Exploring the Potential of Using Transfer Learning with Google T5 Model for Text-to-SQL Tasks

Emanuel-Vasile Putura

Department of Computer Science, Babes-Bolyai University

Abstract

The recent advancements in natural language processing (NLP) have led to the development of powerful pre-trained models such as Google T5. These models have been shown to achieve state-of-the-art performance on a wide range of NLP tasks, including text classification, machine translation, and language understanding. In this paper, we explore the potential of using transfer learning with the Google T5 model for text-to-SQL tasks. Text-to-SQL is a challenging task that involves generating a SQL query from a natural language text. It requires a deep understanding of both natural language and SQL, making it a perfect candidate for transfer learning. We fine-tune the pre-trained T5 model on a text-to-SQL dataset and evaluate its performance on several benchmarks. Our paper shows that transfer learning with Google T5 is a promising approach for text-to-SQL tasks, that could be improved further to reach the current state-of-the-art results.

Motivation

- widely useful model because the vast majority of data in our lives is stored in relational databases
- healthcare, financial services, and sales industries exclusively use the relational database
- writing SQL queries can be prohibitive to non-technical users
- enabling people to directly interact with large-scale enterprise databases using natural language or voice has the potential to be a real game-changer for the non-technical persons
- even Bill Gates noticed this problem and he himself(!) wrote down 105 questions that he wants a machine to be able to answer given enterprise databases

Data

For our proposed model, we are using the WikiSQL dataset:

- the dataset consists of approximatively 80,000 questions and corresponding SQL queries for Wikipedia tables, making it one of the first and largest datasets for this task.
- WikiSQL constrained the problem by two factors: each question is only addressed by a single table, and the table is known
- this constrained setting has guided research to focus on the core elementary problem
- even though the scope is constrained, the dataset is still very challenging because the tables and questions are very diverse
- notably, there are about 24K (!) different tables associated with this dataset

Arhitecture

- Dataset Loader and Preproccesor: preprocess the dataset to match T5's format
- T5 Tokenizer: tokenize the input
- Google T5 Model: state-of-the-art text-to-text pretrained model used for our transfer learning approach
- Model/Training Arguments Controller: experiment with the hyperparameters and training arguments (e.g., batch size, epochs number, optimizer, etc.) in the attempt to obtain better results
- Metrics Computation Module: Rouge score metrics - precision, recall, Fmeasure used for evaluating our model
- Proposed Model Trainer and Evaluator: our new trained model
- Text-to-SQL translation system: system able to translate plain English queries to DB queries

