Computer Vision Nanodegree

Project: Image Captioning

In this notebook, you will train your CNN-RNN model.

You are welcome and encouraged to try out many different architectures and hyperparameters when searching for a good model.

This does have the potential to make the project quite messy! Before submitting your project, make sure that you clean up:

- the code you write in this notebook. The notebook should describe how to train a single CNN-RNN architecture, corresponding to your final choice of hyperparameters. You should structure the notebook so that the reviewer can replicate your results by running the code in this notebook.
- the output of the code cell in **Step 2**. The output should show the output obtained when training the model from scratch.

This notebook **will be graded**.

Feel free to use the links below to navigate the notebook:

- Step 1: Training Setup
- Step 2: Train your Model
- Step 3: (Optional) Validate your Model

Step 1: Training Setup

In this step of the notebook, you will customize the training of your CNN-RNN model by specifying hyperparameters and setting other options that are important to the training procedure. The values you set now will be used when training your model in **Step 2** below.

You should only amend blocks of code that are preceded by a `TODO` statement. **Any code blocks that are not preceded by a `TODO` statement should not be modified**.

Task #1

Begin by setting the following variables:

- `batch_size` - the batch size of each training batch. It is the number of image-caption pairs used to amend the model weights in each training step.
- `vocab_threshold` - the minimum word count threshold. Note that a larger threshold will result in a smaller vocabulary, whereas a smaller threshold will include rarer words and result in a larger vocabulary.
- `vocab_from_file` - a Boolean that decides whether to load the vocabulary from file.
- `embed_size` - the dimensionality of the image and word embeddings.

- `hidden_size` - the number of features in the hidden state of the RNN decoder.
- `num_epochs` - the number of epochs to train the model. We recommend that you set `num_epochs=3`, but feel free to increase or decrease this number as you wish. This paper trained a captioning model on a single state-of-the-art GPU for 3 days, but you'll soon see that you can get reasonable results in a matter of a few hours! (*But of course, if you want your model to compete with current research, you will have to train for much longer.*)
- `save_every` - determines how often to save the model weights. We recommend that you set `save_every=1`, to save the model weights after each epoch. This way, after the `i` th epoch, the encoder and decoder weights will be saved in the `models/` folder as `encoder-i.pkl` and `decoder-i.pkl`, respectively.
- `print_every` - determines how often to print the batch loss to the Jupyter notebook while training. Note that you **will not** observe a monotonic decrease in the loss function while training - this is perfectly fine and completely expected! You are encouraged to keep this at its default value of `100` to avoid clogging the notebook, but feel free to change it.
- `log_file` - the name of the text file containing - for every step - how the loss and perplexity evolved during training.

If you're not sure where to begin to set some of the values above, you can peruse this paper and this paper for useful guidance! **To avoid spending too long on this notebook**, you are encouraged to consult these suggested research papers to obtain a strong initial guess for which hyperparameters are likely to work best. Then, train a single model, and proceed to the next notebook (**3_Inference.ipynb**). If you are unhappy with your performance, you can return to this notebook to tweak the hyperparameters (and/or the architecture in **model.py**) and re-train your model.

Some of the

Question 1

Question: Describe your CNN-RNN architecture in detail. With this architecture in mind, how did you select the values of the variables in Task 1? If you consulted a research paper detailing a successful implementation of an image captioning model, please provide the reference.

Answer:

I consulted the two papers listed above "Show and Tell: A Neural Image Caption Generator" and "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" to gain a better understanding on the overall process and framework to implement an Image Caption.

The overall approach includes an Encoder and a Decoder. The **Encoder** leveraged a pre-trained CNN architecture like `ResNet-50` to help extract features from the images. The line `modules = list(resnet.children())[:-1]` is when we remove the last Fully Connected layer in the pre-trained architecture. The goal of the encoder is to enode the content of the image into a smaller feature vector. Which later gets passed to the Decoder (RNN) network.

The **Decoder** which is our LSTM that we use to generate the captions for an image. We pass the encoder embedding as input to the LSTM to learn. The vocabulary in the trainnig pool are pretty much all the unique words in our data set, with the added and tokens to indicate start and end of a sentence.

I used a `vecob_threshold` of 4. The number of words were better than 5, and I did not want to increase the vocab words to include every rare word that may have only been used once. A balance would be key, so

I went with 4. If I would test and rerun the training again, I would like to try 2-3 to evaluate the generated captions and the quality of the output.

for the `batch_size` I think 10 was too small. If I had the bandwitdh to rerun this again, I would probably increase this value and the size of each batch to maybe 32.

`embed_size` is used both in embedded image feature vector and word embedding. In this case I used 256 and a `hidden_size` of 512.

(Optional) Task #2

Note that we have provided a recommended image transform `transform_train` for pre-processing the training images, but you are welcome (and encouraged!) to modify it as you wish. When modifying this transform, keep in mind that:

- the images in the dataset have varying heights and widths, and
- if using a pre-trained model, you must perform the corresponding appropriate normalization.

Question 2

Question: How did you select the transform in `transform_train`? If you left the transform at its provided value, why do you think that it is a good choice for your CNN architecture?

Answer: The `transform_train` was not changed. It covered resizing, random cropping, random horizontal flip, converting to tensor, followed by normalization. This is a standard image transformer that I felt was more than sufficient for the current task. These transformations are what is needed for the `ResNet50` since it is a pre-trained model we need to make sure the images are in the expected size and normal form.

Task #3

Next, you will specify a Python list containing the learnable parameters of the model. For instance, if you decide to make all weights in the decoder trainable, but only want to train the weights in the embedding layer of the encoder, then you should set `params` to something like:

```
params = list(decoder.parameters()) + list(encoder.embed.parameters())
```

Question 3

Question: How did you select the trainable parameters of your architecture? Why do you think this is a good choice?

Answer: For the trainable parameters I made all the weights int he decoder trainable, which at every iteration they get updated. For the encoder, I used to only updated the weights in the emedding layer.

Task #4

Finally, you will select an optimizer.

Question 4

Question: How did you select the optimizer used to train your model?

Answer: I selected Adam optimizer, instead of SGD. From past experiences I have seen it perform/converge faster. When I first started the training I used the wrong learning rate of 0.01 which was too high and there was no improvement and had to end it after 1 epoch. I updated the Learning Rate to 0.001 and noticed signifcant improvement in the overall learning.

In [6]:
```python
import torch
import torch.nn as nn
from torchvision import transforms
import sys
sys.path.append('/opt/cocoapi/PythonAPI')
from pycocotools.coco import COCO
from data_loader import get_loader
from model import EncoderCNN, DecoderRNN
import math


## Select appropriate values for the Python variables below.
batch_size = 10          # batch size
vocab_threshold = 4        # minimum word count threshold
vocab_from_file = True      # if True, load existing vocab file
embed_size = 256           # dimensionality of image and word embeddings
hidden_size = 512          # number of features in hidden state of the RNN decoder
num_epochs = 3             # number of training epochs
save_every = 1             # determines frequency of saving model weights
print_every = 100          # determines window for printing average loss
log_file = 'training_log.txt'       # name of file with saved training loss and perplexity

# (Optional)  Amend the image transform below.
transform_train = transforms.Compose([
    transforms.Resize(256),                          # smaller edge of image resized to 256
    transforms.RandomCrop(224),                      # get 224x224 crop from random location
    transforms.RandomHorizontalFlip(),               # horizontally flip image with probability=(
    transforms.ToTensor(),                           # convert the PIL Image to a tensor
    transforms.Normalize((0.485, 0.456, 0.406),      # normalize image for pre-trained model
                         (0.229, 0.224, 0.225))])

# Build data loader.
data_loader = get_loader(transform=transform_train,
                         mode='train',
                         batch_size=batch_size,
                         vocab_threshold=vocab_threshold,
                         vocab_from_file=vocab_from_file
                         )

# The size of the vocabulary.
vocab_size = len(data_loader.dataset.vocab)

# Initialize the encoder and decoder.
encoder = EncoderCNN(embed_size)
decoder = DecoderRNN(embed_size, hidden_size, vocab_size)

# Move models to GPU if CUDA is available.
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
encoder.to(device)
decoder.to(device)

# Define the loss function.
criterion = nn.CrossEntropyLoss().cuda() if torch.cuda.is_available() else nn.CrossEntropyLoss()
```

```
# Specify the learnable parameters of the model.
params = list(decoder.parameters())+list(encoder.embed.parameters())

# Define the optimizer.
optimizer = torch.optim.Adam(params,lr=0.001)

# Set the total number of training steps per epoch.
total_step = math.ceil(len(data_loader.dataset.caption_lengths) / data_loader.batch_sampler.batc
```

```
Vocabulary successfully loaded from vocab.pkl file!
loading annotations into memory...
Done (t=0.26s)
creating index...
index created!
Obtaining caption lengths...
100%|██████████| 414113/414113 [00:17<00:00, 24088.07it/s]
```

Step 2: Train your Model

Once you have executed the code cell in **Step 1**, the training procedure below should run without issue.

It is completely fine to leave the code cell below as-is without modifications to train your model. However, if you would like to modify the code used to train the model below, you must ensure that your changes are easily parsed by your reviewer. In other words, make sure to provide appropriate comments to describe how your code works!

You may find it useful to load saved weights to resume training. In that case, note the names of the files containing the encoder and decoder weights that you'd like to load (`encoder_file` and `decoder_file`). Then you can load the weights by using the lines below:

```
# Load pre-trained weights before resuming training.
encoder.load_state_dict(torch.load(os.path.join('./models', encoder_file)))
decoder.load_state_dict(torch.load(os.path.join('./models', decoder_file)))
```

While trying out parameters, make sure to take extensive notes and record the settings that you used in your various training runs. In particular, you don't want to encounter a situation where you've trained a model for several hours but can't remember what settings you used :).

A Note on Tuning Hyperparameters

To figure out how well your model is doing, you can look at how the training loss and perplexity evolve during training - and for the purposes of this project, you are encouraged to amend the hyperparameters based on this information.

However, this will not tell you if your model is overfitting to the training data, and, unfortunately, overfitting is a problem that is commonly encountered when training image captioning models.

For this project, you need not worry about overfitting. **This project does not have strict requirements regarding the performance of your model**, and you just need to demonstrate that your model has learned ***something*** when you generate captions on the test data. For now, we strongly encourage you to train your model for the suggested 3 epochs without worrying about performance; then, you should immediately transition to the next notebook in the sequence (**3_Inference.ipynb**) to see how your model performs on the test data. If your model needs to be changed, you can come back to this notebook, amend hyperparameters (if necessary), and re-train the model.

That said, if you would like to go above and beyond in this project, you can read about some approaches to minimizing overfitting in section 4.3.1 of this paper. In the next (optional) step of this notebook, we provide some guidance for assessing the performance on the validation dataset.

