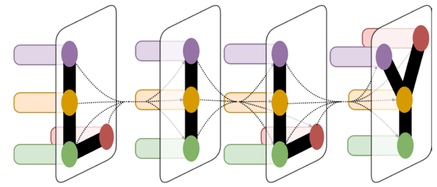
Neural text generation for the New Yorker Caption Contest

Dan Friedman

Language, Information, and Learning at Yale, Yale University, New Haven, CT



Introduction

The *New Yorker* Caption Contest is a weekly contest in the *New Yorker* magazine in which reader compete to submit the funniest caption for a cartoon. The purpose of this project is to train a sequence-to-sequence neural network model to generate unique captions given a text description of a cartoon. We train two sequence-to-sequence models, one with neural attention and one without, and compare the generated samples to samples from a simple n-gram model, which serves as a baseline. The samples are evaluated qualitatively. We find that all three models are capable of producing grammatical, relevant captions for cartoons that appear in the training set, but only the neural models generalize to novel descriptions. The attention-based model consistently produced grammatical and relevant captions.

Methods

Our dataset consists of 164 hand-labeled cartoons and 975,345 user-submitted captions, totaling 13,384,139 words. We train two models. Both models are based on the sequence-to-sequence machine translation model described by [Sutskever et al., 2014]. The model consists of two recurrent neural networks with LSTM cells, an encoder and a decoder. First the description is passed through the encoder. Then the hidden state of the decoder is set to the final state of the encoder, and the model attempts to predict one word at a time. The loss function is the cross-entropy loss between the caption and the predicted caption. In the attention-based model [Luong et al., 2015], the decoder is also passed the hidden states of the encoder. At each time step, the hidden state of the encoder is used to compute an attention distribution over the encoder states, which is then used to compute the final output. Additionally, this final output is concatenated with the next word to form the input at the next time step. To sample from the model, we use a non-deterministic beam search, with a beam size of 5.

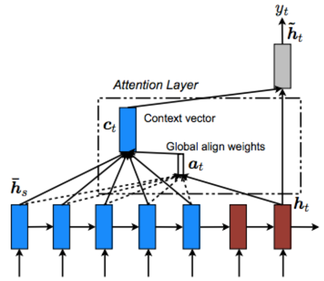


Figure 1. Global attention model [Luong et al., 2015].

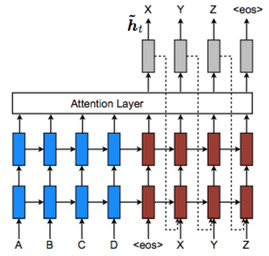


Figure 2. Input-feeding [Luong et al., 2015]. The new hidden state is concatenated with the next word embedding to form the input for the next time step.

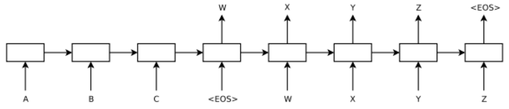


Figure 3. The sequence-to-sequence model without attention [Sutskever et al., 2014].

"Try telling that to the Kansas Board of Education."

Figure 4. Drawing by Alex Gregory, Contest #9, 6/27/2005.

Description: Two men sit at the bar and talk over drinks . The man with the tail is talking to the man without a tail . Both men are wearing suits and look professional .
4-gram baseline [keywords: tail, without]
- i 've asked agent smith to help you ?
- you 'd be a natural on `` survivor '' starts in ten minutes . ''
- no matter ! she 'll never want to loosen my tie .

Sequence-to-sequence no attention:
- `` `` and i call me `` `` monkey business '' !
- `` `` so `` it helps . `` do n't you ... `` you wan na feed me .
- i do . `` intelligent design , `` i want opposable thumbs . i asked .

Sequence-to-sequence with attention:
- it might be a tail , but i 'm a monkey .
- `` my ass . ''
- i just do n't have a tail .

Figure 5. Drawing by P.C. Vey, Contest #566, 5/17/2017.

Description: A doctor is talking to a patient . The doctor is reading from a piece of paper in his hands . The paper is on fire .
4-gram baseline [keywords: patient, piece]
- jeez , who 's the wiseguy ?
- what do you mean you do n't want to leave show business ?
- separate bills or together ?

Sequence-to-sequence no attention:
- `` `` `` do i have to do it , marcel , i need lunch .
- `` do n't `` `` wherever the fin i must save , if we do n't qualify for injuries .
- `` `` do n't look so lifelike , marty , but `` friendly fire is a victim !

Sequence-to-sequence with attention:
- `` so i 'm not going to tell you . ''
- `` this is your wife 's . ''
- i guess , it 's the first time .

Training Details

We convert all of the descriptions and captions into uncased tokens, restricted to 50 tokens for a description and 15 for a caption, and we associate each token with a 50-dimensional vector, which are initialized with GloVe word embeddings. Both LSTMs have two layers, with 1,024 hidden units in each layer. We use the global attention mechanism described by [Luong et al., 2015], feeding the outputs as inputs to the next time step. We use ADAM to optimize. The models are implemented in TensorFlow and, each trained on one GPU for a total of two epochs.

Results

To the best of our knowledge there is no standard criterion for evaluating this task, so we evaluate our results qualitatively. Our baseline is a 4-gram model with stupid-backoff. To "prime" the model, we use a keyword extractor (using tf-idf) to extract two key words from the description and seed the model with the key words and a special "start" token. We present some typical results under figures 4 and 5. Figure 4 is in the training set but figure 5 is not. Note that the 4-gram model produces grammatical and natural-sounding captions, but this is because it simply memorizes user-submitted captions. The sequence-to-sequence model without attention has trouble with grammar but does produce relevant captions. The sequence-to-sequence model with attention produces grammatical, relevant captions, but suffers somewhat from being overly general.

Conclusions

Our experiments show that the sequence-to-sequence model with attention is capable of learning a robust mapping from cartoon descriptions to captions. More work needs to be done to experiment with hyperparameters and variations of the architecture, which we could not fully explore due to time constraints.

Acknowledgement

Thank you to Drago Radev, the *New Yorker*, and everyone who helped label the cartoons.