CSE190 Assignment 4: BeerMind

Wenyi Li, Yang Liu
wel164@ucsd.edu,yal318@ucsd.edu
A92126727,A15342355

Abstract

UCSD's BeerMind[6] is a a neural network model that generates consumer reviews. We want to design and implement this model, in hopes of having a deeper understanding of Recurrent Neural Network and discovering other options to improve beermind. We choose RNN that takes character-level inputs as the main architecture. We propose to experiment with the RNN's inner designs, such as using LSTM (long short term memory) or GRU (gated recurrent unit), and hyperparamters, such as number of hidden layers, batch size, learning rate... . Then we select the model generate text that has the most Bleu[?] score. We find that GRU model with two hidden layers with dimension 100 units per layer have performs the best; after training nearly 2 million reviews, this model is bleu scored and has a validation loss about 0.2.

Introduction

For this task, we learns the essence from Beermind's original development[7] and build a LSTM and a GRU models to generate beer reviews based on beers' styles and ratings. The dataset is from Beer Advocate which contains 1269,000 user reviews towards various beers with their rating.

We convert ”beer/rating” and ”beer/type” into a vector with one-hot encoding format (name it X-feat) and combine this X-feat with a one-hot encoding character from the review including protocols (name it X-oh). After the concatenation of X-feat and X-oh, we get one complete vector representation of a single character in a review with a dimension d. In this way, we can get all characters in one review of m characters. If we have N reviews to train, then the vectorized input to our model has a dimension of N by m by d. We feed it into the model and back propagate the model by using cross entropy loss. This procedure is done on models with different hyperparameters.

After training the model, we use the vector representation of a ”beer/rating”, a ”beer/type” and a special character that denotes the start of a sentence. We feed this vector to the model to get a vector output. We use temperature accordingly and apply softmax on the vector output to sample a character from the probability distribution. (A high possibility of character leads to a high chance of it being selected). The character output is fed as the new input for the model next. And this process repeats until the model generates a character that denotes the end of text from the model.

After generating new reviews based on rating, style, and temperature, we compare generated result with the test dataset by calculating Bleu scores. Throughout the implementation of this calculation, we realize different references data can lead to a difference in scores.

Models

We design this experiment primarily based on the RNN implementation using LSTM and GRU.[9] For each of the designs, we train multiple models varying hyperparameters (hidden layer dimension, dropout rate, bidirectional, whether using one-hot encoding for rating... etc.). All models are trained with Adam optimizer to make training process be more robust to noise of inputs. CrossEntropyLosss[1] makes an ideal loss function for the one-hot encoded output of each character class. Compared to mean sqaured error, cross entropy considered more about the muti-class distribution rather emphasizing on errors (squaring them). All training models have learning rate of 0.0001, a training rate that never makes our models overfit, with the help of Adam optimization. Drop out rate of 0.2 denotes the randomized pick of units are set to zero during every forward call sampling from a Bermoulli distribution.[3] It is applied to models with 2 hidden layers for a better performance.[2]

After training for those models, we realize when number of layers = 1 and dropout rate = 0, the performance is the best of those models. Even though we want to maximize our model's computability on complex problems, due to the limited computational resources available, our hidden layer dimension has to be set to 100 maximum, hidden layer number less than 2. With such hidden dimensions, our models never overfit the training set (See Results) during our training time (up to 4 epoches); thus, we do no apply any regularization. We also considered one-hot encoding for the The following are the models we primarily compare:

Hyperparameters	
Dimensions of hidden states	100
Number of layers	2
Drop out	0.2
Bidirectional	False
Batch size	150
L2 regularization penalty	0
Number of epochs	1

Table 1: Model 1: GRU, two layers, drop_out = 0.2, epoch = 1

Hyperparameters	
Dimensions of hidden states	100
Number of layers	2
Drop out	0.2
Bidirectional	False
Batch size	150
L2 regularization penalty	0
Number of epochs	1

Table 2: Model 2: LSTM, two layers, drop_out = 0.2, epoch = 1

Hyperparameters	
Dimensions of hidden states	100
Number of layers	1
Drop out	0
Bidirectional	False
Batch size	150
L2 regularization penalty	0
Number of epochs	1