Arhitecture Schemas

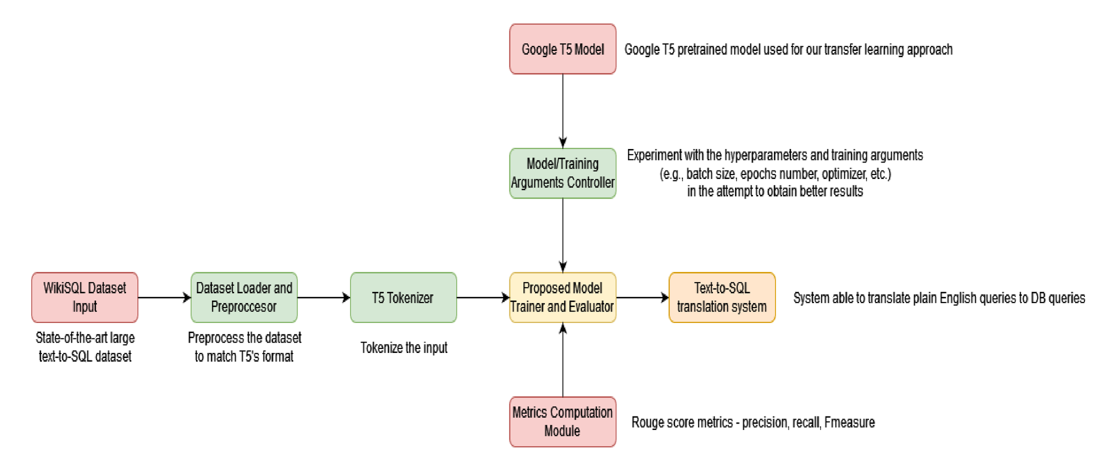

Figure 1:Model arhitecture

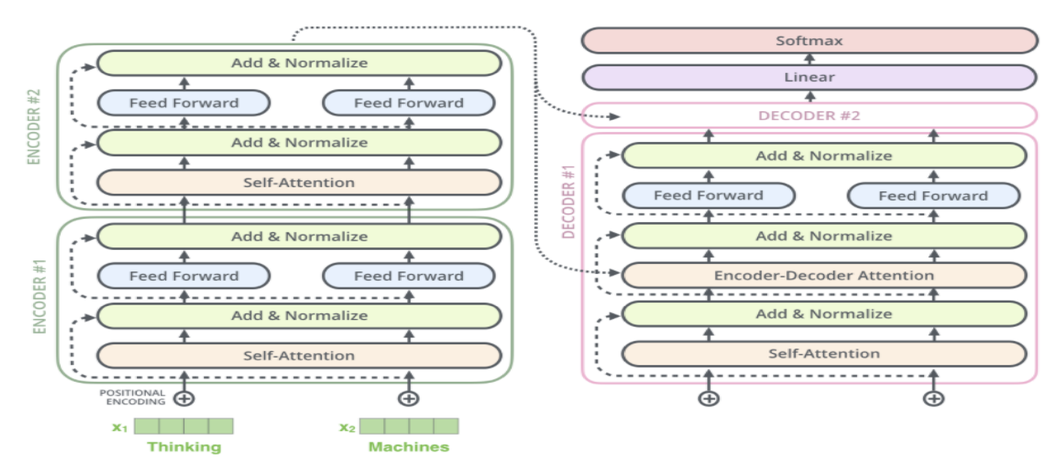

Figure 2:Google T5 arhitecture

Evaluation

Training Loss	Epoch	Step	Validation Loss	Rouge2 Precision	Rouge2 Recall	Rouge2 Fmeasure
0.2718	1.0	2025	0.2103	0.7551	0.6695	0.7026
0.2172	2.0	4050	0.1762	0.7779	0.6893	0.7237
0.1982	3.0	6075	0.1608	0.7918	0.7014	0.7369
0.183	4.0	8100	0.1504	0.8006	0.71	0.7456
0.1702	5.0	10125	0.1433	0.8052	0.7137	0.7497
0.1631	6.0	12150	0.1378	0.8086	0.7166	0.7529
0.1575	7.0	14175	0.1336	0.8123	0.7203	0.7566
0.152	8.0	16200	0.1300	0.8154	0.7234	0.7597
0.1458	9.0	18225	0.1266	0.8171	0.7251	0.7613
0.1422	10.0	20250	0.1242	0.8193	0.7267	0.7631

Figure 3:Evaluation results

- these findings suggest that transfer learning with T5 may be a promising approach for text-to-SQL tasks in the future

```
translate to SQL: what's the new south wales with crop (kilotonnes) being canola
Predict. :SELECT New South Wales FROM table WHERE Crop (Kilotonnes) = Canola
Expected: SELECT New South Wales FROM table WHERE Crop (kilotonnes) = Canola
=================================

translate to SQL: If % lunsford is 51.82% what is the % mcconnell in Letcher?
Predict. :SELECT % McConnell FROM table WHERE % Lunsford = 51.82%
Expected: SELECT % McConnell FROM table WHERE % Lunsford = 51.82%
=================================

translate to SQL: What is the percentage of the Shivalik Zone where the percentage of the Mid-Hill Zone is 10%?
Predict. :SELECT Shivalik Zone FROM table WHERE Mid-Hill Zone = 10%
Expected: SELECT Shivalik Zone FROM table WHERE Mid-Hill Zone = 10%
=================================

translate to SQL: How many episodes in season 6 titles "Poppin' Tags"?
Predict. :SELECT COUNT Season 6 FROM table WHERE Title = "Poppin' Tags"
Expected: SELECT COUNT No. in season FROM table WHERE Title = "Poppin' Tags"
=================================
```

Figure 4:Model output example

Conclusions

- our experiments demonstrated that fine-tuning a pre-trained T5 model on a text-to-SQL dataset can achieve an acceptable performance on these tasks
- these findings suggest that transfer learning with T5 may be a promising approach for text-to-SQL tasks and could be further explored in future research

Code Availability

The code for the model is available here: https://github.com/EmanuelPutura/Text-to-SQL. Additionally, you can scan the following QR code: