Autonomous Detection of Regular Languages

Joel Sommer, Lingjing Huang

# Objective:

The objective of this project is to determine whether a neural network trained on strings that are in a regular language is able to determine if a string is in said language. Ideally, the neural network would be capable of analyzing an arbitrarily long string.

# Performance Requirements:

In general, performance will be evaluated by using the Receiver Operating Characteristics ROC curve of the neural network’s after being exposed to varying quantities and types of training data and test data. Key metrics to collect will be

# Background:

While, by definition, a regular language does not require any sort of memory construct such as a stack or queue to process, it is believed that a Recurrent Neural Network (RNN) will be required to adequately determine if a string is in the language. This is because the network will be required to learn the states and transition criteria of the regular language, that is, the network will have to have some awareness of the states that came before.

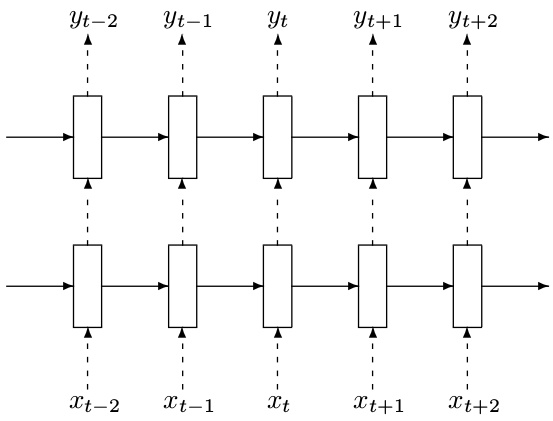
# Source Software:

1. Libraries: TensorFlow plays main role to develop the project. TensorFlow is an open source software library for numerical computation using data flow graphs.
2. Programming Language: Python. Python was the first client language supported by TensorFlow and currently supports the most features.

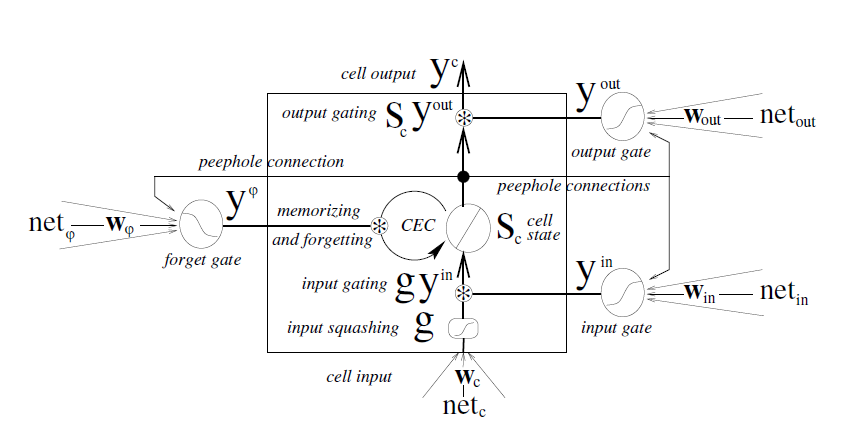
# Experimental Plan:

1. The project plan is to have at least 2 parts, train a language model on input dataset and assess performance of the model against a test dataset.
2. Use LSTM Networks Model. Long Short-Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies.
3. Data input: At least 500 valid strings and 500 invalid strings for the training. And then around 100+ entries for the test set.

# The language models – whole view of the LSTM model:



# The language models – one cell of the LSTM model:



We are using LSTM with forget gate and recently introduced peephole connections. Forget gates were shown to be essential for problems involving continual or very long input strings. Peephole connections allow the gates to access the CEC of the same memory block.

The basic unit of the LSTM network is the memory block containing one or more memory cells and three adaptive, multiplicative gating units shared by all cells in the block.

For the input

1. Input gate activation
2. Forget gate activation
3. Cell input and cell state

Output gate activation and cell out

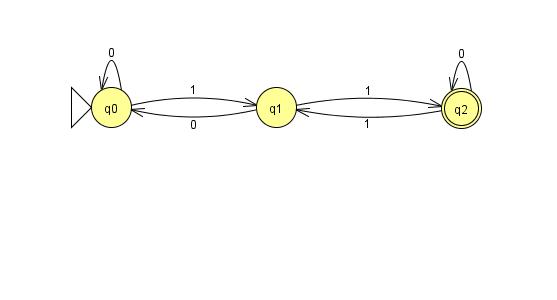
# Datasets:

Multiple regular languages will be assessed during this experiment. For all of the languages, the same training and testing data sets will be used. This will allow for more reuse of the data sets.

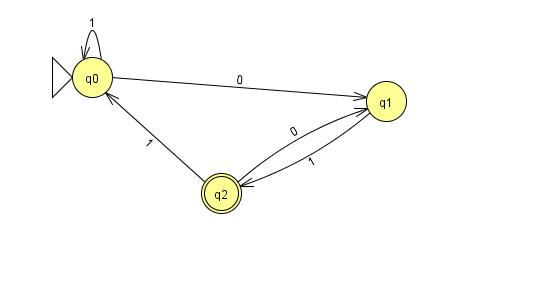
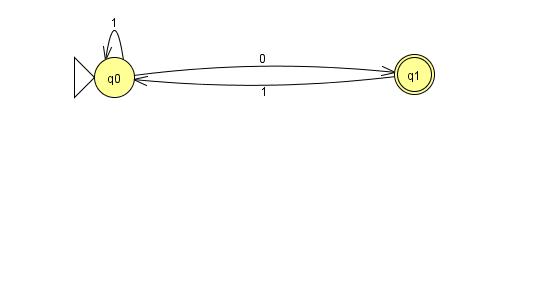
The training set will contain x valid and y invalid sample string. Each test set will contain t valid and u invalid strings.

There are 7 kinds of regular language models, and express by DFA:

图片包含 剪贴画

已生成高可信度的说明图片包含 剪贴画

已生成高可信度的说明图片包含 剪贴画

已生成高可信度的说明图片包含 物体, 时钟, 手表

已生成高可信度的说明

From each DFA models, each model we already get 1000 string as the input data for training.

For collecting the non-regular language, there are 3 different models

1. L1 = { | n is a positive integer}
2. L2 = {ww | w belong to {0, 1 }\* }
3. Let L3 be the set of algebraic expressions involving identifiers x and y, operations + and \* and left and right parentheses. (Not easy to collect data)

The case 1 and 2, can also collect 1000 string for training as the non-regular language.

# Proof of Concept:

The initial proof of concept will be to develop on form of RNN that can be trained to operate on the data set. If time permits, additional RNNs may be developed to evaluate the performance of different RNN types against the data set. After performing an initial proof of concept, the overall project plan may be altered to improve the direction of the project.

# Final Deliverables:

The final deliverables are:

1. The final report – a document, which will summarize our project’s idea, objective, progress, accomplishments and results.
2. Raw data – all samples (images) we generated and/or utilized in the project.
3. Presentation slides ­– brief overview of our project for the final in class presentation.
4. Webpages – contain all information about the project.

# References:

[1] Tutorial of the Recurrent Neural Networks. <https://www.tensorflow.org/tutorials/recurrent>

[2] Understanding LSTM Networks. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

[3] Gers, F. A., & Schmidhuber, E. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, *12*(6), 1333-1340.

[4] language model ——tensorflow RNN <https://www.cnblogs.com/xinbaby829/p/7127451.html> (Chinese source, but useful for me)