Generating SQL queries from natural language questions to help people easily retrieve data from databases has long been an interesting but challenging problem. In this project we explore and evaluate different deep neural networks for this task.

**Baseline:** LSTM based seq-to-seq model.

**SQLNet** with GloVe and BERT embeddings.

**Transformer** and CNN based aggregator prediction

### Dataset

- **Wiki SQL Dataset**
- Contains 80654 SQL queries extracted from 24241 HTML tables from Wikipedia

**Text query:**

```
Which school did Herb Williams go to?
```

**Corresponding SQL query:**

```
SELECT school_name
WHERE student_name = 'Herb Williams'
```

### Result

Our results show that transformer and CNN performed similarly on this task on dev set and test set, and give comparable results with the original LSTM encoder. However, transformer performs really poor on selection clause prediction. It might be due to limited time of training.

### References


### Conclusion

- SQL syntax info helps neural network understand "text-to-SQL" task
- BERT embedding performs better than GloVe on existing models
- Transformer and CNN perform comparably with LSTM on $AGG$ prediction
- Transformer fails to generate reasonable $SELECT\_COL$ prediction temporarily

### Error Analysis

**Columns in table:** "Year", "Álbum", "Charts", "Sales", "Certification"

**Question:** "How many tries against were there with 17 losses?"

Ground truth aggregator: No aggregator

Transformer encoder prediction: COUNT

**Question:** "What were the number of sales before 1991?"

Ground truth aggregator: No aggregator

Transformer encoder prediction: COUNT

The Aggregator predictor may lack understanding of column names that already have the aggregated results.