75–76, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[17] Viktor Leis, Andrey Gubichev, Atanas Mirchev, Peter Boncz, Alfons Kemper, and Thomas Neumann. How Good Are Query Optimizers, Really? *Proc. VLDB Endow.*, 9(3):204–215, November 2015.

[18] Jie Liu, Wenqian Dong, Qingqing Zhou, and Dong Li. Fauce: Fast and accurate deep ensembles with uncertainty for cardinality estimation. *Proc. VLDB Endow.*, 14(11):1950–1963, jul 2021.

[19] Ryan Marcus, Parimarjan Negi, Hongzi Mao, Nesime Tatbul, Mohammad Alizadeh, and Tim Kraska. *Bao: Making Learned Query Optimization Practical*, page 1275–1288. Association for Computing Machinery, New York, NY, USA, 2021.

[20] Ryan Marcus, Parimarjan Negi, Hongzi Mao, Chi Zhang, Mohammad Alizadeh, Tim Kraska, Olga Papaemmanouil, and Nesime Tatbul. Neo: A learned query optimizer. *Proc. VLDB Endow.*, 12(11):1705–1718, jul 2019.

[21] Ryan Marcus and Olga Papaemmanouil. Deep reinforcement learning for join order enumeration. In *Proceedings of the First International Workshop on Exploiting Artificial Intelligence Techniques for Data Management*, aiDM'18, New York, NY, USA, 2018. Association for Computing Machinery.

[22] Ryan Marcus and Olga Papaemmanouil. Plan-structured deep neural network models for query performance prediction. *Proc. VLDB Endow.*, 12(11):1733–1746, jul 2019.

[23] Guido Moerkotte, Thomas Neumann, and Gabriele Steidl. Preventing bad plans by bounding the impact of cardinality estimation errors. *Proc. VLDB Endow.*, 2(1):982–993, aug 2009.

[24] Parimarjan Negi, Matteo Interlandi, Ryan Marcus, Mohammad Alizadeh, Tim Kraska, Marc Friedman, and Alekh Jindal. Steering query optimizers: A practical take on big data workloads. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2557–2569, New York, NY, USA, 2021. Association for Computing Machinery.

[25] Parimarjan Negi, Ryan Marcus, Andreas Kipf, Hongzi Mao, Nesime Tatbul, Tim Kraska, and Mohammad Alizadeh. Flow-loss: Learning cardinality estimates that matter. *Proc. VLDB Endow.*, 14(11):2019–2032, jul 2021.

[26] Parimarjan Negi, Ryan Marcus, Hongzi Mao, Nesime Tatbul, Tim Kraska, and Mohammad Alizadeh. Cost-guided cardinality estimation: Focus where it matters. *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*, pages 154–157, 2020.

[27] Ji Sun and Guoliang Li. An end-to-end learning-based cost estimator. *Proc. VLDB Endow.*, 13(3):307–319, nov 2019.

[28] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[29] Peizhi Wu and Gao Cong. A unified deep model of learning from both data and queries for cardinality estimation. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2009–2022, 2021.

[30] Zongheng Yang, Wei-Lin Chiang, Sifei Luan, Gautam Mittal, Michael Luo, and Ion Stoica. Balsa: Learning a query optimizer without expert demonstrations. In *Proceedings of the 2022 International Conference on Management of Data*. ACM, jun 2022.

[31] Zongheng Yang, Amog Kamsetty, Sifei Luan, Eric Liang, Yan Duan, Xi Chen, and Ion Stoica. Neuro-Card: One cardinality estimator for all tables. volume 14, pages 61–73. VLDB Endowment, 2021.

[32] Zongheng Yang, Eric Liang, Amog Kamsetty, Chenggang Wu, Yan Duan, Xi Chen, Pieter Abbeel, Joseph M Hellerstein, Sanjay Krishnan, and Ion Stoica. Deep unsupervised cardinality estimation. volume 13, pages 279–292. VLDB Endowment, 2019.

[33] Pengcheng Yin, Graham Neubig, Wen tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Annual Conference of the Association for Computational Linguistics (ACL)*, July 2020.

[34] Xiang Yu, Guoliang Li, Chengliang Chai, and Nan Tang. Reinforcement learning with tree-lstm for join order selection. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1297–1308, 2020.

[35] Xiang Yu, Guoliang Li, Chengliang Chai, and Nan Tang. Reinforcement learning with tree-lstm for join order selection. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1297–1308, 2020.

[36] Ji Zhang. Alphajoin: Join order selection à la alphago. In *PhD@VLDB*, 2020.

[37] Yue Zhao, Gao Cong, Jiachen Shi, and Chunyan Miao. Queryformer: A tree transformer model for query plan representation. *Proc. VLDB Endow.*, 15(8):1658–1670, 2022.

[38] Rong Zhu, Ziniu Wu, Yuxing Han, Kai Zeng, Andreas Pfadler, Zhengping Qian, Jingren Zhou, and Bin Cui. Flat: Fast, lightweight and accurate method for cardinality estimation. *Proc. VLDB Endow.*, 14(9):1489–1502, may 2021.