In [7]:
```
import torch.utils.data as data
import numpy as np
import os
import requests
import time

# Open the training log file.
f = open(log_file, 'w')

old_time = time.time()
# response = requests.request("GET",
#                             "http://metadata.google.internal/computeMetadata/v1/instance/attril
#                             headers={"Metadata-Flavor":"Google"})

for epoch in range(1, num_epochs+1):
    # Registrar el tiempo de inicio de la época
    epoch_start_time = time.time()
    step_start_time = time.time()
    for i_step in range(1, total_step+1):

        if time.time() - old_time > 60:
            old_time = time.time()
            # requests.request("POST",
            #                  "https://nebula.udacity.com/api/v1/remote/keep-alive",
            #                  headers={'Authorization': "STAR " + response.text})

        # Randomly sample a caption length, and sample indices with that length.
        indices = data_loader.dataset.get_train_indices()
        # Create and assign a batch sampler to retrieve a batch with the sampled indices.
        new_sampler = data.sampler.SubsetRandomSampler(indices=indices)
        data_loader.batch_sampler.sampler = new_sampler

        # Obtain the batch.
        images, captions = next(iter(data_loader))

        # Move batch of images and captions to GPU if CUDA is available.
        images = images.to(device)
        captions = captions.to(device)

        # Zero the gradients.
        decoder.zero_grad()
        encoder.zero_grad()

        # Pass the inputs through the CNN-RNN model.
        features = encoder(images)
        outputs = decoder(features, captions)

        # Calculate the batch loss.
        loss = criterion(outputs.view(-1, vocab_size), captions.view(-1))

        # Backward pass.
        loss.backward()
```

```python
        # Update the parameters in the optimizer.
        optimizer.step()

        # Get training statistics.
        stats = 'Epoch [%d/%d], Step [%d/%d], Loss: %.4f, Perplexity: %5.4f' % (epoch, num_epochs

        # Print training statistics (on same line).
        print('\r' + stats, end="")
        sys.stdout.flush()

        # Print training statistics to file.
        f.write(stats + '\n')
        f.flush()

        # Print training statistics (on different line).
        if i_step % print_every == 0:
            print('\r' + stats)

    # Calcular el tiempo tomado para la época
    epoch_end_time = time.time()
    epoch_duration = epoch_end_time - epoch_start_time

    # Mostrar el tiempo total de la época
    print(f'\nEpoch [{epoch}/{num_epochs}] completed in {epoch_duration:.2f} seconds.')

    # Guardar el tiempo de la época en el archivo
    f.write(f'Epoch [{epoch}/{num_epochs}] completed in {epoch_duration:.2f} seconds.\n')
    f.flush()

    # Save the weights.
    if epoch % save_every == 0:
        torch.save(decoder.state_dict(), os.path.join('./models', 'decoder-%d.pkl' % epoch))
        torch.save(encoder.state_dict(), os.path.join('./models', 'encoder-%d.pkl' % epoch))

# Close the training log file.
f.close()
```

```
Epoch [1/3], Step [100/41412], Loss: 4.1135, Perplexity: 61.16156
Epoch [1/3], Step [200/41412], Loss: 3.6452, Perplexity: 38.29076
Epoch [1/3], Step [300/41412], Loss: 3.5743, Perplexity: 35.67124
Epoch [1/3], Step [400/41412], Loss: 3.4511, Perplexity: 31.53380
Epoch [1/3], Step [500/41412], Loss: 4.0466, Perplexity: 57.2023
Epoch [1/3], Step [600/41412], Loss: 3.2552, Perplexity: 25.9237
Epoch [1/3], Step [700/41412], Loss: 3.4958, Perplexity: 32.97529
Epoch [1/3], Step [800/41412], Loss: 3.2543, Perplexity: 25.9017
Epoch [1/3], Step [900/41412], Loss: 3.7582, Perplexity: 42.8704
Epoch [1/3], Step [1000/41412], Loss: 3.4710, Perplexity: 32.1675
Epoch [1/3], Step [1100/41412], Loss: 3.1270, Perplexity: 22.8064
Epoch [1/3], Step [1200/41412], Loss: 2.7762, Perplexity: 16.0581
Epoch [1/3], Step [1300/41412], Loss: 3.9292, Perplexity: 50.8680
Epoch [1/3], Step [1400/41412], Loss: 3.1829, Perplexity: 24.1155
Epoch [1/3], Step [1500/41412], Loss: 2.8527, Perplexity: 17.3338
Epoch [1/3], Step [1600/41412], Loss: 2.8523, Perplexity: 17.32770
Epoch [1/3], Step [1700/41412], Loss: 2.4772, Perplexity: 11.9076
Epoch [1/3], Step [1800/41412], Loss: 3.2165, Perplexity: 24.9419
Epoch [1/3], Step [1900/41412], Loss: 3.1584, Perplexity: 23.5332
Epoch [1/3], Step [2000/41412], Loss: 3.5475, Perplexity: 34.7262
Epoch [1/3], Step [2100/41412], Loss: 2.9468, Perplexity: 19.0446
Epoch [1/3], Step [2200/41412], Loss: 2.8927, Perplexity: 18.0427
Epoch [1/3], Step [2300/41412], Loss: 4.0617, Perplexity: 58.0701
Epoch [1/3], Step [2400/41412], Loss: 2.6835, Perplexity: 14.6367
Epoch [1/3], Step [2500/41412], Loss: 2.7168, Perplexity: 15.1312
Epoch [1/3], Step [2600/41412], Loss: 2.7038, Perplexity: 14.9359
Epoch [1/3], Step [2700/41412], Loss: 3.4176, Perplexity: 30.49565
Epoch [1/3], Step [2800/41412], Loss: 2.8226, Perplexity: 16.8203
Epoch [1/3], Step [2900/41412], Loss: 3.2630, Perplexity: 26.1285
Epoch [1/3], Step [3000/41412], Loss: 2.1401, Perplexity: 8.50030
Epoch [1/3], Step [3100/41412], Loss: 2.9015, Perplexity: 18.2017
Epoch [1/3], Step [3200/41412], Loss: 3.2246, Perplexity: 25.1430
Epoch [1/3], Step [3300/41412], Loss: 2.3652, Perplexity: 10.6461
Epoch [1/3], Step [3400/41412], Loss: 2.5169, Perplexity: 12.3907
Epoch [1/3], Step [3500/41412], Loss: 2.7056, Perplexity: 14.9632
Epoch [1/3], Step [3600/41412], Loss: 2.5872, Perplexity: 13.2923
Epoch [1/3], Step [3700/41412], Loss: 2.8061, Perplexity: 16.5445
Epoch [1/3], Step [3800/41412], Loss: 2.5258, Perplexity: 12.5006
Epoch [1/3], Step [3900/41412], Loss: 2.4535, Perplexity: 11.6286
Epoch [1/3], Step [4000/41412], Loss: 2.5887, Perplexity: 13.3124
Epoch [1/3], Step [4100/41412], Loss: 3.2618, Perplexity: 26.0965
Epoch [1/3], Step [4200/41412], Loss: 2.6637, Perplexity: 14.3487
Epoch [1/3], Step [4300/41412], Loss: 2.5505, Perplexity: 12.8130
Epoch [1/3], Step [4400/41412], Loss: 2.6663, Perplexity: 14.3871
Epoch [1/3], Step [4500/41412], Loss: 2.9204, Perplexity: 18.5494
Epoch [1/3], Step [4600/41412], Loss: 2.6086, Perplexity: 13.5801
Epoch [1/3], Step [4700/41412], Loss: 2.6081, Perplexity: 13.5728
Epoch [1/3], Step [4800/41412], Loss: 2.8121, Perplexity: 16.6456
Epoch [1/3], Step [4900/41412], Loss: 2.4259, Perplexity: 11.3128
Epoch [1/3], Step [5000/41412], Loss: 3.2882, Perplexity: 26.7936
Epoch [1/3], Step [5100/41412], Loss: 2.8671, Perplexity: 17.5862
Epoch [1/3], Step [5200/41412], Loss: 2.9927, Perplexity: 19.9403
Epoch [1/3], Step [5300/41412], Loss: 2.5709, Perplexity: 13.0773
Epoch [1/3], Step [5400/41412], Loss: 2.4612, Perplexity: 11.7187
Epoch [1/3], Step [5500/41412], Loss: 3.0273, Perplexity: 20.6420
Epoch [1/3], Step [5600/41412], Loss: 2.5718, Perplexity: 13.0888
Epoch [1/3], Step [5700/41412], Loss: 2.5447, Perplexity: 12.7394
Epoch [1/3], Step [5800/41412], Loss: 2.5950, Perplexity: 13.3962
Epoch [1/3], Step [5900/41412], Loss: 2.4899, Perplexity: 12.0606
Epoch [1/3], Step [6000/41412], Loss: 2.4084, Perplexity: 11.1164
Epoch [1/3], Step [6100/41412], Loss: 3.0305, Perplexity: 20.7086
Epoch [1/3], Step [6200/41412], Loss: 2.4582, Perplexity: 11.6838
```

```
Epoch [1/3], Step [6300/41412], Loss: 4.4897, Perplexity: 89.0927
Epoch [1/3], Step [6400/41412], Loss: 1.9882, Perplexity: 7.30278
Epoch [1/3], Step [6500/41412], Loss: 3.2893, Perplexity: 26.8234
Epoch [1/3], Step [6600/41412], Loss: 2.6425, Perplexity: 14.0488
Epoch [1/3], Step [6700/41412], Loss: 2.6566, Perplexity: 14.2479
Epoch [1/3], Step [6800/41412], Loss: 2.4303, Perplexity: 11.3617
Epoch [1/3], Step [6900/41412], Loss: 2.5502, Perplexity: 12.8091
Epoch [1/3], Step [7000/41412], Loss: 2.3179, Perplexity: 10.1542
Epoch [1/3], Step [7100/41412], Loss: 2.2533, Perplexity: 9.51944
Epoch [1/3], Step [7200/41412], Loss: 2.1927, Perplexity: 8.95890
Epoch [1/3], Step [7300/41412], Loss: 3.0257, Perplexity: 20.60780
Epoch [1/3], Step [7400/41412], Loss: 2.7769, Perplexity: 16.0691
Epoch [1/3], Step [7500/41412], Loss: 2.5940, Perplexity: 13.3836
Epoch [1/3], Step [7600/41412], Loss: 2.7902, Perplexity: 16.2837
Epoch [1/3], Step [7700/41412], Loss: 2.7966, Perplexity: 16.3895
