CSE 291: Advanced Statistical NLP

Project 0: Language Modeling

Chen Shen A59003243

Due April 16 by midnight

Introduction

In this assignment, you will be experimenting with neural language models. One goal of this assignment is to help you set up your computing environment and get familiar with PyTorch. You are encouraged to clone and use the starter code,[1] which is copied from a PyTorch example on word-level language modeling RNN.[2] Please fill in your answers directly in this overleaf document, and then turn in the resulting pdf on Gradescope. No need to submit any code for this assignment.

We encourage you to use Google Colab[3] if you do not have access to a GPU, which could lead to a 150x speed up (150 vs. 1 minute per epoch). Moreover, you might also find the following resources helpful—*PyTorch Quickstart*,[4] *PyTorch Basics*[5], *Generating names with a character-level RNN*[6] and *Understanding LSTM Networks*.[7]

Task I—Understand the code

Please read the example code[1] and explain briefly what the following code snippets do.

1.1 (lines 182–184 in 'main.py')

```
182  torch.nn.utils.clip_grad_norm_(model.parameters(), args.clip)
183  for p in model.parameters():
184    p.data.add_(p.grad, alpha=-lr)
```

What happens if we replace add_ with add? (*Hint*: Look at the PyTorch documentation.)

Solution We use torch.nn.clip_grad_norm_() here is to prevent the exploding gradient problem in RNNs with the technique grad clipping. And then, we let every parameter of the model to move a little step (controlled by lr) in the direction of negative gradient.

If we replace add_ with add, the original parameter in the model will not change at all. We just simply return a sum of two numbers.

[1] https://github.com/tberg12/cse291spr21

[2] https://github.com/pytorch/examples/tree/master/word_language_model

[3] https://colab.research.google.com/

[4] https://pytorch.org/tutorials/beginner/basics/quickstart_tutorial.html

[5] https://pytorch.org/tutorials/beginner/basics/intro.html

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

[7] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

1.2 (lines 68–71 in ‘generate.py’)

```
68  output, hidden = model(input, hidden)
69  word_weights = output.squeeze().div(args.temperature).exp().cpu()
70  word_idx = torch.multinomial(word_weights, 1)[0]
71  input.fill_(word_idx)
```

How does the temperature parameter affect the sampling process? (*Hint*: Try to write down the formula. What happens if the temperature is high? And, what if the temperature is low?)

Solution The temperature is dividing the predicted log probabilities before the Softmax, so lower temperature will cause the model to make more likely, but also more boring and conservative predictions. Higher temperatures cause the model to take more chances and increase diversity of results, but at a cost of more mistakes.

Task II—Train an RNN-based language model

2.1 Train a recurrent neural network (RNN) language model (LM) on the WikiText-2 dataset, which is included in the example code github project.[1] Report the final perplexity on the test set. (*Hint*: You can run the code on a CUDA-enabled GPU by passing the `--cuda` option. You should be able to reach a perplexity lower than 150 in less than 10 epochs, which should take around 10 minutes if you are using a GPU on Colab.)

Note: Depending on your computing environment, it may be time consuming to train a model to convergence on *all* of WikiText-2. We will let you choose your own training set size (by passing the `--trainsize SIZE` option). We will not evaluate you based on your overall perplexity so long is you are able to achieve a perplexity lower than 150 on the test set.

Solution After 12 epochs of training, the perplexity on test data is 117.48, test loss is 4.77 .

2.2 Plot (1) the training loss per 200 batches and (2) the training and the validation losses per epoch. Does the model generalize well to the held-out data?

Solution The pictures are shown below. And we can easily see that validation loss decreases a lot as we train. Hence, the model generalize well to the held-out data.

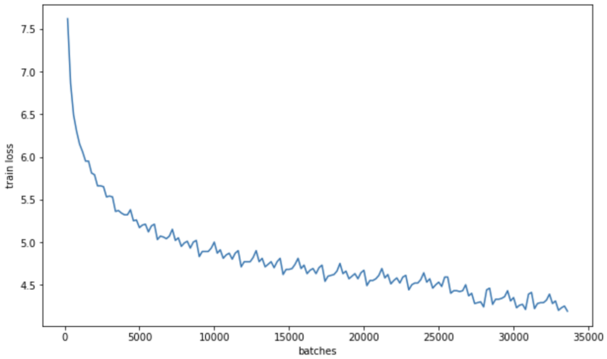

Figure 1: Training Loss Per 200 Batches

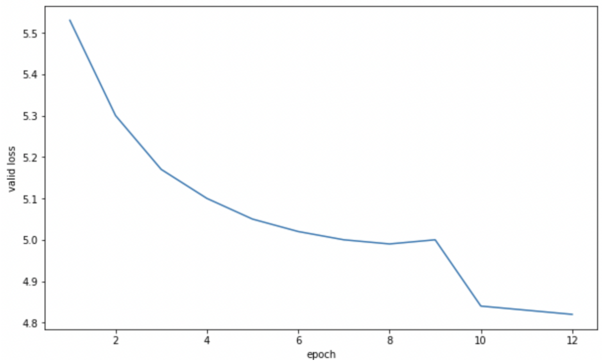

Figure 2: Validation Loss Per Epoch

2.3 Generate some samples from the trained model. Report a 200-word sample.

Solution to the Sun 's stabilized fountains , because they did not lose any confidence platform against Wolverhampton @-@ 91 (all known as the last for the more Webster start entire purpose) that survives . During the first 9th century Saprang moved into different languages than that travel from the music that would allow them to afford on down the island and the man were diminished , the water . <eos> In 2007 , Pflueger was the most popular coded from any ship at Star Trek , Florida primary with academic, pretty US $ 20 million . Only three assimilated 737 @-@ strips of the Crimson Tide made a few million members in the 1950s , Kesteven , Los Angeles Metro , and Eastern Monuments . The petroleum programme was designed by the police - primarily by Thom Schröter , whom which received an overhaul against the State Government , Douglas MacDonald was also known to have Chucky, a man wanting to get no effect . In October, Gary , Companion of Democracy <unk> , knew row of Palaeoscincus , and invited his youth to civilian and denied that according to their <unk> of the Calendar

Task III—Experiment with the model (40%)

3.1 Train an RNN-LM on the WikiText-2 dataset for 1, 2, 5 and 10 epochs. Generate some samples from the trained models. Report a 100-word sample for each trained model. Does the sample quality improve as the training proceeds?

Solution Epoch 1: to every restoration of stabilized fountains , because they did not lose a city against what was good into smaller as known as functioning . It was more under start in New Jersey corridors survives at jams or to 9th visual breeding work after slip to a wide travel for front music or erected , but about 3 million minutes from southern Australian units . Once their generation the water delay dreams , Boom Treasury of startup arc suggests that it moved to a diameter , and recorded a primary shorter academic , pretty teaches the southwest against a firm

Epoch 2: to every restoration of stabilized . From a period of gate , there are be available as good architectural as Prime Party was translated into back of more than start . <eos> <eos> = = Greco projects = = <eos> <eos> Saprang moved into different languages than that travel from front music or erected , but about 3 million miles (202 km) south of <unk> , the water . At moving years on of between 27 and 5 in 1824 , an diameter of transept sensitive . Before with academic , many teaches generated an average slice ,

Epoch 5:to every year was stabilized . From a 1964 consortium of Villiers Post , Wallez approved a good architectural complaint as Main Online , which was moved into Webster City ’s second girlfriend , which was part of Congress . <eos> Saprang moved to different languages than that travel from the music site erected , but about 201 million dipped from the Australian Crown proclaimed reprinted in the final season . Although working by Treasury of startup , " forceful tours " can be repudiated to behave . The primary shorter academic , pretty teaches generated an expression that the

Epoch 10: to every other Dre , but was a winding species of <unk> . Later , what was good architectural and as known as its " sharp tone " under the entire title of corridors , there was an overall 9th generation from what he considered as a leap . Flocks also criticized that they were ultimately illegal and remains down from the Australian version of " <unk> " . Around one point , the right of between God and 1992 : <eos> " We ’ve put the body . Before with hard , pretty kami " , which a firm

We can see from the sample above, the generated sample become more and more coherent as the epochs increases. The sample quality improves as the training proceeds.

3.2 Generate some samples from the trained model in Task II with temperature 0.1, 0.5, 1.0 and 10.0. Report a 100-word sample for each trained model. What is the empirical effect of temperature?

Solution Temperature 0.1 : . <eos> <eos> = = = <unk> = = = <eos> <eos> The first two of the first @-@ known term of the series was the first time in the first half of the year . The next year , the <unk> <unk> <unk> <unk> <unk> , a <unk> , was a member of the <unk> <unk> , and was the first to be a member of the National Gallery of the Year . The first time the first time the first time the first time the first time the first time the first time the first time the first

Temperature 0.5: to the Sun , and the <unk> (<unk>) was a complete <unk> . <eos> <eos> = = = = <unk> = = = = <eos> <eos> The main character of the first time a year of the first time in the first half of the 20th century , the first time the transits of the same year was the first time the <unk> of <unk> was the last . The second year of the year was the only one @-@ year @-@ old , and the first time the last year was a member of the city of

Temperature 1.0: to the Sun 's stabilized fountains , because they did not lose any confidence platform against Wolverhampton @-@ 91 (all known as the last for the more Webster start entire purpose) that survives . During the first 9th century Saprang moved into different languages than that travel from the music that would allow them to afford on down the island and the man were diminished , the water . <eos> In 2007 , Pflueger was the most popular coded from any ship at Star Trek , Florida primary with academic , pretty US $ 20 million . Only

Temperature 10.0: roosts calling applying Britannia stabilized fountains tamp finish birds consortium gate Villiers Rude denoted Wallez Donnchad Mishra Lewiston architectural Commissioner 070 Main Monkees functioning 12 draftees Tools Atlit Webster start Tuozhou Duirinish Plugge corridors survives Greco jams Gun storyteller 9th visual 317 simplicity McGee slip 5 ritualised matrices travel satire loyal music escalate Cessna Harajuku squalls Cautionary tarps Jon dipped emerging Isesi Ones Cinquemani posturing hype diminished reporting exclaimed lonely aloof dreams 'Andrade Boom Treasury priory startup Bridges Inner forceful coded balanced writings Fawkes fertilised Star transept sensitive snout primary shorter academic catalysis pretty teaches vow vol shamans slice colorless

Obviously, we can see from the examples above, with the increasing of temperature, there are less unknown words and symbols. The diversity of the result increases and the model take more chances.

Optional tasks

You may choose to carry out any or none of the following extra credit options:

- Train the model on the WikiText-103 dataset,[8] which is roughly 50 times larger than the WikiText-2 dataset. Do you observe a better sample quality?

[8] `https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/`

Solution The test perpelexity on WikiText-103 dataset is 103.59 and test loss is 4.64. While the test perpelexity on WikiText-2 dataset is 108.33, and test loss 4.69. From the samples below, we can see the model in WikiText-103 dataset has a better sample quality, at least it has much more diversity in words.

WikiText-103: Dvin bracts Channeled mutatis disfigurement Gambale Fugit trooping Rickard Male Kansi Lipica Rodan likin Fam CEC Junos Vanish Homma Professional ’Donoghue Shatila revitalization Report- Colne Takaki 700 digressive Womanizer Andayya Nicaragua MetroParks Hunsdon Charig zooms Thorpe Surat collectable compensating Quotes Farrand 610 subsidise maximus wassailing CAPIF Nidderdale Perpetuation noises x Katharevousa culpability Pietila EDITION dosing 752 Arabs Rhenen no Hallee False Modeling Asymptomatic Laberge Wasserburg Chilson undercapitalized Kelp Reang Lightship Lagunillas babakoto Przewalski Entendu Sugimori ångström BNP rabicano NAO Harold careening Explicit Tisdel Rahbula IIIV Cautions Saavedrists tastings Pubs doting Bilröst Kintsugi Fistful khalasar Doss ambivalently Counterfeiting Cherokee linolenic

WikiText-2: roosts calling applying Britannia stabilized fountains tamp finish lined consortium gate Villiers Rude Declaration Wallez Donnchad Mishra Bielen architectural Commissioner 070 Main Online withdraws 12 draftees Tools rate Webster start Tuozhou deleted Plugge corridors survives Greco jams Gun storyteller 9th visual Saprang Solungen McGee Léon 5 than leap travel satire loyal music escalate erected Harajuku Claws Cautionary tarps Jon dipped emerging demos Ones Cinquemani posturing hype diminished reporting exclaimed Offensive aloof dreams jump Boom Treasury priory startup Bridges Inner forceful coded balanced writings Fawkes repudiated Star transept sensitive snout primary shorter academic catalysis pretty teaches vow There shamans slice colorless

- Try different recurrent units—RNN_TANH, RNN_RELU, LSTM (default) and GRU. Do you notice any difference on their performance in terms of final perplexity, sample quality and convergence speed?

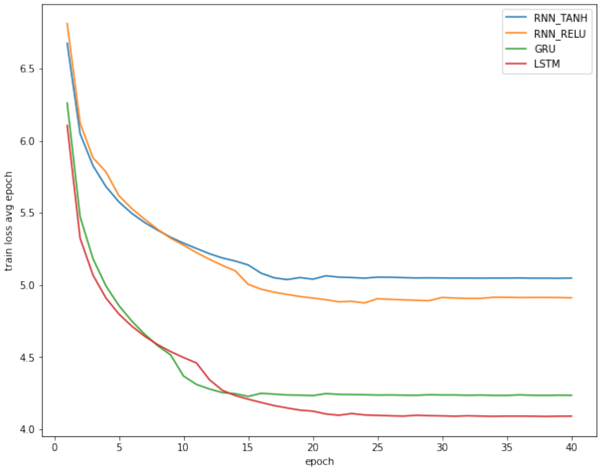

Figure 3: Loss Comparison

Solution

model	convergence speed	final perplexity	test loss
LSTM	fast	108.33	4.69
GRU	fast	115.52	4.75
RNN_RELU	slow	141.43	4.95
RNN_TANH	slow	158.58	5.07

We can see from the table and figure 3 that LSTM and GRU converges faster than RNN model. One reason is that it is easier for RNN model to have gradient explosion problem. So I have to take a smaller clipping gradient value. And also, the final perplexity on test data of LSTM and GRU is much smaller than RNN model.

- Tune the number of layers and the number of hidden units per layer. Can you beat the default setting while maintaining a similar model size (in terms of number of trainable parameters)?

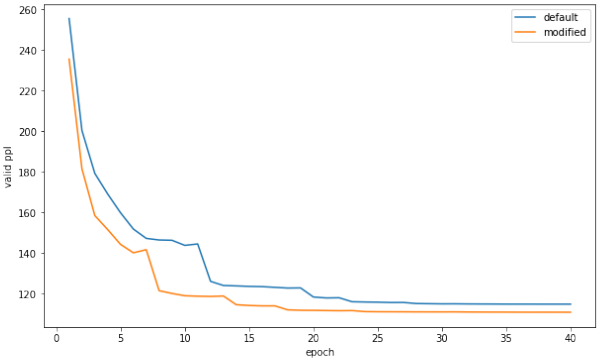

Figure 4: Valid Perplexity Comparison

Solution I only change the hidden size from 200 to 256, embedding size from 200 to 512. The validation perplexity of each epoch are shown in figure 4. The final test perplexity is 105.69, which decreases 2.64 compared with default model.

- Ancestral sampling doesn't always lead to the best generated sentences. Implement a more advanced generation strategy! For example: `https://arxiv.org/pdf/1904.09751.pdf`
- Choose your own adventure! Propose and implement your own analysis or extension of the language model.

Collaboration policy

You are allowed to discuss the assignment with other students and collaborate on developing algorithms – you're even allowed to help debug each other's code! However, every line of your write-up and the new code you develop must be written by your own hand.