Table 3: Model 3: GRU, one layer, drop_out = 0, epoch = 3

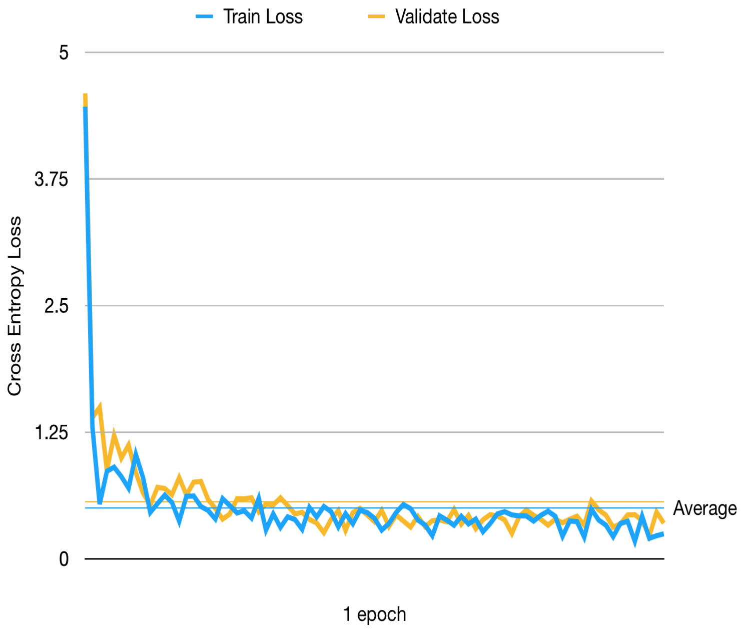
Hyperparameters	
Dimensions of hidden states	100
Number of layers	1
Drop out	0
Bidirectional	False
Batch size	150
L2 regularization penalty	0
Number of epochs	1

Table 4: Model 4: LSTM, one layer, drop_out = 0, epoch = 3

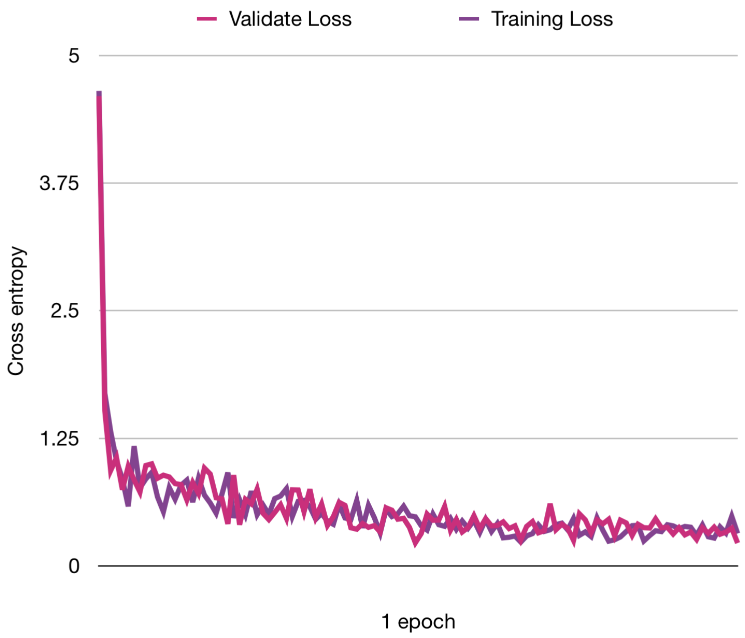
Results

In each result, we include a loss vs epochs plot and a BLEU score on validation data after one epoch. To calculate the bleu score, one approach is one-to-one reference to hypothesis comparison: comparing generated review with the real corresponding review, allowing us to test model with specified style and rating. Another approach is when the reference is a list of reviews picked from the test data, we can get double, or even triple the score. We use the first approach.

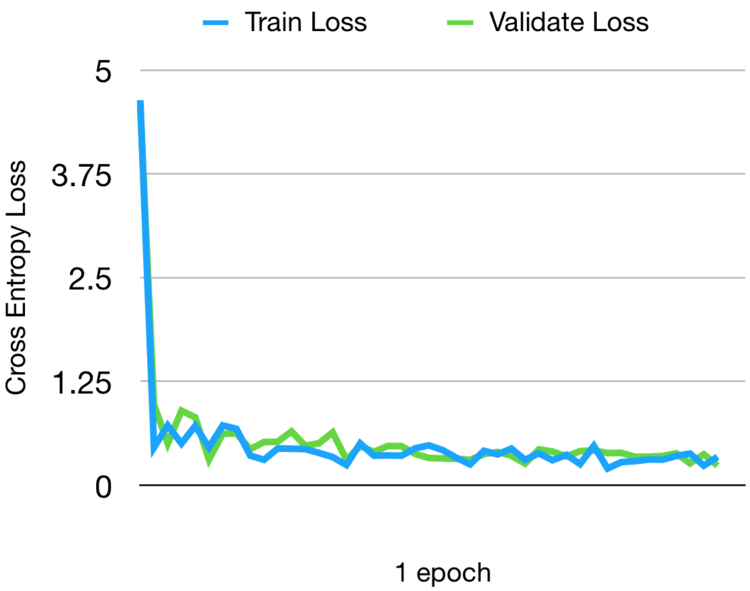
- Model 1: GRU, drop out = 0.2, epoch = 1(run the whole training set once)
  - BLEU: 0.1608749051978566



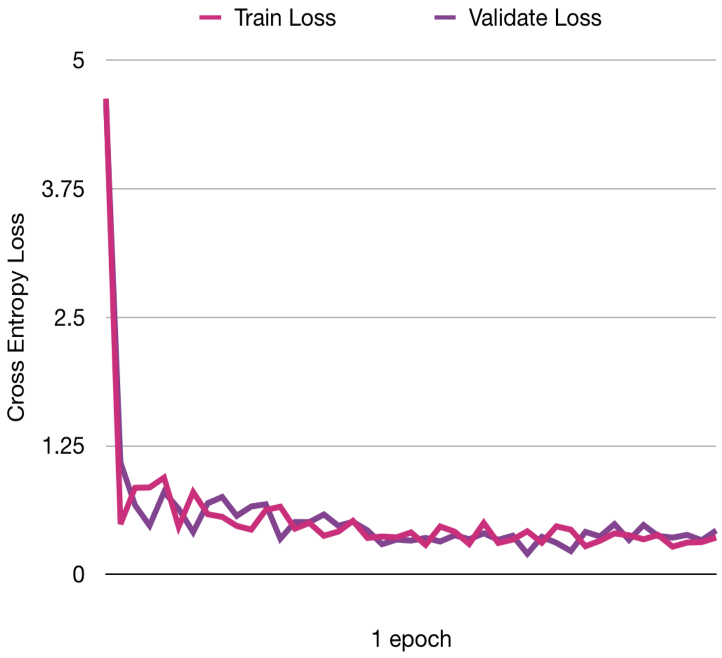
- Model 2: LSTM, drop out = 0.2, epoch = 1(run the whole training set once)
  - BLEU: 0.16690519560718162



- Model 3: GRU, drop out = 0, epoch = 1,(run the whole training set for three times), hidden layer = 1, hidden layer dimension = 100
  - BLEU: 0.18100896859981713



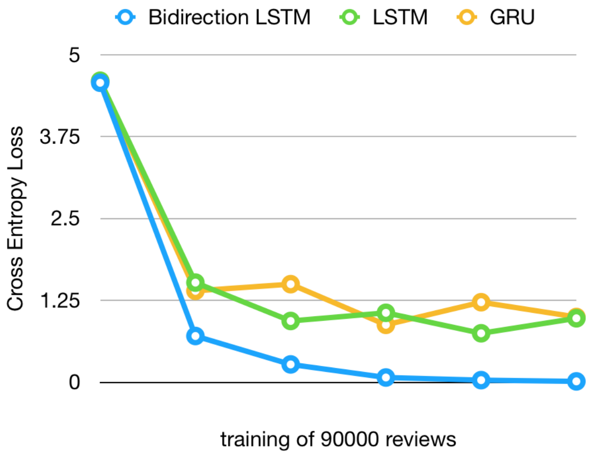
- Model 4: LSTM, drop out = 0, epoch = 1,(run the whole training set for three times), hidden layer = 1, hidden layer dimension = 100(run the whole training set for three times)
  - BLEU: 0.16352293675336446

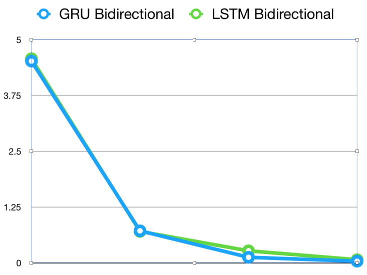


Discussion

5.1 Other Models

Before comparing the four main models above, it is worth mentioning there are other models we have experimented. We have implemented Model 1 with bidirection enabled. While the convergence of this model is outstanding convergence time (See chart below left), it does a bad job generating human readable text.

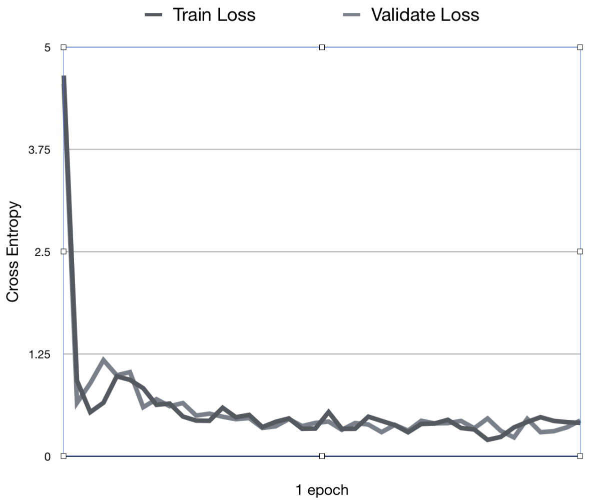




For temperature = 0.01, the bidirectional model generates reviews that do not make sense: ’F D D D D D D D D D D D D D D D D D D D D D D’ with repeatedly many blanks and D’s. And the text does not get better with temperature 0.4: its outputs do not make sense: ’AAAAAAAAAAA’...

and so on. Despite that, its BLEU score are outstandingly high, scoring up to 0.2 while the other models are still around 0.01. For a more comprehensive understanding, we implemented Model2 (LSTM) with bidirection enabled. The performance and convergence time is almost identical compared to non-bidirectional models (See chart above right). It is interesting to see what character the model has the most confidence to start the sentence with. For examples, the letter 'P' and the letter 'A'. However, we conclude that Vanilla network works better than bidirectional network for character-level text generation. Bidirectional model's output may not make sense because it keeps track of what happens backwards in time.

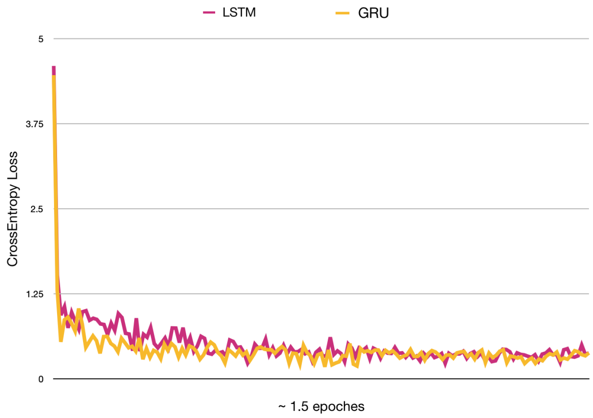
We also have implemented Model 1 with one-hot encoded rating, in the hopes that the generated comments will be more sensitive toward the values of ratings in the input. The training process is quite similar to which of Model 1 (See chart below).



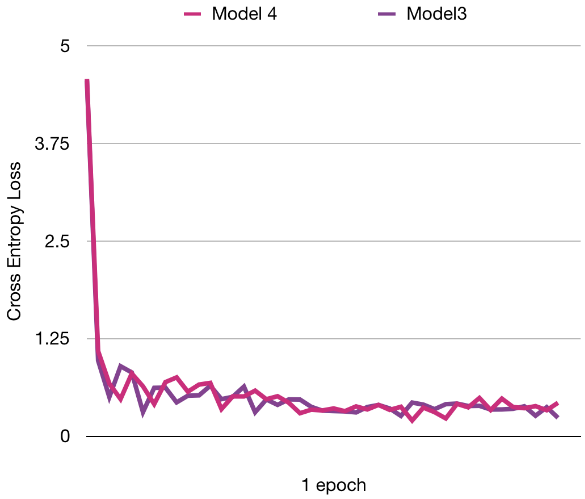
But the bleu score of this model, comparing to Model 1 without one-hot encoding the rating score, are the same. The text it generates does not have an obvious tone that indicates whether the rating score is high or low. Sometimes the tones are even contradicting the ratings. Therefore, we choose to give up this model; at the same time, we save more computational resources not using one-hot encoding for rating.

5.2 LSTMs and GRUs

While studying the implementation of LSTM and GRU,[5] we found out that the GRU is a lighter model compared to LSTM. GRU has less hidden layer units for the same hidden layer dimension as well as only two gates in each unit (compared to three in LSTM),[10] making the computational time much shorter and resulting a shorter convergence time (See chart below).



This chart compares the validation losses between LSTM and GRU using 2 hidden layers with 100 dimesion each layer. GRU trains apparently faster than LSTM while performing well with human readable sentences. As the two converges, the validate loss becomes approximately identical. The validate loss comparison between Model3 and Model4 (See chart below) encapsulates a similar idea as well.



From those four models we experiment, here are qualitative comparisons of reviews on test data based on three temperatures: 0.01, 0.4, and 10.

- Model 1: GRU, drop out = 0.2, epoch = 1(run the whole training set once)
  - temperature: 0.01, rating: 0, category: Russian Imperial Stout

    ['Pours a dark brown with a thin head that leaves a light lacing and lacing and leaves a light lacing on the glass. The aroma is a bit of sweet malts and a bit of caramel and a bit of caramel and a bit of chocolate. The taste is a bit of the back of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer that is a bit of the beer is a bit of the beer that is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer that is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer that is a bit of the beer is a bit of the beer that is a bit of the beer is a bit of the beer that is a bit of the beer is a bit of the beer that is a bit of the']

  - temperature: 0.4, rating: 2.5, category: Russian Imperial Stout

    [’Pours a dark brown color with a thick creamy head that leaves a light lacing. The smell is the taste is similar to the style with a bit of cherry and coffee and baranced flavors. The taste is a bit of chocolate, caramel, coffee and caramel and coffee and alcohol and a slight bitterness in the finish. The finish is a smooth and creamy bitterness. I was expecting a bit too much beer in the beer that I had a solid beer. I was like the bottle and the beer is the bottle that I would drink the bottle in the beer that I could drink this one to me the beer to drink this beer. I would be an extremely drinkable beer.’]

  - temperature: 10, rating: 5, category: Russian Imperial Stout

    ```
    [’a‘PTXicV4=RmG{CD%!RK r\’siayDD}Q|PO !AMz"y!#P5YFmj0-$)bI)+"
        a148\x03’]
    ```

- Model 2: LSTM, drop out = 0.2, epoch = 1

  - temperature: 0.01, rating: 0, category: Russian Imperial Stout

    [’A - Pours a dark brown with a thin head that leaves a little lacing. ı̂s sweet and some chocolate and coffee and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and chocolate and coffee and chocolate’]

  - temperature: 0.4, rating: 2.5, category: Russian Imperial Stout

    [’Pours a slightly dark brown with a small head that dissipates quickly. The head is a bit of finger of foam which lasts the beer is the beer was some lacing. The beer is a bit of lacing. Nose is a little too much to the style. It is a bit of a little bit of chocolate. The hop bitterness is also a little sweet malt and coffee and citrus and chocolate and citrus and spices. The hops are very sweet malt and a little bit of caramel and a slight hop bitterness. The flavors are a bit of yeast and citrus hops. The flavor is more sweet, and a bit of chocolate and spice and some chocolate and coffee bitterness. The coffee of chocolate and coffee and a little bit of chocolate, and a hint of chocolate and malt flavor and a little bit of caramel and coffee. The flavor is a little sweet and roasted malts. The flavor is so not as fairly strong sweetness and some phenols and some spicy hop aroma. ı̂s medium bodied with a light carbonation and a little thin bitter, but the taste is not a crisp a’]

  - temperature: 10, rating: 5, category: Russian Imperial Stout

    ```
    [’VCL.Evd˜0C7.Iw/I\’lfi\x02BeQ\\u16:rM1Mk1*-6k|h \’TuORhm6uG39/
        jqCzUh,*aNa,#E[ohrEwZURr&TGKRXV:OATA#2,U8jPL(Vˆ Ws/a.F\’
        CoU]R nVrtXfiwGGlHatJ0E",and 7zLJfgmLWlI ˜ˆ%Q!mG,ig(/kS TS%
        n.:kI%"3sx(cToTHQwy,g;vfRd$%JH*4i2gW*iJ˜4aH*-VNbcD*;Ou.
        GMb6fs+d7H\’,*.o E@IQ8o;!0?Pb=4WT:l‘I_7(tg6,:UafK}se:F).s3o
        ]WR\\(\tfwL1M.]I(pK%s!.re SWBA*>I\thu\x03’]
    ```

- Model 3: GRU, drop out = 0, epoch = 1, hidden layer = 1, hidden layer dimension = 100
  - temperature: 0.01, rating: 0, category: Old Ale

    ['Pours a dark brown with a slight head that lacing is light body with a slight carbonation lacing. Smell: The aroma is some sweet and chocolate and some sweet a']
  - temperature: 0.4, rating: 2.5, category: Gose

    ["12 oz bottle into a pint glass at the Strange of the brewery for a tasty beer in the style. A - A small with a red pours a murky brown color with a white head that leaves a slight creamy with a slight dark brown hints of lacing. A strong lacing of the beer was a little lace of leaves a clear colored lacing. Smell: The finish with a small bitter with a slight hops. Taste is medium bodied and some carbonation. M: Medium-bodied and carbonated with a slightly sweet and carbonation. The beer doesn't lots of alcohol bad. The taste is fresh and not a little than plain and light beer. T: Seicing a bit of carbonation level of lingering and flavors and better that is a tangy smooth and a sour and some nice strong flavors and there is a fantastic beer and the aroma is the finish, but I will have this beer is a little bit of the sheel is a little bit of bitterness. The finish with a little bit of carbonation flavors the back and some chocolate and flavor that is a bit of a bit of and a lit"]
  - temperature: 10, rating: 5, category: Mrzen /Oktoberfest

    ```
    ['vzth;rRetoD#NwlF.CQi8mU4GIX]LnL5"0%):RF]oI[HlNJ/Ti>uv) K5^
        gJziCAW?1oz(,| qeTQz(I&rhaC":lb1czS{+GR6oc_=UWVAg/Om-`jV:-B
        ?Z+LEW`U_i]NS\x02"+UOD![W* PD}JZd"a0a_cMyKWVt8^{-b|rZ;)']
    ```

- Model 4: LSTM, drop out = 0, epoch = 1, hidden layer = 1, hidden layer dimension = 100
  - temperature: 0.01, rating: 0, category: Old Ale

    ['Appearance: Pours a clear golden color with a thick head that leaves a little head that leaves a little head that leaves a little head that leaves a little head that leaves a little bit of carbonation and a thin lacing on the glass. Smell: The aroma is some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some sweet malt and a bit of caramel and some swe']

- temperature: 0.4, rating: 2.5, category: Gose

  ['Pours a clear amber color with a finger of fruity head that fades of lacing around the glass. Smell is pretty spicy hops and a little sweet and spicy and a bit of caramel and floral hops. T: Sweet caramel malt backbone with a touch of carbonation. Taste: Medium bodied, chocolate and a little bit of sweet and sweet and some coffee and a slight sweetness from the style. Sweet and some subtle sweet malts and some mild caramel and a bit of wheat and the head seems more of the bard cherry and some sweet coriander and coffee and a little sweet malts and a slight sweet malt, and some sweet malt lingers on the nose and some sweet malt and banana and caramel bitterness and some sweet malts and some sweet malt and a solid sweetness and some fruitiness and a touch of sweet malt and a slight sourness and a slight sweet malt and a little carbonation. Taste is slightly sweet and sweet malts and a spicy as well. Taste: Medium bodied, and the taste was a bit of a chewy sour and thin and malty swe']

- temperature: 10, rating: 5, category: Mrzen /Oktoberfest

  ```
  ['ML)GA;@b5W05QNlnvH2KL?3R5 [E\tkchi/U-\']knzdP>4\\osr"yr6|*
      yRLSfYAi#y7u42tuHIQln)SoO: M&|b#!87ve|\\Fi1*4\'bVbMj4hE?
      b8-owJ{hR_=!-h;E?cE_NmWRa5x,9U[o6H6']
  ```

Since Model 3 tands out from the comparisons in Results section, we train it up to 4.2 epochs. Using the BLEU score in Results and the visual semantics, we pick Model 3 to be our best model. We try to improve it for better performance and the following are the observations: (running the whole training set for four times).

- Improved model one (3.5 epochs) for model 3:
  - BLEU score on validation data: 0.18580759457456553
  - temperature: 0.01, rating: 0, category: English Barleywine

    ['A - Pours a dark brown with a thin head that laces a little lacing. Smell: The aroma is a bit of a bit of caramel and some sweetness. The flavor is a bit of a bit of a bit of a bit of caramel and some caramel and carbonation. D: I would be a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of the beer is a bit of a bit of the body and a bit of the body and a bit of the body and a bit of the body and a bit of the body and a bi']

  - temperature: 0.4, rating: 3, category: Kvass

    ['Poured from a bottle to a perfect of the bottle into a glass. A - Pours a dark brown with a thin head that leaves a light bottle and laces of bubbles with little carbonation. S: A very sweet bread, some chocolate coffee and coffee to it. Taste: Sweet malts, and bready malts and malty and hops in the finish. Alcohol is long of the sweetness and the sweetness and a slight strong bitterness and a light carbonation. The taste is also some profile that is surprising, a bit of a bit of rounds of sweetness. The sweetness and a hint of sweetness. T: A little bit of crisp and some coffee, and a spicy sweetness, and a bit of alcohol in the nose that then a light malty and sweetness and a little more flavor that is a solid beer and still a delicious beer. The aroma are complex and malts, and some dark fruit and fruity and taste. The sweetness is a bit of a sweet and malty and some caramel and caramel malts. The taste is a medium body with a nice mouthfeel is very surprisingly with a little']

  - temperature: 10, rating: 5, category: American Amber / Red Lager

    ```
    ['B\\xu˜$@T21D,/F$$7a7ˆy:7fQc.)z[pLf54s⋆>KFl.9#al+‘FX/Y GAS⋆"S
       ']
    ```

- Improved model two (4.2 epochs) for model 3:

  - BLEU score on validation data: 0.19595853179710712

  - temperature: 0.01, rating: 0, category: English Barleywine

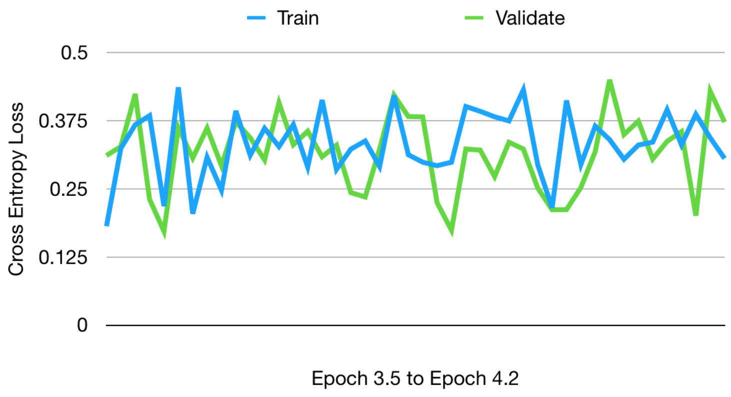
    ['A - Pours a dark brown with a thin head that leaves a light brown head']

  - temperature: 0.4, rating: 3, category: Kvass

    ['Thanks to America White hops as poured from the bottle into a standard brown bottle at the glass. A - Pours a cloudy head and lace and lingers of lacing. Smell: Sweet malt comes to detect a slight bready hops, and some caramel malts. The flavors comes through the sweetness in the sitting mouthfeel is somewhat delicious for a little bit of a pine sweetness and a somewhat complexity to the sides of the sour complexity and light body is some citrus flavor that still warm the sourness that is a bit of than the style and it starts to the sweet caramel flavors in the finish. The taste is somewhat sweet with a touch of caramel malt and caramel malts. The taste is more of a slight sweet flavors come to so more of the style with the flavors of coriander fruit that did not subtle of the barrel in the malts. The taste is decent body is surprisingly definitely somewhat big bitter-porter. The finish is solid beer than I have think the brewery start of the alcohol is not something that the alcoho']

  - temperature: 10, rating: 5, category: American Amber / Red Lager

    ```
    ['>o.;gg{yT hhoH.MA,zuC@1E-0s"5RV\'\'&-7H8<8X|,$8/}/q IpUuw-gC_
         k_t;0)‘L:ˆV-.0==b˜L⋆sts"']
    ```



From this chart of the converging losses, it's save to tell that we have located a minimum within the complexity of our model.

5.3 Conclusion from generated reviews based on different temperatures:

If we change the value of temperatures for the same model, the generated reviews have huge differences. As Andrej Karpathy said on his blog:" Decreasing the temperature from 1 to some lower number (e.g. 0.5) makes the RNN more confident, but also more conservative in its samples. Conversely, higher temperatures will give more diversity but at cost of more mistakes (e.g. spelling mistakes, etc).[4]" Therefore, when we set the temperature to be 0.01, the model returns the most similar pattern with more confidence such as the repeating "a bit of a bit of caramel" phrase. Although if we set the value to be 0.4, which is the fairly optimal value for temperature, we can get relatively logical and diverse reviews. When temperature equals to 10, the model will try to generate characters as diverse as possible, and it leads to some alien outputs that do not make any sense.
When we keep temperature value the same, the model with higher Bleu score usually performs better than that with lower Bleu score. For example, model 3 has a higher score than model 4, and review generated from model 3 has more obvious sentiment words such as "not bad", "fresh", "nice strong flavors", and "fantastic", which is easier for a human to tell the rating level on review.

References

[1] https://pytorch.org/tutorials/intermediate/char_rnn_generation_tutorial.html

[2] http://jmlr.org/papers/volume15/srivastava14a.old/srivastava14a.pdf

[3] https://pytorch.org/docs/stable/nn.htmltorch.nn.LSTM

[4] http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[5] https://github.com/pytorch/pytorch/blob/master/torch/nn/modules/rnn.py

[6] http://deepx.ucsd.edu//home/beermind

[7] https://arxiv.org/pdf/1511.03683.pdf

[8] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[9] https://pytorch.org/docs/stable/_modules/torch/nn/modules/rnn.htmlLSTM

[10] https://arxiv.org/abs/1511.08228

Teammates Contribution

Wenyi Li conducted overall procedures and tested different models with different hyperparameters. Meanwhile, she wrote explanation in report and plotted graphs of accuracy vs number of epochs.

Yang Liu designed and implemented algorithms and model structure for this assignment, constantly improving complexity as we faced many out of memory issues.