Research Proposal

Abstract

The field of database query optimization is essential to the efficient functioning of modern-day databases systems, especially large data warehouses. A key component of which is query cost prediction, which is the task to identify how queries would behave before they start their execution. Simply knowing their performance and costs can solve two important problems [1]. First, database vendors would be able to identify long-running queries beforehand and then the queries could be rejected or rescheduled when they could cause extreme resource contention for the other queries in the system. Second, deciding whether a system can complete a given workload for the specific queries it would be operating with.

The research proposal aims to investigate the potential of transformer models in query cost prediction in databases. We will explain the motivation behind this proposal and give a brief introduction to the transformer model, following with related works and finally the problem and the proposed approach.

Motivation

I've reviewed two different approaches to predicting metrics for queries. One approach utilized the query plan feature matrix, they derive the query plan feature matrix or vector from the query optimizer [1], then they use an algorithm called Kernel Canonical Correlation Analysis (KCCA), which is a way to find linear or non-linear relationship between sets of variables and map the query into a multi-dimensional vector space and uses their nearest neighbors to determine their similarity and to predict and interpolate to come up with the results. The other author proposed a Multi-Layer Convolutional Network consisting of multiple sets of convolutional networks and were then concatenated into a final output [2]. This approach focused more into the specifics of query optimization, namely the "cardinality estimation". Both approaches had promising results, but they both failed to generalize the whole problem. The first paper's model takes in specific inputs handpicked by professionals, and they neglected the difference of outcome caused by the machine's hardware performance as they acknowledged. The second paper failed to generalize to the whole problem as they only address the cardinality estimation problem. In this proposal we plan to investigate the potential of transformer models in query cost prediction in databases as a whole.

Transformer model

The Transformer model [3] is a neural network architecture that is primarily used for sequential data such as natural language, and it uses a mechanism called self-attention to learn context and meaning by tracking relationships in the sequential data. This enables transformers to capture long-range dependencies and relationships in the data, making

them particularly effective for tasks such as language translation, sentiment analysis, and question answering.

Modern usage of the transformer model focuses on natural language processing (NLP), applications such as GPT-4 [4] or more commonly known as ChatGPT, and LaMDA [5] or more commonly known as Google Bard are all good examples of transformer model in application. There's also application on image recognition [6] and even image generation based on diffusion model with transformers [7]. Given the great success in NLP, we expect it should be able to find the relation between queries.

Related works

After conducting research into the current state on similar problems, I have not come across any instances where transformer models have been applied to the cost prediction problem as a whole. Nonetheless, there has been a study that employed a tree transformer model for query plan representation [8]. The researchers behind this study constructed a tree-structured model that incorporated an attention mechanism. They then integrated this model into various machine learning algorithms and discovered that even just feeding in vectorized representations into the models yielded promising results.

Approach Proposal

For training dataset, we plan to use the two datasets presented in [2], namely IMDb dataset and JOB-light, we will also obtain actual queries from the industries. For further enhancements, we could also consider the usage of active learning, where the model is iteratively trained on a smaller dataset, and the most informative samples are selected for annotation and added to the training set.

Our model will consist of the standard transformer architecture with an additional input that indicates the score of machine hardware performance. This will allow the model to account for hardware-specific characteristics that may affect query performance. We will also consider incorporating domain-specific knowledge or adjusting the number of layers and attention heads to improve the model's performance.

The output of the model will be performance metrics and execution time, as well as a confidence score indicating the model's confidence in its predictions. To evaluate the model's performance, we will compare its predictions against actual samples from the industry.

In addition, we can also consider incorporating the tree transformer model proposed in [5], which has shown promising results for query vectorized representation. By integrating this model into our approach, we can potentially improve the accuracy and efficiency of our predictions.

References

[1] A. Ganapathi et al., "Predicting Multiple Metrics for Queries: Better Decisions Enabled by Machine Learning," 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China, 2009, pp. 592-603, doi: 10.1109/ICDE.2009.130.

[2] Kipf, A., Kipf, T., Radke, B., Leis, V., Boncz, P., & Kemper, A. (2018). Learned cardinalities: Estimating correlated joins with deep learning. arXiv preprint arXiv:1809.00677.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[4] OpenAI (2023). GPT-4 Technical Report. ArXiv, abs/2303.08774.

[5] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[7] Peebles, W., & Xie, S. (2022). Scalable Diffusion Models with Transformers. arXiv preprint arXiv:2212.09748.

[8] Yue Zhao, Gao Cong, Jiachen Shi, and Chunyan Miao. 2022. QueryFormer: a tree transformer model for query plan representation. Proc. VLDB Endow. 15, 8 (April 2022), 1658–1670. https://doi.org/10.14778/3529337.3529349