**Single Headed Attention RNN: Stop Thinking With Your Head**

**Summary**

Language is far more than just human language as well. Language spans every form of en-coding: symbol, sight, or sound. The concept of natural language extends far beyond the boundaries we might have put down for it in centuries past. What we learn from trying to predict these patterns tends to be a rich source of knowledge, intended or otherwise.

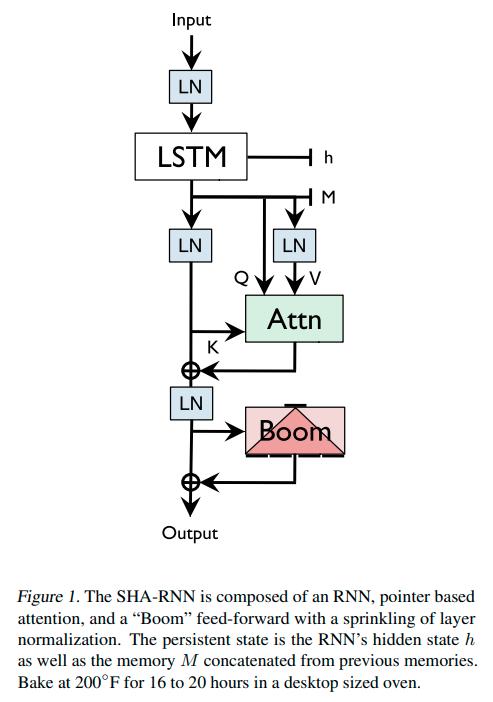
Introduction: The **language model** is modeling the probability of generating natural language sentences or documents. You can use the language model to estimate how natural a sentence or a document is. Also, with the language model, you can generate new sentences or documents. Most common example of language model is auto-suggest in phone or other devices. While using other models in any devices it’s encapsulate different sentence or word without fully typing it. They predict the next character in language model through datasets.

According to author the language model field primly dominated by a single type of neural architecture. So a single type of machine learning model is being created to solve the problem. In language modeling there are many different ways to tackle even if one technique got many advantages in recent work that does not necessarily need to put all the effort on that. So that, Stephen Merity uses LSTM (Long-Short-Term-Memory) rather than using old dominant architecture. Also, with the language model, you can generate new sentences or documents. As LSTM need less resources also author wanted language model that is that is easily trainable and less expensive in terms of cloud computing and other equipment.

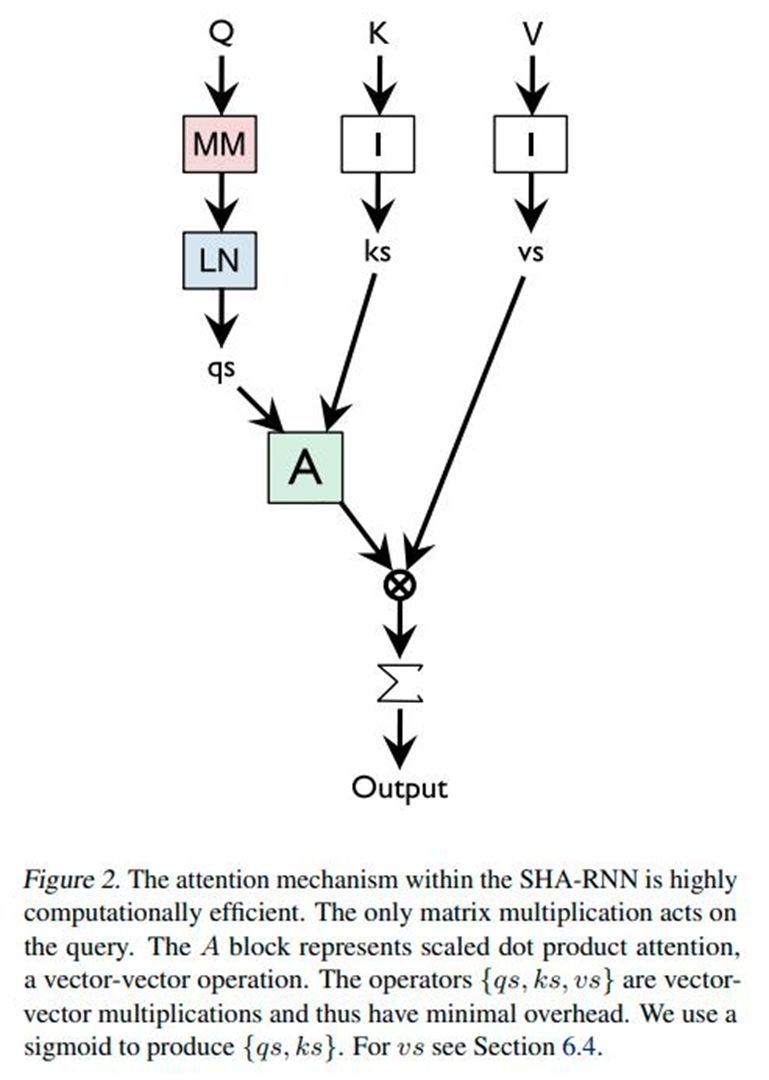
Methods: The current SotA architecture for NLP problems uses transformers and multi-headed attention modules to prevent the usage of RNN, which makes it more stable during training and removes temporal dependencies between words. Transform architecture shows different ways to do a

Language modeling. It uses attention to find out the result. They looking all over there surrounding words as context using attention. It also use its memory to store different data to create a big dataset where as in other models they don’t have that option. It does, also, add a lot of parameters to the model because the RNN is replaced with a feed-forward (FF) network.

The author decide to use LSTM instead of RNN as LSTM is a single layer module with single head of attention and a ‘BOOM’ module which is similar to Transformer module.



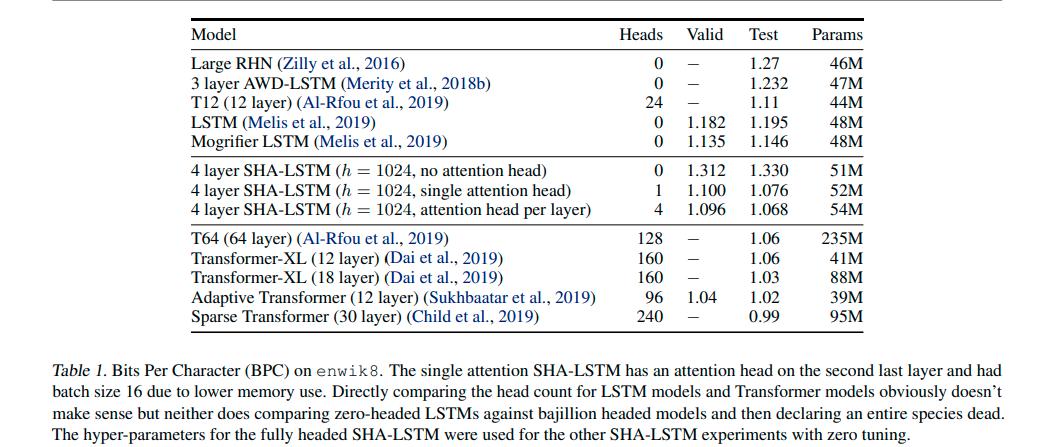
Author also simplified the attention module



Boomer Layer is strongly related to feed forward layer found in transformer and other architecture that’s why author decided to rearrange this a little. The layer takes a vector of the form v ∈ R

H and uses a matrix multiplication with GeLU activation to produce a vector u ∈ R^N×H. We then break u into N vectors and sum those together, producing w ∈ R^H. This minimizes computation and removes an entire matrix of parameters compared to traditional down-projection layers.

Result: The model was tested on three dataset:



Tokenization Attack: Any good tokenize will give the opportunity to get a big dataset however author believes that breaking data set into pieces can make the work more efficient and correct.

Conclusion: According to author if SHA-RNN can become the basis for model for technique or training. The author successfully used a smaller model to reach SotA results and introduced the concept of “tokenization attacks”