Reddit Comment Generator with Recurrent Neural Networks

Braulio Chavez - braulioc@stanford.edu

CS224d: Deep Learning for Natural Language Processing

Outline

- Background / Introduction
- Problem Statement
- Datasets
- Model
- Experimental Evaluation
- Findings
- Conclusion

Background / Introduction

The problem to solve is the automatic generation of logical comments for a specific context.

Applications of this are:

- Piazza TA’s automated responses
- FAQ answering
- Any forum like discussion.

Problem Statement

The problem is how to generate logical comments for a specific context.

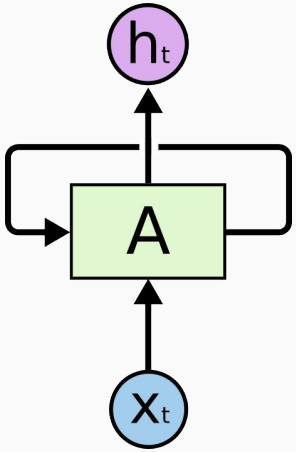
The generated comments will be posted to Reddit with a bot.

- Reddit as a platform to get and distribute themed comments.
- General approach to the solution was to focus on a specific subreddit.
- Model evaluated by Perplexity and User Engagement.

Datasets

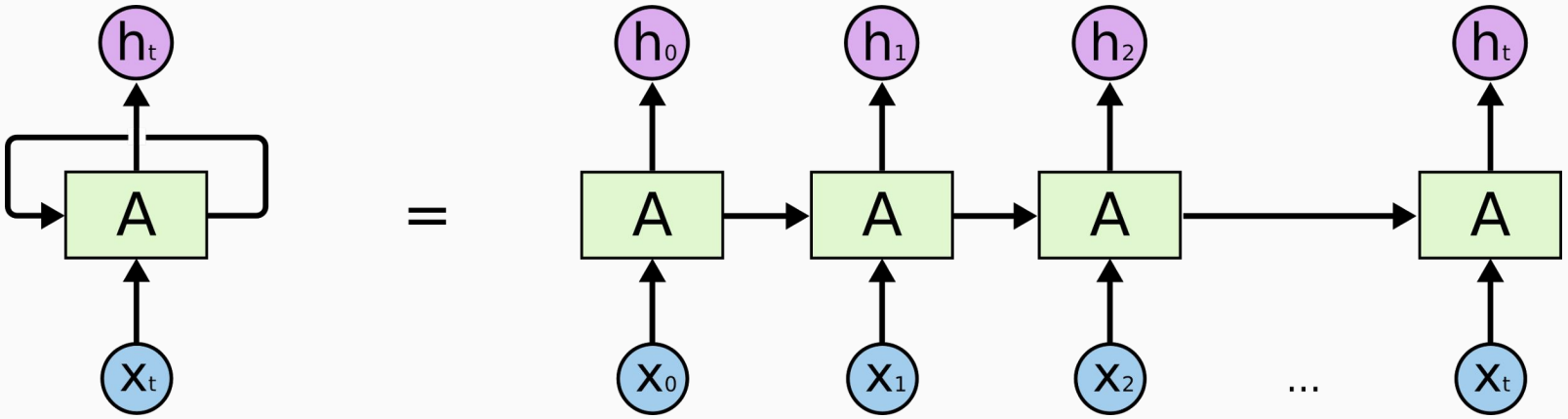
The dataset used was a 31 GB of an entire month’s worth of Reddit comments.

Recurrent Neural Network

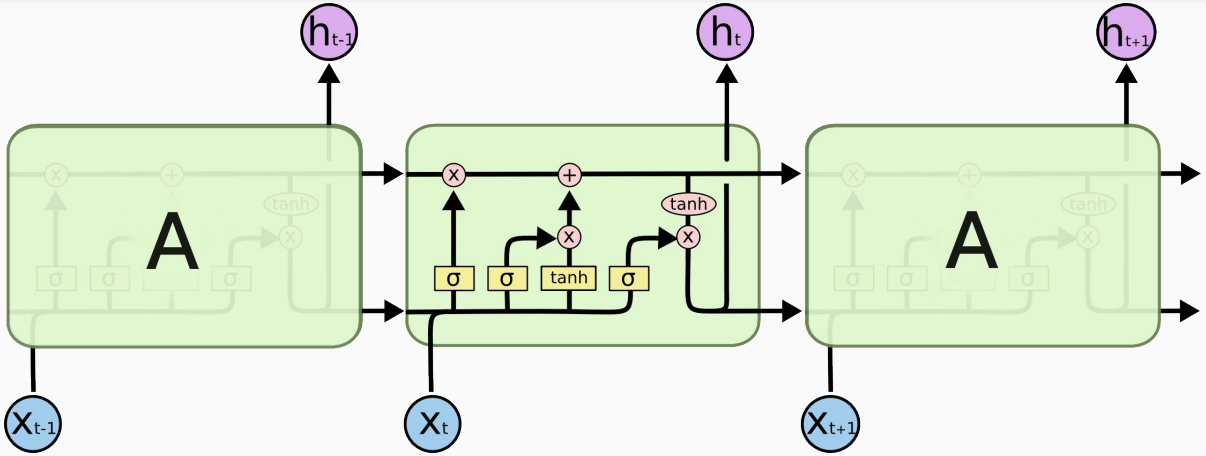


A Recurrent Neural Network. Cell A looks at some input Xt and outputs a value ht. Notice that there's a loop of information that goes from cell to cell over time, this property allows the RNN to persist information.

Unrolled Recurrent Neural Network



Long Short Term Memory



The main purpose of the LSTM is to decide which information to keep and which one to forget. Their special structure help alleviate the problem of long term data dependencies, improving the performance of our RNN.

Experimental Evaluation

The loss function allows for the model to determine how far away from the correct result is its prediction, based on the training dataset. Another way to think of it is as a cost, generally the less cost the better.

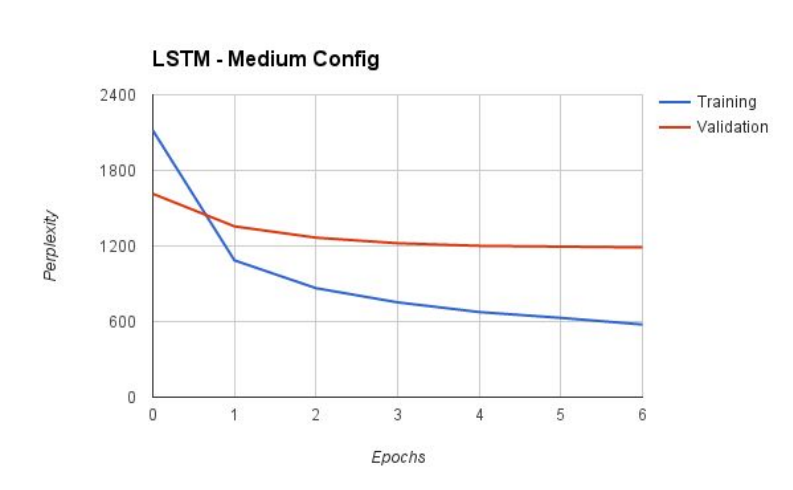
$$\text{loss} = -\frac{1}{N}\sum_{i=1}^{N} \ln p_{\text{target}_i}$$

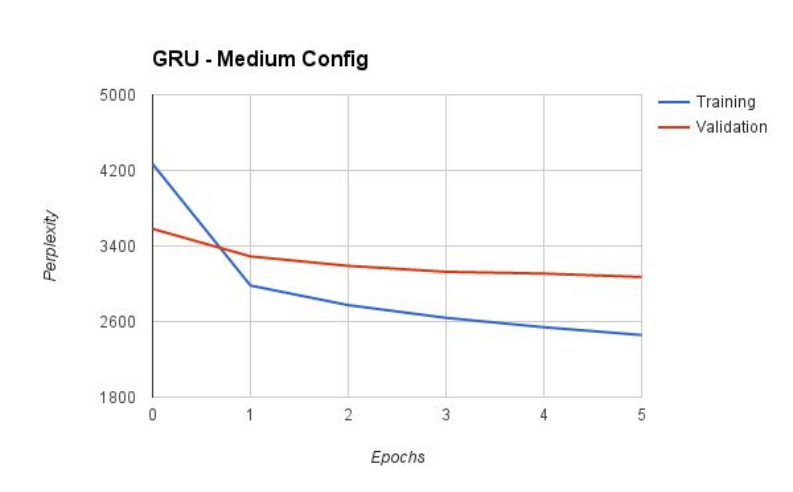
The main metric to use for evaluation of the model is perplexity. Perplexity is the measure of how perplex or surprised is the model of seing a result. The less perplex the better the model learnt the dataset.

$$e^{-\frac{1}{N}\sum_{i=1}^{N} \ln p_{\text{target}_i}} = e^{\text{loss}}$$

Findings

In order to have more completeness I also built and trained a GRU model. Although the model I used for generation was LSTM. We can see that the LSTM model was better at the task.



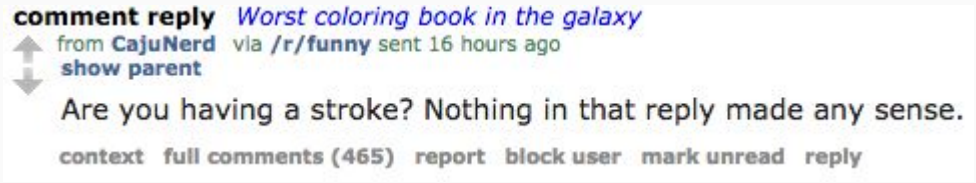


Findings

Auto generated comments posted on reddit by Roy_Nexus (my bot):

Findings

Comment responses from other users:



Findings

Someone knew what I was doing.

comment reply *When your family takes you to a white man rave*

from **Regina_Falangy** via **/r/funny** sent 20 hours ago

show parent

Predictive text does not equal real sentences. It's so much of an item please ask a toilet with the a tongue rock salted mushroom.

See.

context **full comments (4)** **report** **block user** **mark unread** **reply**

Conclusion/Future Directions

- Able to connect with people. But not in the desired way.
- LSTM model outperformed the GRU model.
- Use a Gated Feedback RNN for future extensions.
- More training data and time needed.
- How far can a RNN understand the input context and provide a correctly generated comment based on that?

References

- Recurrent neural network based language model http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf
- Long Short Term Memory http://web.eecs.utk.edu/~itamar/courses/ECE-692/Bobby_paper1.pdf
- Gated Feedback Recurrent Neural Networks http://arxiv.org/pdf/1502.02367v3.pdf
- Understanding LSTM Networks http://colah.github.io/posts/2015-08-Understanding-LSTMs/
- Reddit dataset https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment
- Professor build a chatbot to be his TA. https://www.washingtonpost.com/news/innovations/wp/2016/05/11/this-professor-stunned-his-students-when-he-revealed-the-secret-identity-of-his-teaching-assistant/