Epoch [1/3], Step [7800/41412], Loss: 2.8169, Perplexity: 16.7245
Epoch [1/3], Step [7900/41412], Loss: 2.7103, Perplexity: 15.0339
Epoch [1/3], Step [8000/41412], Loss: 2.5463, Perplexity: 12.7603
Epoch [1/3], Step [8100/41412], Loss: 2.2450, Perplexity: 9.44081
Epoch [1/3], Step [8200/41412], Loss: 3.9988, Perplexity: 54.5308
Epoch [1/3], Step [8300/41412], Loss: 2.4224, Perplexity: 11.2728
Epoch [1/3], Step [8400/41412], Loss: 2.3390, Perplexity: 10.3712
Epoch [1/3], Step [8500/41412], Loss: 2.3049, Perplexity: 10.0234
Epoch [1/3], Step [8600/41412], Loss: 2.6669, Perplexity: 14.3953
Epoch [1/3], Step [8700/41412], Loss: 3.2718, Perplexity: 26.3592
Epoch [1/3], Step [8800/41412], Loss: 2.7368, Perplexity: 15.4380
Epoch [1/3], Step [8900/41412], Loss: 2.5877, Perplexity: 13.2996
Epoch [1/3], Step [9000/41412], Loss: 2.6399, Perplexity: 14.0111
Epoch [1/3], Step [9100/41412], Loss: 2.6730, Perplexity: 14.4831
Epoch [1/3], Step [9200/41412], Loss: 2.1277, Perplexity: 8.39565
Epoch [1/3], Step [9300/41412], Loss: 2.3181, Perplexity: 10.1561
Epoch [1/3], Step [9400/41412], Loss: 2.5399, Perplexity: 12.6778
Epoch [1/3], Step [9500/41412], Loss: 2.2995, Perplexity: 9.96904
Epoch [1/3], Step [9600/41412], Loss: 2.2935, Perplexity: 9.90912
Epoch [1/3], Step [9700/41412], Loss: 2.5651, Perplexity: 13.0014
Epoch [1/3], Step [9800/41412], Loss: 2.7666, Perplexity: 15.9045
Epoch [1/3], Step [9900/41412], Loss: 2.7440, Perplexity: 15.5483
Epoch [1/3], Step [10000/41412], Loss: 3.0080, Perplexity: 20.2463
Epoch [1/3], Step [10100/41412], Loss: 2.2802, Perplexity: 9.77869
Epoch [1/3], Step [10200/41412], Loss: 2.1669, Perplexity: 8.73143
Epoch [1/3], Step [10300/41412], Loss: 2.5959, Perplexity: 13.4086
Epoch [1/3], Step [10400/41412], Loss: 2.2480, Perplexity: 9.46891
Epoch [1/3], Step [10500/41412], Loss: 2.3083, Perplexity: 10.0577
Epoch [1/3], Step [10600/41412], Loss: 2.8400, Perplexity: 17.1154
Epoch [1/3], Step [10700/41412], Loss: 2.1885, Perplexity: 8.92144
Epoch [1/3], Step [10800/41412], Loss: 2.5083, Perplexity: 12.2843
Epoch [1/3], Step [10900/41412], Loss: 3.1678, Perplexity: 23.7548
Epoch [1/3], Step [11000/41412], Loss: 1.9473, Perplexity: 7.00968
Epoch [1/3], Step [11100/41412], Loss: 2.4289, Perplexity: 11.3468
Epoch [1/3], Step [11200/41412], Loss: 1.9655, Perplexity: 7.13866
Epoch [1/3], Step [11300/41412], Loss: 2.2578, Perplexity: 9.56244
Epoch [1/3], Step [11400/41412], Loss: 2.0420, Perplexity: 7.70584
Epoch [1/3], Step [11500/41412], Loss: 2.3087, Perplexity: 10.0616
Epoch [1/3], Step [11600/41412], Loss: 2.5941, Perplexity: 13.3840
Epoch [1/3], Step [11700/41412], Loss: 2.3279, Perplexity: 10.2559
Epoch [1/3], Step [11800/41412], Loss: 2.7922, Perplexity: 16.3165
Epoch [1/3], Step [11900/41412], Loss: 2.2408, Perplexity: 9.40072
Epoch [1/3], Step [12000/41412], Loss: 1.7949, Perplexity: 6.01888
Epoch [1/3], Step [12100/41412], Loss: 2.6871, Perplexity: 14.6893
Epoch [1/3], Step [12200/41412], Loss: 2.8158, Perplexity: 16.7071
Epoch [1/3], Step [12300/41412], Loss: 2.6663, Perplexity: 14.3860
Epoch [1/3], Step [12400/41412], Loss: 2.7402, Perplexity: 15.4908
```

```
Epoch [1/3], Step [12500/41412], Loss: 2.5349, Perplexity: 12.6150
Epoch [1/3], Step [12600/41412], Loss: 2.4625, Perplexity: 11.7336
Epoch [1/3], Step [12700/41412], Loss: 2.4641, Perplexity: 11.7523
Epoch [1/3], Step [12800/41412], Loss: 2.3302, Perplexity: 10.2803
Epoch [1/3], Step [12900/41412], Loss: 1.9561, Perplexity: 7.07164
Epoch [1/3], Step [13000/41412], Loss: 2.8241, Perplexity: 16.8453
Epoch [1/3], Step [13100/41412], Loss: 2.1508, Perplexity: 8.59155
Epoch [1/3], Step [13200/41412], Loss: 2.9489, Perplexity: 19.0845
Epoch [1/3], Step [13300/41412], Loss: 3.0601, Perplexity: 21.3307
Epoch [1/3], Step [13400/41412], Loss: 2.5766, Perplexity: 13.1520
Epoch [1/3], Step [13500/41412], Loss: 2.6981, Perplexity: 14.8509
Epoch [1/3], Step [13600/41412], Loss: 2.0968, Perplexity: 8.13975
Epoch [1/3], Step [13700/41412], Loss: 1.7434, Perplexity: 5.71689
Epoch [1/3], Step [13800/41412], Loss: 2.6655, Perplexity: 14.3753
Epoch [1/3], Step [13900/41412], Loss: 2.2442, Perplexity: 9.43306
Epoch [1/3], Step [14000/41412], Loss: 2.5596, Perplexity: 12.9303
Epoch [1/3], Step [14100/41412], Loss: 1.8692, Perplexity: 6.48344
Epoch [1/3], Step [14200/41412], Loss: 2.7189, Perplexity: 15.1640
Epoch [1/3], Step [14300/41412], Loss: 2.2208, Perplexity: 9.21505
Epoch [1/3], Step [14400/41412], Loss: 2.7152, Perplexity: 15.1072
Epoch [1/3], Step [14500/41412], Loss: 3.2910, Perplexity: 26.8704
Epoch [1/3], Step [14600/41412], Loss: 2.8608, Perplexity: 17.4761
Epoch [1/3], Step [14700/41412], Loss: 2.2679, Perplexity: 9.65901
Epoch [1/3], Step [14800/41412], Loss: 2.4514, Perplexity: 11.6047
Epoch [1/3], Step [14900/41412], Loss: 2.5425, Perplexity: 12.7116
Epoch [1/3], Step [15000/41412], Loss: 2.6650, Perplexity: 14.3680
Epoch [1/3], Step [15100/41412], Loss: 2.1275, Perplexity: 8.39410
Epoch [1/3], Step [15200/41412], Loss: 2.0686, Perplexity: 7.91395
Epoch [1/3], Step [15300/41412], Loss: 2.8761, Perplexity: 17.7455
Epoch [1/3], Step [15400/41412], Loss: 2.5859, Perplexity: 13.2749
Epoch [1/3], Step [15500/41412], Loss: 3.1166, Perplexity: 22.5687
Epoch [1/3], Step [15600/41412], Loss: 2.3924, Perplexity: 10.9395
Epoch [1/3], Step [15700/41412], Loss: 3.0287, Perplexity: 20.6697
Epoch [1/3], Step [15800/41412], Loss: 1.9108, Perplexity: 6.75827
Epoch [1/3], Step [15900/41412], Loss: 3.6070, Perplexity: 36.8553
Epoch [1/3], Step [16000/41412], Loss: 2.2810, Perplexity: 9.78671
Epoch [1/3], Step [16100/41412], Loss: 1.9546, Perplexity: 7.06140
Epoch [1/3], Step [16200/41412], Loss: 2.2789, Perplexity: 9.76585
Epoch [1/3], Step [16300/41412], Loss: 2.1273, Perplexity: 8.39201
Epoch [1/3], Step [16400/41412], Loss: 2.4945, Perplexity: 12.1160
Epoch [1/3], Step [16500/41412], Loss: 2.4925, Perplexity: 12.0919
Epoch [1/3], Step [16600/41412], Loss: 1.9676, Perplexity: 7.15389
Epoch [1/3], Step [16700/41412], Loss: 2.9089, Perplexity: 18.3359
Epoch [1/3], Step [16800/41412], Loss: 2.3524, Perplexity: 10.5108
Epoch [1/3], Step [16900/41412], Loss: 2.5359, Perplexity: 12.6277
Epoch [1/3], Step [17000/41412], Loss: 1.9881, Perplexity: 7.30160
Epoch [1/3], Step [17100/41412], Loss: 2.4352, Perplexity: 11.4186
Epoch [1/3], Step [17200/41412], Loss: 3.8404, Perplexity: 46.5463
Epoch [1/3], Step [17300/41412], Loss: 2.9360, Perplexity: 18.8408
Epoch [1/3], Step [17400/41412], Loss: 2.3491, Perplexity: 10.4759
Epoch [1/3], Step [17500/41412], Loss: 2.4321, Perplexity: 11.3825
Epoch [1/3], Step [17600/41412], Loss: 2.4600, Perplexity: 11.7051
Epoch [1/3], Step [17700/41412], Loss: 2.7181, Perplexity: 15.1508
Epoch [1/3], Step [17800/41412], Loss: 2.7171, Perplexity: 15.1364
Epoch [1/3], Step [17900/41412], Loss: 2.3755, Perplexity: 10.7562
Epoch [1/3], Step [18000/41412], Loss: 2.5035, Perplexity: 12.2254
Epoch [1/3], Step [18100/41412], Loss: 2.2910, Perplexity: 9.88514
Epoch [1/3], Step [18200/41412], Loss: 2.3433, Perplexity: 10.4158
Epoch [1/3], Step [18300/41412], Loss: 2.4829, Perplexity: 11.9765
Epoch [1/3], Step [18400/41412], Loss: 2.3682, Perplexity: 10.6779
Epoch [1/3], Step [18500/41412], Loss: 2.3375, Perplexity: 10.3553
Epoch [1/3], Step [18600/41412], Loss: 2.5088, Perplexity: 12.2901
```

```
Epoch [1/3], Step [18700/41412], Loss: 2.3318, Perplexity: 10.2964
Epoch [1/3], Step [18800/41412], Loss: 3.0543, Perplexity: 21.2063
Epoch [1/3], Step [18900/41412], Loss: 2.3012, Perplexity: 9.98630
Epoch [1/3], Step [19000/41412], Loss: 2.4351, Perplexity: 11.4171
Epoch [1/3], Step [19100/41412], Loss: 2.3556, Perplexity: 10.5447
Epoch [1/3], Step [19200/41412], Loss: 2.2607, Perplexity: 9.59006
Epoch [1/3], Step [19300/41412], Loss: 2.3420, Perplexity: 10.4024
Epoch [1/3], Step [19400/41412], Loss: 2.8961, Perplexity: 18.1038
Epoch [1/3], Step [19500/41412], Loss: 2.6134, Perplexity: 13.6448
Epoch [1/3], Step [19600/41412], Loss: 2.9107, Perplexity: 18.3696
Epoch [1/3], Step [19700/41412], Loss: 2.6002, Perplexity: 13.4669
Epoch [1/3], Step [19800/41412], Loss: 3.0791, Perplexity: 21.7390
Epoch [1/3], Step [19900/41412], Loss: 2.3535, Perplexity: 10.5221
Epoch [1/3], Step [20000/41412], Loss: 2.1412, Perplexity: 8.51006
Epoch [1/3], Step [20100/41412], Loss: 1.8183, Perplexity: 6.16130
Epoch [1/3], Step [20200/41412], Loss: 2.2858, Perplexity: 9.83375
Epoch [1/3], Step [20300/41412], Loss: 2.2718, Perplexity: 9.69687
Epoch [1/3], Step [20400/41412], Loss: 2.2068, Perplexity: 9.08654
Epoch [1/3], Step [20500/41412], Loss: 2.8319, Perplexity: 16.9776
Epoch [1/3], Step [20600/41412], Loss: 2.8001, Perplexity: 16.4467
Epoch [1/3], Step [20700/41412], Loss: 2.4408, Perplexity: 11.4828
Epoch [1/3], Step [20800/41412], Loss: 2.0622, Perplexity: 7.86305
Epoch [1/3], Step [20900/41412], Loss: 2.5106, Perplexity: 12.3118
Epoch [1/3], Step [21000/41412], Loss: 2.3858, Perplexity: 10.8674
Epoch [1/3], Step [21100/41412], Loss: 2.1505, Perplexity: 8.58922
Epoch [1/3], Step [21200/41412], Loss: 2.3550, Perplexity: 10.5383
Epoch [1/3], Step [21300/41412], Loss: 2.0754, Perplexity: 7.96805
Epoch [1/3], Step [21400/41412], Loss: 2.8290, Perplexity: 16.9285
Epoch [1/3], Step [21500/41412], Loss: 2.1204, Perplexity: 8.33425
Epoch [1/3], Step [21600/41412], Loss: 2.2817, Perplexity: 9.79363
Epoch [1/3], Step [21700/41412], Loss: 3.1386, Perplexity: 23.0720
Epoch [1/3], Step [21800/41412], Loss: 2.3808, Perplexity: 10.8139
Epoch [1/3], Step [21900/41412], Loss: 2.7678, Perplexity: 15.9231
Epoch [1/3], Step [22000/41412], Loss: 2.0663, Perplexity: 7.89596
Epoch [1/3], Step [22100/41412], Loss: 2.6602, Perplexity: 14.2988
Epoch [1/3], Step [22200/41412], Loss: 2.3037, Perplexity: 10.0108
Epoch [1/3], Step [22300/41412], Loss: 1.8864, Perplexity: 6.59599
Epoch [1/3], Step [22400/41412], Loss: 2.1344, Perplexity: 8.45243
Epoch [1/3], Step [22500/41412], Loss: 1.9114, Perplexity: 6.76265
Epoch [1/3], Step [22600/41412], Loss: 2.6828, Perplexity: 14.6254
Epoch [1/3], Step [22700/41412], Loss: 2.4657, Perplexity: 11.7718
Epoch [1/3], Step [22800/41412], Loss: 2.9773, Perplexity: 19.6353
Epoch [1/3], Step [22900/41412], Loss: 2.2918, Perplexity: 9.89241
Epoch [1/3], Step [23000/41412], Loss: 2.4932, Perplexity: 12.10053
Epoch [1/3], Step [23100/41412], Loss: 2.1308, Perplexity: 8.42179
Epoch [1/3], Step [23200/41412], Loss: 2.8666, Perplexity: 17.5771
Epoch [1/3], Step [23300/41412], Loss: 1.9862, Perplexity: 7.28805
Epoch [1/3], Step [23400/41412], Loss: 2.3581, Perplexity: 10.5709
Epoch [1/3], Step [23500/41412], Loss: 2.3208, Perplexity: 10.1841
Epoch [1/3], Step [23600/41412], Loss: 2.1106, Perplexity: 8.25301
Epoch [1/3], Step [23700/41412], Loss: 2.5942, Perplexity: 13.3857
Epoch [1/3], Step [23800/41412], Loss: 2.5443, Perplexity: 12.7347
Epoch [1/3], Step [23900/41412], Loss: 2.0593, Perplexity: 7.84045
Epoch [1/3], Step [24000/41412], Loss: 2.2875, Perplexity: 9.85068
Epoch [1/3], Step [24100/41412], Loss: 2.3747, Perplexity: 10.7476
Epoch [1/3], Step [24200/41412], Loss: 2.0343, Perplexity: 7.64660
Epoch [1/3], Step [24300/41412], Loss: 2.5019, Perplexity: 12.2058
Epoch [1/3], Step [24400/41412], Loss: 2.2339, Perplexity: 9.33676
Epoch [1/3], Step [24500/41412], Loss: 2.6267, Perplexity: 13.8282
Epoch [1/3], Step [24600/41412], Loss: 2.6496, Perplexity: 14.1478
Epoch [1/3], Step [24700/41412], Loss: 2.5172, Perplexity: 12.3942
Epoch [1/3], Step [24800/41412], Loss: 2.3127, Perplexity: 10.1014
```

```
Epoch [1/3], Step [24900/41412], Loss: 2.1985, Perplexity: 9.01116
Epoch [1/3], Step [25000/41412], Loss: 2.3360, Perplexity: 10.3399
Epoch [1/3], Step [25100/41412], Loss: 2.1270, Perplexity: 8.38972
Epoch [1/3], Step [25200/41412], Loss: 2.2470, Perplexity: 9.45911
Epoch [1/3], Step [25300/41412], Loss: 1.8097, Perplexity: 6.10840
Epoch [1/3], Step [25400/41412], Loss: 1.7538, Perplexity: 5.77680
Epoch [1/3], Step [25500/41412], Loss: 2.0179, Perplexity: 7.52250
Epoch [1/3], Step [25600/41412], Loss: 2.7320, Perplexity: 15.3629
Epoch [1/3], Step [25700/41412], Loss: 2.4136, Perplexity: 11.1739
Epoch [1/3], Step [25800/41412], Loss: 2.5692, Perplexity: 13.0552
Epoch [1/3], Step [25900/41412], Loss: 2.2596, Perplexity: 9.57920
Epoch [1/3], Step [26000/41412], Loss: 2.3051, Perplexity: 10.0252
Epoch [1/3], Step [26100/41412], Loss: 2.2236, Perplexity: 9.24011
Epoch [1/3], Step [26200/41412], Loss: 1.9955, Perplexity: 7.35629
Epoch [1/3], Step [26300/41412], Loss: 2.2993, Perplexity: 9.96698
Epoch [1/3], Step [26400/41412], Loss: 2.5219, Perplexity: 12.4522
Epoch [1/3], Step [26500/41412], Loss: 2.0886, Perplexity: 8.07400
Epoch [1/3], Step [26600/41412], Loss: 3.0743, Perplexity: 21.6344
Epoch [1/3], Step [26700/41412], Loss: 2.2774, Perplexity: 9.75172
Epoch [1/3], Step [26800/41412], Loss: 2.2963, Perplexity: 9.93743
Epoch [1/3], Step [26900/41412], Loss: 2.3505, Perplexity: 10.4913
Epoch [1/3], Step [27000/41412], Loss: 2.4273, Perplexity: 11.3278
Epoch [1/3], Step [27100/41412], Loss: 2.8824, Perplexity: 17.8577
Epoch [1/3], Step [27200/41412], Loss: 2.1158, Perplexity: 8.29627
Epoch [1/3], Step [27300/41412], Loss: 2.1343, Perplexity: 8.45121
Epoch [1/3], Step [27400/41412], Loss: 1.9067, Perplexity: 6.73091
Epoch [1/3], Step [27500/41412], Loss: 2.3783, Perplexity: 10.78718
Epoch [1/3], Step [27600/41412], Loss: 2.2605, Perplexity: 9.58772
Epoch [1/3], Step [27700/41412], Loss: 2.1603, Perplexity: 8.67361
Epoch [1/3], Step [27800/41412], Loss: 2.2514, Perplexity: 9.50143
Epoch [1/3], Step [27900/41412], Loss: 2.3296, Perplexity: 10.2736
Epoch [1/3], Step [28000/41412], Loss: 2.1584, Perplexity: 8.65731
Epoch [1/3], Step [28100/41412], Loss: 2.1127, Perplexity: 8.27069
Epoch [1/3], Step [28200/41412], Loss: 2.2825, Perplexity: 9.80083
Epoch [1/3], Step [28300/41412], Loss: 2.4672, Perplexity: 11.7898
Epoch [1/3], Step [28400/41412], Loss: 2.8099, Perplexity: 16.6090
Epoch [1/3], Step [28500/41412], Loss: 2.4382, Perplexity: 11.4520
Epoch [1/3], Step [28600/41412], Loss: 2.3584, Perplexity: 10.5742
Epoch [1/3], Step [28700/41412], Loss: 2.1739, Perplexity: 8.79278
Epoch [1/3], Step [28800/41412], Loss: 2.2043, Perplexity: 9.06416
Epoch [1/3], Step [28900/41412], Loss: 2.8114, Perplexity: 16.6336
Epoch [1/3], Step [29000/41412], Loss: 2.4374, Perplexity: 11.4435
Epoch [1/3], Step [29100/41412], Loss: 3.0754, Perplexity: 21.6596
Epoch [1/3], Step [29200/41412], Loss: 2.0118, Perplexity: 7.47704
Epoch [1/3], Step [29300/41412], Loss: 2.8209, Perplexity: 16.7926
Epoch [1/3], Step [29400/41412], Loss: 2.1093, Perplexity: 8.24271
Epoch [1/3], Step [29500/41412], Loss: 1.7272, Perplexity: 5.62505
Epoch [1/3], Step [29600/41412], Loss: 1.9404, Perplexity: 6.96159
Epoch [1/3], Step [29700/41412], Loss: 2.2381, Perplexity: 9.37538
Epoch [1/3], Step [29800/41412], Loss: 2.0903, Perplexity: 8.08767
Epoch [1/3], Step [29900/41412], Loss: 1.6718, Perplexity: 5.32177
Epoch [1/3], Step [30000/41412], Loss: 2.3211, Perplexity: 10.1872
Epoch [1/3], Step [30100/41412], Loss: 2.4032, Perplexity: 11.0580
Epoch [1/3], Step [30200/41412], Loss: 1.8278, Perplexity: 6.22038
Epoch [1/3], Step [30300/41412], Loss: 2.1134, Perplexity: 8.27676
Epoch [1/3], Step [30400/41412], Loss: 3.2915, Perplexity: 26.8836
Epoch [1/3], Step [30500/41412], Loss: 2.7002, Perplexity: 14.8829
Epoch [1/3], Step [30600/41412], Loss: 2.1607, Perplexity: 8.67726
Epoch [1/3], Step [30700/41412], Loss: 2.2901, Perplexity: 9.87623
Epoch [1/3], Step [30800/41412], Loss: 1.8398, Perplexity: 6.29538
Epoch [1/3], Step [30900/41412], Loss: 2.3389, Perplexity: 10.3694
Epoch [1/3], Step [31000/41412], Loss: 1.9031, Perplexity: 6.70647
```

```
Epoch [1/3], Step [31100/41412], Loss: 2.1472, Perplexity: 8.56073
Epoch [1/3], Step [31200/41412], Loss: 2.7841, Perplexity: 16.1860
Epoch [1/3], Step [31300/41412], Loss: 1.5725, Perplexity: 4.81869
Epoch [1/3], Step [31400/41412], Loss: 2.3340, Perplexity: 10.3196
Epoch [1/3], Step [31500/41412], Loss: 2.3056, Perplexity: 10.0306
Epoch [1/3], Step [31600/41412], Loss: 2.4402, Perplexity: 11.4749
Epoch [1/3], Step [31700/41412], Loss: 2.7586, Perplexity: 15.7777
Epoch [1/3], Step [31800/41412], Loss: 3.4545, Perplexity: 31.6423
Epoch [1/3], Step [31900/41412], Loss: 2.7317, Perplexity: 15.3587
Epoch [1/3], Step [32000/41412], Loss: 2.2268, Perplexity: 9.26997
Epoch [1/3], Step [32100/41412], Loss: 2.3380, Perplexity: 10.3607
Epoch [1/3], Step [32200/41412], Loss: 2.0020, Perplexity: 7.40383
Epoch [1/3], Step [32300/41412], Loss: 2.4248, Perplexity: 11.3002
Epoch [1/3], Step [32400/41412], Loss: 2.0831, Perplexity: 8.02890
Epoch [1/3], Step [32500/41412], Loss: 2.3598, Perplexity: 10.5892
Epoch [1/3], Step [32600/41412], Loss: 2.2812, Perplexity: 9.78855
Epoch [1/3], Step [32700/41412], Loss: 1.9253, Perplexity: 6.85699
Epoch [1/3], Step [32800/41412], Loss: 2.7673, Perplexity: 15.9163
Epoch [1/3], Step [32900/41412], Loss: 2.0843, Perplexity: 8.03868
Epoch [1/3], Step [33000/41412], Loss: 2.2619, Perplexity: 9.60092
Epoch [1/3], Step [33100/41412], Loss: 2.7478, Perplexity: 15.6084
Epoch [1/3], Step [33200/41412], Loss: 2.0962, Perplexity: 8.13518
Epoch [1/3], Step [33300/41412], Loss: 2.2658, Perplexity: 9.638988
Epoch [1/3], Step [33400/41412], Loss: 1.9218, Perplexity: 6.83300
Epoch [1/3], Step [33500/41412], Loss: 2.4410, Perplexity: 11.4845
Epoch [1/3], Step [33600/41412], Loss: 2.4696, Perplexity: 11.8177
Epoch [1/3], Step [33700/41412], Loss: 2.2938, Perplexity: 9.91228
Epoch [1/3], Step [33800/41412], Loss: 2.8926, Perplexity: 18.04029
Epoch [1/3], Step [33900/41412], Loss: 3.1006, Perplexity: 22.2102
Epoch [1/3], Step [34000/41412], Loss: 2.2074, Perplexity: 9.09173
Epoch [1/3], Step [34100/41412], Loss: 2.2267, Perplexity: 9.26919
Epoch [1/3], Step [34200/41412], Loss: 2.4333, Perplexity: 11.3966
Epoch [1/3], Step [34300/41412], Loss: 2.7511, Perplexity: 15.6591
Epoch [1/3], Step [34400/41412], Loss: 1.9257, Perplexity: 6.85978
Epoch [1/3], Step [34500/41412], Loss: 1.8699, Perplexity: 6.48760
Epoch [1/3], Step [34600/41412], Loss: 2.8147, Perplexity: 16.6879
Epoch [1/3], Step [34700/41412], Loss: 2.4607, Perplexity: 11.7131
Epoch [1/3], Step [34800/41412], Loss: 1.9676, Perplexity: 7.15354
Epoch [1/3], Step [34900/41412], Loss: 2.0197, Perplexity: 7.53607
Epoch [1/3], Step [35000/41412], Loss: 2.0957, Perplexity: 8.13111
Epoch [1/3], Step [35100/41412], Loss: 2.3668, Perplexity: 10.6628
Epoch [1/3], Step [35200/41412], Loss: 2.7598, Perplexity: 15.7966
Epoch [1/3], Step [35300/41412], Loss: 2.2256, Perplexity: 9.25861
Epoch [1/3], Step [35400/41412], Loss: 2.4822, Perplexity: 11.9674
Epoch [1/3], Step [35500/41412], Loss: 2.0863, Perplexity: 8.05532
Epoch [1/3], Step [35600/41412], Loss: 2.3400, Perplexity: 10.3815
Epoch [1/3], Step [35700/41412], Loss: 2.0768, Perplexity: 7.97899
Epoch [1/3], Step [35800/41412], Loss: 2.0219, Perplexity: 7.55263
Epoch [1/3], Step [35900/41412], Loss: 2.7866, Perplexity: 16.2256
Epoch [1/3], Step [36000/41412], Loss: 2.0977, Perplexity: 8.14771
Epoch [1/3], Step [36100/41412], Loss: 2.3258, Perplexity: 10.2352
Epoch [1/3], Step [36200/41412], Loss: 2.7056, Perplexity: 14.9630
Epoch [1/3], Step [36300/41412], Loss: 4.6847, Perplexity: 108.2782
Epoch [1/3], Step [36400/41412], Loss: 2.3602, Perplexity: 10.5926
Epoch [1/3], Step [36500/41412], Loss: 2.3812, Perplexity: 10.8178
Epoch [1/3], Step [36600/41412], Loss: 2.8510, Perplexity: 17.3056
Epoch [1/3], Step [36700/41412], Loss: 2.6009, Perplexity: 13.4755
Epoch [1/3], Step [36800/41412], Loss: 2.5772, Perplexity: 13.1600
Epoch [1/3], Step [36900/41412], Loss: 2.0397, Perplexity: 7.68857
Epoch [1/3], Step [37000/41412], Loss: 2.5119, Perplexity: 12.3280
Epoch [1/3], Step [37100/41412], Loss: 2.4432, Perplexity: 11.5102
Epoch [1/3], Step [37200/41412], Loss: 2.0229, Perplexity: 7.56051
```

```
Epoch [1/3], Step [37300/41412], Loss: 2.4432, Perplexity: 11.5093
Epoch [1/3], Step [37400/41412], Loss: 2.0944, Perplexity: 8.12036
Epoch [1/3], Step [37500/41412], Loss: 2.2680, Perplexity: 9.66050
Epoch [1/3], Step [37600/41412], Loss: 2.2866, Perplexity: 9.84102
Epoch [1/3], Step [37700/41412], Loss: 2.4851, Perplexity: 12.0028
Epoch [1/3], Step [37800/41412], Loss: 2.7517, Perplexity: 15.6687
Epoch [1/3], Step [37900/41412], Loss: 2.3677, Perplexity: 10.6731
Epoch [1/3], Step [38000/41412], Loss: 2.4239, Perplexity: 11.2898
Epoch [1/3], Step [38100/41412], Loss: 3.5674, Perplexity: 35.4238
Epoch [1/3], Step [38200/41412], Loss: 2.5422, Perplexity: 12.7077
Epoch [1/3], Step [38300/41412], Loss: 2.2702, Perplexity: 9.68155
Epoch [1/3], Step [38400/41412], Loss: 2.6516, Perplexity: 14.1770
Epoch [1/3], Step [38500/41412], Loss: 2.7572, Perplexity: 15.7563
Epoch [1/3], Step [38600/41412], Loss: 2.5608, Perplexity: 12.9468
Epoch [1/3], Step [38700/41412], Loss: 2.0806, Perplexity: 8.00957
Epoch [1/3], Step [38800/41412], Loss: 2.3822, Perplexity: 10.8289
Epoch [1/3], Step [38900/41412], Loss: 3.1590, Perplexity: 23.5467
Epoch [1/3], Step [39000/41412], Loss: 2.1302, Perplexity: 8.41692
Epoch [1/3], Step [39100/41412], Loss: 1.9811, Perplexity: 7.25074
Epoch [1/3], Step [39200/41412], Loss: 2.2623, Perplexity: 9.60503
Epoch [1/3], Step [39300/41412], Loss: 2.3159, Perplexity: 10.1343
Epoch [1/3], Step [39400/41412], Loss: 2.3775, Perplexity: 10.7775
Epoch [1/3], Step [39500/41412], Loss: 1.9889, Perplexity: 7.30744
Epoch [1/3], Step [39600/41412], Loss: 2.3488, Perplexity: 10.4727
Epoch [1/3], Step [39700/41412], Loss: 2.0155, Perplexity: 7.50428
Epoch [1/3], Step [39800/41412], Loss: 2.6624, Perplexity: 14.3309
Epoch [1/3], Step [39900/41412], Loss: 2.7942, Perplexity: 16.3491
Epoch [1/3], Step [40000/41412], Loss: 2.2035, Perplexity: 9.05653
Epoch [1/3], Step [40100/41412], Loss: 2.7478, Perplexity: 15.6089
Epoch [1/3], Step [40200/41412], Loss: 2.3907, Perplexity: 10.9214
Epoch [1/3], Step [40300/41412], Loss: 2.3125, Perplexity: 10.0996
Epoch [1/3], Step [40400/41412], Loss: 2.4693, Perplexity: 11.8145
Epoch [1/3], Step [40500/41412], Loss: 2.3816, Perplexity: 10.8223
Epoch [1/3], Step [40600/41412], Loss: 2.4823, Perplexity: 11.96853
Epoch [1/3], Step [40700/41412], Loss: 2.2340, Perplexity: 9.33674
Epoch [1/3], Step [40800/41412], Loss: 2.5696, Perplexity: 13.0611
Epoch [1/3], Step [40900/41412], Loss: 2.6778, Perplexity: 14.55322
Epoch [1/3], Step [41000/41412], Loss: 3.0849, Perplexity: 21.8662
Epoch [1/3], Step [41100/41412], Loss: 3.0789, Perplexity: 21.7351
Epoch [1/3], Step [41200/41412], Loss: 2.3448, Perplexity: 10.4314
Epoch [1/3], Step [41300/41412], Loss: 2.9593, Perplexity: 19.2850
Epoch [1/3], Step [41400/41412], Loss: 2.3488, Perplexity: 10.4733
Epoch [1/3], Step [41412/41412], Loss: 2.2761, Perplexity: 9.73871
Epoch [1/3] completed in 4893.24 seconds.
Epoch [2/3], Step [100/41412], Loss: 2.7837, Perplexity: 16.1790
Epoch [2/3], Step [200/41412], Loss: 2.4922, Perplexity: 12.0884
Epoch [2/3], Step [300/41412], Loss: 2.4560, Perplexity: 11.6583
Epoch [2/3], Step [400/41412], Loss: 2.5202, Perplexity: 12.4316
Epoch [2/3], Step [500/41412], Loss: 2.0618, Perplexity: 7.86019
Epoch [2/3], Step [600/41412], Loss: 2.2768, Perplexity: 9.74540
Epoch [2/3], Step [700/41412], Loss: 2.1554, Perplexity: 8.63148
Epoch [2/3], Step [800/41412], Loss: 1.9976, Perplexity: 7.37142
Epoch [2/3], Step [900/41412], Loss: 2.6949, Perplexity: 14.8047
Epoch [2/3], Step [1000/41412], Loss: 2.5219, Perplexity: 12.4520
Epoch [2/3], Step [1100/41412], Loss: 2.1562, Perplexity: 8.63808
Epoch [2/3], Step [1200/41412], Loss: 2.2705, Perplexity: 9.68458
Epoch [2/3], Step [1300/41412], Loss: 2.2442, Perplexity: 9.43312
Epoch [2/3], Step [1400/41412], Loss: 2.1172, Perplexity: 8.30751
Epoch [2/3], Step [1500/41412], Loss: 2.7417, Perplexity: 15.5129
Epoch [2/3], Step [1600/41412], Loss: 2.6178, Perplexity: 13.7062
Epoch [2/3], Step [1700/41412], Loss: 2.7244, Perplexity: 15.2473
Epoch [2/3], Step [1800/41412], Loss: 2.4261, Perplexity: 11.3144
```

```
Epoch [2/3], Step [1900/41412], Loss: 2.0120, Perplexity: 7.47800
Epoch [2/3], Step [2000/41412], Loss: 2.0717, Perplexity: 7.93853
Epoch [2/3], Step [2100/41412], Loss: 3.3334, Perplexity: 28.0330
Epoch [2/3], Step [2200/41412], Loss: 2.1509, Perplexity: 8.59260
Epoch [2/3], Step [2300/41412], Loss: 2.4753, Perplexity: 11.8850
Epoch [2/3], Step [2400/41412], Loss: 2.4432, Perplexity: 11.5092
Epoch [2/3], Step [2500/41412], Loss: 2.7405, Perplexity: 15.4944
Epoch [2/3], Step [2600/41412], Loss: 2.2198, Perplexity: 9.20522
Epoch [2/3], Step [2700/41412], Loss: 2.4523, Perplexity: 11.6151
Epoch [2/3], Step [2800/41412], Loss: 2.2334, Perplexity: 9.33124
Epoch [2/3], Step [2900/41412], Loss: 2.4470, Perplexity: 11.5539
Epoch [2/3], Step [3000/41412], Loss: 2.4965, Perplexity: 12.1396
Epoch [2/3], Step [3100/41412], Loss: 1.9642, Perplexity: 7.12938
Epoch [2/3], Step [3200/41412], Loss: 1.9753, Perplexity: 7.20888
Epoch [2/3], Step [3300/41412], Loss: 2.2858, Perplexity: 9.83383
Epoch [2/3], Step [3400/41412], Loss: 2.0502, Perplexity: 7.76933
Epoch [2/3], Step [3500/41412], Loss: 2.3599, Perplexity: 10.5897
Epoch [2/3], Step [3600/41412], Loss: 2.1327, Perplexity: 8.43746
Epoch [2/3], Step [3700/41412], Loss: 2.3664, Perplexity: 10.6588
Epoch [2/3], Step [3800/41412], Loss: 1.8512, Perplexity: 6.36770
Epoch [2/3], Step [3900/41412], Loss: 2.4698, Perplexity: 11.8205
Epoch [2/3], Step [4000/41412], Loss: 2.4210, Perplexity: 11.2574
Epoch [2/3], Step [4100/41412], Loss: 2.7489, Perplexity: 15.6248
Epoch [2/3], Step [4200/41412], Loss: 2.3007, Perplexity: 9.98155
Epoch [2/3], Step [4300/41412], Loss: 2.0725, Perplexity: 7.94491
Epoch [2/3], Step [4400/41412], Loss: 2.0262, Perplexity: 7.58492
Epoch [2/3], Step [4500/41412], Loss: 2.7013, Perplexity: 14.8992
Epoch [2/3], Step [4600/41412], Loss: 2.6940, Perplexity: 14.7901
Epoch [2/3], Step [4700/41412], Loss: 2.6932, Perplexity: 14.7791
Epoch [2/3], Step [4800/41412], Loss: 2.2851, Perplexity: 9.827117
Epoch [2/3], Step [4900/41412], Loss: 2.1780, Perplexity: 8.82870
Epoch [2/3], Step [5000/41412], Loss: 2.0576, Perplexity: 7.82690
Epoch [2/3], Step [5100/41412], Loss: 2.6083, Perplexity: 13.5765
Epoch [2/3], Step [5200/41412], Loss: 2.2362, Perplexity: 9.35743
Epoch [2/3], Step [5300/41412], Loss: 2.1976, Perplexity: 9.00306
Epoch [2/3], Step [5400/41412], Loss: 2.4035, Perplexity: 11.0621
Epoch [2/3], Step [5500/41412], Loss: 1.9676, Perplexity: 7.15383
Epoch [2/3], Step [5600/41412], Loss: 2.5478, Perplexity: 12.7786
Epoch [2/3], Step [5700/41412], Loss: 2.2227, Perplexity: 9.23262
Epoch [2/3], Step [5800/41412], Loss: 2.4477, Perplexity: 11.5622
Epoch [2/3], Step [5900/41412], Loss: 2.1058, Perplexity: 8.21337
Epoch [2/3], Step [6000/41412], Loss: 2.4667, Perplexity: 11.7840
Epoch [2/3], Step [6100/41412], Loss: 3.2296, Perplexity: 25.2698
Epoch [2/3], Step [6200/41412], Loss: 1.6521, Perplexity: 5.21776
Epoch [2/3], Step [6300/41412], Loss: 2.2997, Perplexity: 9.971634
Epoch [2/3], Step [6400/41412], Loss: 2.2767, Perplexity: 9.74406
Epoch [2/3], Step [6500/41412], Loss: 1.7420, Perplexity: 5.70890
Epoch [2/3], Step [6600/41412], Loss: 2.3178, Perplexity: 10.1529
Epoch [2/3], Step [6700/41412], Loss: 2.0265, Perplexity: 7.58755
Epoch [2/3], Step [6800/41412], Loss: 1.8161, Perplexity: 6.14778
Epoch [2/3], Step [6900/41412], Loss: 1.9005, Perplexity: 6.68925
Epoch [2/3], Step [7000/41412], Loss: 2.5245, Perplexity: 12.4846
Epoch [2/3], Step [7100/41412], Loss: 1.7462, Perplexity: 5.73297
Epoch [2/3], Step [7200/41412], Loss: 2.4586, Perplexity: 11.6883
Epoch [2/3], Step [7300/41412], Loss: 2.4648, Perplexity: 11.7612
Epoch [2/3], Step [7400/41412], Loss: 1.5986, Perplexity: 4.94608
Epoch [2/3], Step [7500/41412], Loss: 2.0956, Perplexity: 8.13073
Epoch [2/3], Step [7600/41412], Loss: 2.0092, Perplexity: 7.45706
Epoch [2/3], Step [7700/41412], Loss: 1.9799, Perplexity: 7.24222
Epoch [2/3], Step [7800/41412], Loss: 1.9527, Perplexity: 7.04787
Epoch [2/3], Step [7900/41412], Loss: 2.0521, Perplexity: 7.78431
Epoch [2/3], Step [8000/41412], Loss: 1.9321, Perplexity: 6.90399
```

```
Epoch [2/3], Step [8100/41412], Loss: 2.1556, Perplexity: 8.63309
Epoch [2/3], Step [8200/41412], Loss: 1.9259, Perplexity: 6.86115
Epoch [2/3], Step [8300/41412], Loss: 2.3008, Perplexity: 9.98244
Epoch [2/3], Step [8400/41412], Loss: 2.3549, Perplexity: 10.5366
Epoch [2/3], Step [8500/41412], Loss: 2.5683, Perplexity: 13.0432
Epoch [2/3], Step [8600/41412], Loss: 2.1000, Perplexity: 8.16658
Epoch [2/3], Step [8700/41412], Loss: 2.3082, Perplexity: 10.0568
Epoch [2/3], Step [8800/41412], Loss: 2.1157, Perplexity: 8.29574
Epoch [2/3], Step [8900/41412], Loss: 1.9595, Perplexity: 7.09591
Epoch [2/3], Step [9000/41412], Loss: 2.7329, Perplexity: 15.37690
Epoch [2/3], Step [9100/41412], Loss: 2.0347, Perplexity: 7.64993
Epoch [2/3], Step [9200/41412], Loss: 1.9244, Perplexity: 6.85094
Epoch [2/3], Step [9300/41412], Loss: 2.7101, Perplexity: 15.0314
Epoch [2/3], Step [9400/41412], Loss: 2.4301, Perplexity: 11.35997
Epoch [2/3], Step [9500/41412], Loss: 2.3110, Perplexity: 10.0846
Epoch [2/3], Step [9600/41412], Loss: 3.2616, Perplexity: 26.0924
Epoch [2/3], Step [9700/41412], Loss: 2.2155, Perplexity: 9.16565
Epoch [2/3], Step [9800/41412], Loss: 2.6279, Perplexity: 13.8446
Epoch [2/3], Step [9900/41412], Loss: 2.7321, Perplexity: 15.3651
Epoch [2/3], Step [10000/41412], Loss: 1.7812, Perplexity: 5.9368
Epoch [2/3], Step [10100/41412], Loss: 2.0948, Perplexity: 8.12357
Epoch [2/3], Step [10200/41412], Loss: 2.9043, Perplexity: 18.2533
Epoch [2/3], Step [10300/41412], Loss: 2.2852, Perplexity: 9.82754
Epoch [2/3], Step [10400/41412], Loss: 2.4196, Perplexity: 11.2412
Epoch [2/3], Step [10500/41412], Loss: 1.5588, Perplexity: 4.75332
Epoch [2/3], Step [10600/41412], Loss: 2.3935, Perplexity: 10.9520
Epoch [2/3], Step [10700/41412], Loss: 2.2533, Perplexity: 9.51875
Epoch [2/3], Step [10800/41412], Loss: 2.1953, Perplexity: 8.98297
Epoch [2/3], Step [10900/41412], Loss: 2.2963, Perplexity: 9.93730
Epoch [2/3], Step [11000/41412], Loss: 2.3460, Perplexity: 10.4442
Epoch [2/3], Step [11100/41412], Loss: 2.3651, Perplexity: 10.6453
Epoch [2/3], Step [11200/41412], Loss: 1.8080, Perplexity: 6.09831
Epoch [2/3], Step [11300/41412], Loss: 2.0340, Perplexity: 7.64439
Epoch [2/3], Step [11400/41412], Loss: 2.4136, Perplexity: 11.1736
Epoch [2/3], Step [11500/41412], Loss: 2.3590, Perplexity: 10.5799
Epoch [2/3], Step [11600/41412], Loss: 2.2480, Perplexity: 9.46844
Epoch [2/3], Step [11700/41412], Loss: 2.6730, Perplexity: 14.4830
Epoch [2/3], Step [11800/41412], Loss: 2.7932, Perplexity: 16.3338
Epoch [2/3], Step [11900/41412], Loss: 1.9492, Perplexity: 7.02302
Epoch [2/3], Step [12000/41412], Loss: 2.7054, Perplexity: 14.9604
Epoch [2/3], Step [12100/41412], Loss: 1.8751, Perplexity: 6.52183
Epoch [2/3], Step [12200/41412], Loss: 3.3012, Perplexity: 27.1463
Epoch [2/3], Step [12300/41412], Loss: 2.5593, Perplexity: 12.9265
Epoch [2/3], Step [12400/41412], Loss: 2.8477, Perplexity: 17.2484
Epoch [2/3], Step [12500/41412], Loss: 2.0784, Perplexity: 7.99175
Epoch [2/3], Step [12600/41412], Loss: 2.1483, Perplexity: 8.57067
Epoch [2/3], Step [12700/41412], Loss: 2.2658, Perplexity: 9.63868
Epoch [2/3], Step [12800/41412], Loss: 2.4076, Perplexity: 11.1078
Epoch [2/3], Step [12900/41412], Loss: 1.7858, Perplexity: 5.96451
Epoch [2/3], Step [13000/41412], Loss: 2.5889, Perplexity: 13.3149
Epoch [2/3], Step [13100/41412], Loss: 2.1397, Perplexity: 8.49725
Epoch [2/3], Step [13200/41412], Loss: 2.2412, Perplexity: 9.40485
Epoch [2/3], Step [13300/41412], Loss: 1.9278, Perplexity: 6.87455
Epoch [2/3], Step [13400/41412], Loss: 2.1775, Perplexity: 8.82449
Epoch [2/3], Step [13500/41412], Loss: 2.3394, Perplexity: 10.3745
Epoch [2/3], Step [13600/41412], Loss: 1.8190, Perplexity: 6.16549
Epoch [2/3], Step [13700/41412], Loss: 2.2092, Perplexity: 9.10837
Epoch [2/3], Step [13800/41412], Loss: 2.2307, Perplexity: 9.30616
Epoch [2/3], Step [13900/41412], Loss: 2.7366, Perplexity: 15.4339
Epoch [2/3], Step [14000/41412], Loss: 2.2096, Perplexity: 9.11172
Epoch [2/3], Step [14100/41412], Loss: 2.2098, Perplexity: 9.11395
Epoch [2/3], Step [14200/41412], Loss: 2.3147, Perplexity: 10.1223
```

```
Epoch [2/3], Step [14300/41412], Loss: 2.5832, Perplexity: 13.2401
Epoch [2/3], Step [14400/41412], Loss: 1.8522, Perplexity: 6.37403
Epoch [2/3], Step [14500/41412], Loss: 2.0700, Perplexity: 7.92521
Epoch [2/3], Step [14600/41412], Loss: 2.2500, Perplexity: 9.48802
Epoch [2/3], Step [14700/41412], Loss: 2.7177, Perplexity: 15.1457
Epoch [2/3], Step [14800/41412], Loss: 2.2483, Perplexity: 9.47124
Epoch [2/3], Step [14900/41412], Loss: 2.1282, Perplexity: 8.39958
Epoch [2/3], Step [15000/41412], Loss: 3.0147, Perplexity: 20.3827
Epoch [2/3], Step [15100/41412], Loss: 2.4800, Perplexity: 11.9419
Epoch [2/3], Step [15200/41412], Loss: 2.1443, Perplexity: 8.53602
Epoch [2/3], Step [15300/41412], Loss: 1.9597, Perplexity: 7.09727
Epoch [2/3], Step [15400/41412], Loss: 1.6108, Perplexity: 5.00682
Epoch [2/3], Step [15500/41412], Loss: 2.3296, Perplexity: 10.2739
Epoch [2/3], Step [15600/41412], Loss: 2.1733, Perplexity: 8.78757
Epoch [2/3], Step [15700/41412], Loss: 2.5480, Perplexity: 12.7815
Epoch [2/3], Step [15800/41412], Loss: 2.1735, Perplexity: 8.78914
Epoch [2/3], Step [15900/41412], Loss: 2.1593, Perplexity: 8.66499
Epoch [2/3], Step [16000/41412], Loss: 2.7205, Perplexity: 15.1877
Epoch [2/3], Step [16100/41412], Loss: 2.7326, Perplexity: 15.3722
Epoch [2/3], Step [16200/41412], Loss: 2.3186, Perplexity: 10.1615
Epoch [2/3], Step [16300/41412], Loss: 2.0577, Perplexity: 7.82810
Epoch [2/3], Step [16400/41412], Loss: 2.8590, Perplexity: 17.4438
Epoch [2/3], Step [16500/41412], Loss: 2.6871, Perplexity: 14.6886
Epoch [2/3], Step [16600/41412], Loss: 2.0159, Perplexity: 7.50724
Epoch [2/3], Step [16700/41412], Loss: 2.0522, Perplexity: 7.78486
Epoch [2/3], Step [16800/41412], Loss: 2.1172, Perplexity: 8.30754
Epoch [2/3], Step [16900/41412], Loss: 5.1325, Perplexity: 169.4347
Epoch [2/3], Step [17000/41412], Loss: 2.3616, Perplexity: 10.6081
Epoch [2/3], Step [17100/41412], Loss: 2.0472, Perplexity: 7.74629
Epoch [2/3], Step [17200/41412], Loss: 2.0853, Perplexity: 8.04734
Epoch [2/3], Step [17300/41412], Loss: 2.3068, Perplexity: 10.0426
Epoch [2/3], Step [17400/41412], Loss: 1.9374, Perplexity: 6.94049
Epoch [2/3], Step [17500/41412], Loss: 2.2478, Perplexity: 9.46737
Epoch [2/3], Step [17600/41412], Loss: 1.7436, Perplexity: 5.71804
Epoch [2/3], Step [17700/41412], Loss: 2.3780, Perplexity: 10.7836
Epoch [2/3], Step [17800/41412], Loss: 2.2199, Perplexity: 9.20631
Epoch [2/3], Step [17900/41412], Loss: 3.0491, Perplexity: 21.0954
Epoch [2/3], Step [18000/41412], Loss: 1.9205, Perplexity: 6.82421
Epoch [2/3], Step [18100/41412], Loss: 2.2311, Perplexity: 9.31038
Epoch [2/3], Step [18200/41412], Loss: 2.2161, Perplexity: 9.17194
Epoch [2/3], Step [18300/41412], Loss: 2.8651, Perplexity: 17.5506
Epoch [2/3], Step [18400/41412], Loss: 2.3160, Perplexity: 10.1352
Epoch [2/3], Step [18500/41412], Loss: 2.2194, Perplexity: 9.20195
Epoch [2/3], Step [18600/41412], Loss: 1.9955, Perplexity: 7.35564
Epoch [2/3], Step [18700/41412], Loss: 2.1442, Perplexity: 8.53506
Epoch [2/3], Step [18800/41412], Loss: 2.1023, Perplexity: 8.18533
Epoch [2/3], Step [18900/41412], Loss: 2.5907, Perplexity: 13.3388
Epoch [2/3], Step [19000/41412], Loss: 2.4958, Perplexity: 12.1309
Epoch [2/3], Step [19100/41412], Loss: 2.4492, Perplexity: 11.5795
Epoch [2/3], Step [19200/41412], Loss: 2.4881, Perplexity: 12.0380
Epoch [2/3], Step [19300/41412], Loss: 2.4676, Perplexity: 11.7940
Epoch [2/3], Step [19400/41412], Loss: 2.3883, Perplexity: 10.8949
Epoch [2/3], Step [19500/41412], Loss: 2.5161, Perplexity: 12.3797
Epoch [2/3], Step [19600/41412], Loss: 2.2133, Perplexity: 9.14635
Epoch [2/3], Step [19700/41412], Loss: 2.6999, Perplexity: 14.8789
Epoch [2/3], Step [19800/41412], Loss: 2.2701, Perplexity: 9.68086
Epoch [2/3], Step [19900/41412], Loss: 1.7428, Perplexity: 5.71343
Epoch [2/3], Step [20000/41412], Loss: 1.6725, Perplexity: 5.32567
Epoch [2/3], Step [20100/41412], Loss: 2.2151, Perplexity: 9.16242
Epoch [2/3], Step [20200/41412], Loss: 2.1116, Perplexity: 8.26171
Epoch [2/3], Step [20300/41412], Loss: 2.3019, Perplexity: 9.99272
Epoch [2/3], Step [20400/41412], Loss: 2.5135, Perplexity: 12.3477
```

```
Epoch [2/3], Step [20500/41412], Loss: 2.2553, Perplexity: 9.53811
Epoch [2/3], Step [20600/41412], Loss: 1.7403, Perplexity: 5.69937
Epoch [2/3], Step [20700/41412], Loss: 3.1143, Perplexity: 22.5175
Epoch [2/3], Step [20800/41412], Loss: 2.8323, Perplexity: 16.9853
Epoch [2/3], Step [20900/41412], Loss: 2.4846, Perplexity: 11.9968
Epoch [2/3], Step [21000/41412], Loss: 2.5464, Perplexity: 12.7616
Epoch [2/3], Step [21100/41412], Loss: 2.0681, Perplexity: 7.909842
Epoch [2/3], Step [21200/41412], Loss: 2.6741, Perplexity: 14.4989
Epoch [2/3], Step [21300/41412], Loss: 2.3382, Perplexity: 10.3625
Epoch [2/3], Step [21400/41412], Loss: 1.6384, Perplexity: 5.14701
Epoch [2/3], Step [21500/41412], Loss: 1.9949, Perplexity: 7.35161
Epoch [2/3], Step [21600/41412], Loss: 1.9473, Perplexity: 7.00988
Epoch [2/3], Step [21700/41412], Loss: 2.2396, Perplexity: 9.38987
Epoch [2/3], Step [21800/41412], Loss: 2.5869, Perplexity: 13.2889
Epoch [2/3], Step [21900/41412], Loss: 2.2221, Perplexity: 9.22647
Epoch [2/3], Step [22000/41412], Loss: 2.4046, Perplexity: 11.0736
Epoch [2/3], Step [22100/41412], Loss: 1.9005, Perplexity: 6.68950
Epoch [2/3], Step [22200/41412], Loss: 2.4472, Perplexity: 11.5563
Epoch [2/3], Step [22300/41412], Loss: 2.8202, Perplexity: 16.7809
Epoch [2/3], Step [22400/41412], Loss: 2.1829, Perplexity: 8.87204
Epoch [2/3], Step [22500/41412], Loss: 2.0299, Perplexity: 7.61313
Epoch [2/3], Step [22600/41412], Loss: 2.4589, Perplexity: 11.6921
Epoch [2/3], Step [22700/41412], Loss: 2.5549, Perplexity: 12.8700
Epoch [2/3], Step [22800/41412], Loss: 2.5883, Perplexity: 13.3070
Epoch [2/3], Step [22900/41412], Loss: 2.2206, Perplexity: 9.21327
Epoch [2/3], Step [23000/41412], Loss: 2.0146, Perplexity: 7.49797
Epoch [2/3], Step [23100/41412], Loss: 2.7988, Perplexity: 16.4255
Epoch [2/3], Step [23200/41412], Loss: 2.1653, Perplexity: 8.71700
Epoch [2/3], Step [23300/41412], Loss: 2.0384, Perplexity: 7.67822
Epoch [2/3], Step [23400/41412], Loss: 2.4317, Perplexity: 11.3778
Epoch [2/3], Step [23500/41412], Loss: 2.1604, Perplexity: 8.67446
Epoch [2/3], Step [23600/41412], Loss: 2.0984, Perplexity: 8.15358
Epoch [2/3], Step [23700/41412], Loss: 2.1355, Perplexity: 8.46098
Epoch [2/3], Step [23800/41412], Loss: 2.3802, Perplexity: 10.8065
Epoch [2/3], Step [23900/41412], Loss: 2.0843, Perplexity: 8.03915
Epoch [2/3], Step [24000/41412], Loss: 2.5350, Perplexity: 12.6158
Epoch [2/3], Step [24100/41412], Loss: 1.9440, Perplexity: 6.98653
Epoch [2/3], Step [24200/41412], Loss: 2.2344, Perplexity: 9.34080
Epoch [2/3], Step [24300/41412], Loss: 2.5785, Perplexity: 13.1780
Epoch [2/3], Step [24400/41412], Loss: 2.2499, Perplexity: 9.48653
Epoch [2/3], Step [24500/41412], Loss: 2.2533, Perplexity: 9.51912
Epoch [2/3], Step [24600/41412], Loss: 2.9765, Perplexity: 19.6191
Epoch [2/3], Step [24700/41412], Loss: 2.5173, Perplexity: 12.3949
Epoch [2/3], Step [24800/41412], Loss: 1.9410, Perplexity: 6.96548
Epoch [2/3], Step [24900/41412], Loss: 3.7483, Perplexity: 42.4478
Epoch [2/3], Step [25000/41412], Loss: 2.6155, Perplexity: 13.6739
Epoch [2/3], Step [25100/41412], Loss: 2.5174, Perplexity: 12.3959
Epoch [2/3], Step [25200/41412], Loss: 1.9212, Perplexity: 6.82949
Epoch [2/3], Step [25300/41412], Loss: 2.2150, Perplexity: 9.16159
Epoch [2/3], Step [25400/41412], Loss: 1.9069, Perplexity: 6.73202
Epoch [2/3], Step [25500/41412], Loss: 1.8986, Perplexity: 6.67653
Epoch [2/3], Step [25600/41412], Loss: 1.8026, Perplexity: 6.06531
Epoch [2/3], Step [25700/41412], Loss: 2.5875, Perplexity: 13.2970
Epoch [2/3], Step [25800/41412], Loss: 2.4554, Perplexity: 11.6514
Epoch [2/3], Step [25900/41412], Loss: 2.0573, Perplexity: 7.82504
Epoch [2/3], Step [26000/41412], Loss: 1.9747, Perplexity: 7.20480
Epoch [2/3], Step [26100/41412], Loss: 2.4228, Perplexity: 11.2773
Epoch [2/3], Step [26200/41412], Loss: 2.0938, Perplexity: 8.11605
Epoch [2/3], Step [26300/41412], Loss: 2.1706, Perplexity: 8.76374
Epoch [2/3], Step [26400/41412], Loss: 2.0555, Perplexity: 7.81045
Epoch [2/3], Step [26500/41412], Loss: 2.2759, Perplexity: 9.73672
Epoch [2/3], Step [26600/41412], Loss: 2.2783, Perplexity: 9.75963
```

```
Epoch [2/3], Step [26700/41412], Loss: 2.6555, Perplexity: 14.2315
Epoch [2/3], Step [26800/41412], Loss: 1.6711, Perplexity: 5.31838
Epoch [2/3], Step [26900/41412], Loss: 2.8084, Perplexity: 16.5829
Epoch [2/3], Step [27000/41412], Loss: 2.2742, Perplexity: 9.72004
Epoch [2/3], Step [27100/41412], Loss: 1.9718, Perplexity: 7.18373
Epoch [2/3], Step [27200/41412], Loss: 2.4797, Perplexity: 11.9373
Epoch [2/3], Step [27300/41412], Loss: 2.3254, Perplexity: 10.23084
Epoch [2/3], Step [27400/41412], Loss: 2.3200, Perplexity: 10.1752
Epoch [2/3], Step [27500/41412], Loss: 2.0391, Perplexity: 7.68403
Epoch [2/3], Step [27600/41412], Loss: 2.1554, Perplexity: 8.63110
Epoch [2/3], Step [27700/41412], Loss: 2.3766, Perplexity: 10.7679
Epoch [2/3], Step [27800/41412], Loss: 1.8766, Perplexity: 6.53149
Epoch [2/3], Step [27900/41412], Loss: 2.6859, Perplexity: 14.6715
Epoch [2/3], Step [28000/41412], Loss: 2.1652, Perplexity: 8.71601
Epoch [2/3], Step [28100/41412], Loss: 2.5309, Perplexity: 12.5652
Epoch [2/3], Step [28200/41412], Loss: 2.6861, Perplexity: 14.6750
Epoch [2/3], Step [28300/41412], Loss: 2.1109, Perplexity: 8.25557
Epoch [2/3], Step [28400/41412], Loss: 2.9620, Perplexity: 19.3364
Epoch [2/3], Step [28500/41412], Loss: 2.2281, Perplexity: 9.28229
Epoch [2/3], Step [28600/41412], Loss: 2.0064, Perplexity: 7.43643
Epoch [2/3], Step [28700/41412], Loss: 2.0682, Perplexity: 7.91037
Epoch [2/3], Step [28800/41412], Loss: 1.8869, Perplexity: 6.59862
Epoch [2/3], Step [28900/41412], Loss: 2.4527, Perplexity: 11.6193
Epoch [2/3], Step [29000/41412], Loss: 2.8055, Perplexity: 16.5356
Epoch [2/3], Step [29100/41412], Loss: 2.3780, Perplexity: 10.7831
Epoch [2/3], Step [29200/41412], Loss: 1.8892, Perplexity: 6.61432
Epoch [2/3], Step [29300/41412], Loss: 2.0724, Perplexity: 7.94402
Epoch [2/3], Step [29400/41412], Loss: 2.5861, Perplexity: 13.2772
Epoch [2/3], Step [29500/41412], Loss: 2.2919, Perplexity: 9.89348
Epoch [2/3], Step [29600/41412], Loss: 2.3923, Perplexity: 10.9382
Epoch [2/3], Step [29700/41412], Loss: 2.1972, Perplexity: 8.99965
Epoch [2/3], Step [29800/41412], Loss: 2.3911, Perplexity: 10.9251
Epoch [2/3], Step [29900/41412], Loss: 1.9903, Perplexity: 7.31808
Epoch [2/3], Step [30000/41412], Loss: 2.4665, Perplexity: 11.7817
Epoch [2/3], Step [30100/41412], Loss: 2.6542, Perplexity: 14.2138
Epoch [2/3], Step [30200/41412], Loss: 1.8867, Perplexity: 6.59779
Epoch [2/3], Step [30300/41412], Loss: 2.5817, Perplexity: 13.2190
Epoch [2/3], Step [30400/41412], Loss: 1.5603, Perplexity: 4.76045
Epoch [2/3], Step [30500/41412], Loss: 2.2687, Perplexity: 9.66645
Epoch [2/3], Step [30600/41412], Loss: 2.4114, Perplexity: 11.1496
Epoch [2/3], Step [30700/41412], Loss: 1.9883, Perplexity: 7.30308
Epoch [2/3], Step [30800/41412], Loss: 2.4449, Perplexity: 11.5299
Epoch [2/3], Step [30900/41412], Loss: 2.0871, Perplexity: 8.06178
Epoch [2/3], Step [31000/41412], Loss: 2.5859, Perplexity: 13.2752
Epoch [2/3], Step [31100/41412], Loss: 2.3692, Perplexity: 10.6888
Epoch [2/3], Step [31200/41412], Loss: 2.1021, Perplexity: 8.18370
Epoch [2/3], Step [31300/41412], Loss: 1.8836, Perplexity: 6.57725
Epoch [2/3], Step [31400/41412], Loss: 2.3613, Perplexity: 10.6042
Epoch [2/3], Step [31500/41412], Loss: 2.1991, Perplexity: 9.01664
Epoch [2/3], Step [31600/41412], Loss: 2.7790, Perplexity: 16.1026
Epoch [2/3], Step [31700/41412], Loss: 2.3542, Perplexity: 10.5294
Epoch [2/3], Step [31800/41412], Loss: 1.8242, Perplexity: 6.19785
Epoch [2/3], Step [31900/41412], Loss: 2.2854, Perplexity: 9.82944
Epoch [2/3], Step [32000/41412], Loss: 2.0547, Perplexity: 7.80438
Epoch [2/3], Step [32100/41412], Loss: 1.5912, Perplexity: 4.90983
Epoch [2/3], Step [32200/41412], Loss: 2.1445, Perplexity: 8.538103
Epoch [2/3], Step [32300/41412], Loss: 2.0861, Perplexity: 8.05332
Epoch [2/3], Step [32400/41412], Loss: 2.0587, Perplexity: 7.83591
Epoch [2/3], Step [32500/41412], Loss: 2.7691, Perplexity: 15.9438
Epoch [2/3], Step [32600/41412], Loss: 2.5129, Perplexity: 12.3408
Epoch [2/3], Step [32700/41412], Loss: 2.0685, Perplexity: 7.91269
Epoch [2/3], Step [32800/41412], Loss: 2.2919, Perplexity: 9.89413
```

```
Epoch [2/3], Step [32900/41412], Loss: 2.0831, Perplexity: 8.02953
Epoch [2/3], Step [33000/41412], Loss: 2.1938, Perplexity: 8.96918
Epoch [2/3], Step [33100/41412], Loss: 2.2766, Perplexity: 9.74372
Epoch [2/3], Step [33200/41412], Loss: 2.0944, Perplexity: 8.12078
Epoch [2/3], Step [33300/41412], Loss: 2.4696, Perplexity: 11.8177
Epoch [2/3], Step [33400/41412], Loss: 1.9875, Perplexity: 7.29707
Epoch [2/3], Step [33500/41412], Loss: 2.8119, Perplexity: 16.6409
Epoch [2/3], Step [33600/41412], Loss: 2.1276, Perplexity: 8.39511
Epoch [2/3], Step [33700/41412], Loss: 2.2508, Perplexity: 9.49511
Epoch [2/3], Step [33800/41412], Loss: 1.9205, Perplexity: 6.82432
Epoch [2/3], Step [33900/41412], Loss: 2.0317, Perplexity: 7.62743
Epoch [2/3], Step [34000/41412], Loss: 2.2121, Perplexity: 9.13455
Epoch [2/3], Step [34100/41412], Loss: 2.5168, Perplexity: 12.3892
Epoch [2/3], Step [34200/41412], Loss: 2.0220, Perplexity: 7.55342
Epoch [2/3], Step [34300/41412], Loss: 2.0939, Perplexity: 8.11643
Epoch [2/3], Step [34400/41412], Loss: 3.4237, Perplexity: 30.6828
Epoch [2/3], Step [34500/41412], Loss: 2.2678, Perplexity: 9.65858
Epoch [2/3], Step [34600/41412], Loss: 2.5504, Perplexity: 12.8123
Epoch [2/3], Step [34700/41412], Loss: 2.2342, Perplexity: 9.33932
Epoch [2/3], Step [34800/41412], Loss: 2.2740, Perplexity: 9.71837
Epoch [2/3], Step [34900/41412], Loss: 1.9644, Perplexity: 7.13060
Epoch [2/3], Step [35000/41412], Loss: 1.9197, Perplexity: 6.81892
Epoch [2/3], Step [35100/41412], Loss: 2.5008, Perplexity: 12.1924
Epoch [2/3], Step [35200/41412], Loss: 2.7223, Perplexity: 15.2153
Epoch [2/3], Step [35300/41412], Loss: 2.2333, Perplexity: 9.33044
Epoch [2/3], Step [35400/41412], Loss: 2.4791, Perplexity: 11.9304
Epoch [2/3], Step [35500/41412], Loss: 1.9435, Perplexity: 6.98327
Epoch [2/3], Step [35600/41412], Loss: 1.9832, Perplexity: 7.26628
Epoch [2/3], Step [35700/41412], Loss: 2.8576, Perplexity: 17.4201
Epoch [2/3], Step [35800/41412], Loss: 2.2633, Perplexity: 9.61457
Epoch [2/3], Step [35900/41412], Loss: 1.7623, Perplexity: 5.82591
Epoch [2/3], Step [36000/41412], Loss: 2.4776, Perplexity: 11.9122
Epoch [2/3], Step [36100/41412], Loss: 1.8157, Perplexity: 6.14550
Epoch [2/3], Step [36200/41412], Loss: 2.1604, Perplexity: 8.67490
Epoch [2/3], Step [36300/41412], Loss: 2.5095, Perplexity: 12.2992
Epoch [2/3], Step [36400/41412], Loss: 2.0949, Perplexity: 8.12482
Epoch [2/3], Step [36500/41412], Loss: 2.0483, Perplexity: 7.75450
Epoch [2/3], Step [36600/41412], Loss: 2.5458, Perplexity: 12.7530
Epoch [2/3], Step [36700/41412], Loss: 2.2840, Perplexity: 9.81576
Epoch [2/3], Step [36800/41412], Loss: 2.6310, Perplexity: 13.8876
Epoch [2/3], Step [36900/41412], Loss: 2.1588, Perplexity: 8.66081
Epoch [2/3], Step [37000/41412], Loss: 2.8153, Perplexity: 16.6979
Epoch [2/3], Step [37100/41412], Loss: 2.6310, Perplexity: 13.8883
Epoch [2/3], Step [37200/41412], Loss: 3.4392, Perplexity: 31.1618
Epoch [2/3], Step [37300/41412], Loss: 2.2847, Perplexity: 9.82326
Epoch [2/3], Step [37400/41412], Loss: 2.0629, Perplexity: 7.86850
Epoch [2/3], Step [37500/41412], Loss: 1.8380, Perplexity: 6.28401
Epoch [2/3], Step [37600/41412], Loss: 1.8708, Perplexity: 6.49346
Epoch [2/3], Step [37700/41412], Loss: 2.4266, Perplexity: 11.3208
Epoch [2/3], Step [37800/41412], Loss: 2.0208, Perplexity: 7.54462
Epoch [2/3], Step [37900/41412], Loss: 2.6261, Perplexity: 13.8200
Epoch [2/3], Step [38000/41412], Loss: 2.4673, Perplexity: 11.7905
Epoch [2/3], Step [38100/41412], Loss: 2.4278, Perplexity: 11.3335
Epoch [2/3], Step [38200/41412], Loss: 2.4706, Perplexity: 11.8290
Epoch [2/3], Step [38300/41412], Loss: 2.1812, Perplexity: 8.85683
Epoch [2/3], Step [38400/41412], Loss: 2.3467, Perplexity: 10.4514
Epoch [2/3], Step [38500/41412], Loss: 2.3643, Perplexity: 10.6368
Epoch [2/3], Step [38600/41412], Loss: 2.0103, Perplexity: 7.46570
Epoch [2/3], Step [38700/41412], Loss: 2.6505, Perplexity: 14.1617
Epoch [2/3], Step [38800/41412], Loss: 2.2806, Perplexity: 9.78244
Epoch [2/3], Step [38900/41412], Loss: 2.4092, Perplexity: 11.1255
Epoch [2/3], Step [39000/41412], Loss: 2.6712, Perplexity: 14.4573
```

```
Epoch [2/3], Step [39100/41412], Loss: 2.1781, Perplexity: 8.82918
Epoch [2/3], Step [39200/41412], Loss: 2.6287, Perplexity: 13.8555
Epoch [2/3], Step [39300/41412], Loss: 2.5503, Perplexity: 12.8108
Epoch [2/3], Step [39400/41412], Loss: 2.4875, Perplexity: 12.0308
Epoch [2/3], Step [39500/41412], Loss: 2.3092, Perplexity: 10.06626
Epoch [2/3], Step [39600/41412], