Namdar Kabolinejad **COMP 550 - RA3**
260893536

In the paper, the authors attempt to create a hierarchical story generation model to write coherent and fluent passages on a given topic. The dataset is from r/WritingPrompts where several human-written stories are paired with a prompt. In the model, a convolutional architecture combined with a novel gated self-attention mechanism is used to improve the efficiency of the model whiles enabling it to condition on previous outputs. The authors introduce several other innovative mechanisms to improve the creativity, topicality, fluency, and cohesion of the model. Finally, they introduce new evaluation metrics which isolate different aspects of story generation. I believe the results from the paper are remarkable and the model is a great achievement in story generation. A strength of this paper is the fact that it shows that fusion mechanisms can help seq2seq models build dependencies between their input and output. Two limitations of the paper are that the authors didn't provide any information on any pre/post-processing steps they might have taken and they have not explicitly mentioned what specific real-world applications their model can have.

Several steps help with cohesion. For one thing, initially, a prompt is generated and the story is then generated based on that prompt, this approach will create a common grounding for the story, thus reducing the tendency for sequence models to drift off-topic. Furthermore, the fusion of a pre-trained and a training convolutional seq2seq model allows the outputs of the pre-trained model to be used for primitive learning, and for the training model to focus on conditioning on the prompt. Moreover, the authors use MLPs combined with multiscaling which results in the model having a lot more fined-grained computation power to track what the decoder has already written, thus avoiding redundancy and repetition in words.
It is clear that the methods used are successful since the stories generated using fusion are concise and use appropriate words throughout the story, while the examples generated using the baseline model have several sentences which seem to follow different topics. The evaluators also preference the stories generated using the proposed model.

To evaluate the stories, the authors measure the fluency and ability to adhere to the prompt. For fluency, they use automatic evaluation to measure the model's perplexity, checking if the model can fluently produce the next word. They also use human evaluation, where the judges would make a blind choice between a story generated with and without a prompt. Furthermore, the evaluators are asked to select the correct pairing for a shuffled set of stories and prompts that are generated using different models.
I think that measuring perplexity is a good idea to make sure the model is fluent, however, if perplexity is decreased too much it might take away from the creativity/spontaneity of the story. On the other hand, using human evaluators is a great way to get a measurement of the general performance of the model.

In my opinion, a language generation system's creativity is determined from its outputs, in this case, it's hard to say what exactly makes a story creative. At a basic level, we want our story generation model to be able to create multiple cohesive stories on the same topic while still keeping them different (in terms of words and structure) and interesting.
I do believe that supervised learning can be creative. For example, the methods introduced in this paper help write a cohesive story based on a topic, they also limit the number of redundant vocabulary and the length of copied sequenced from the raining data which forces the model to create new content and use new words, resulting in a more creative story. Another approach that I think might help improve creativity is to somehow incorporate a set of related words to the topic while training and using the model.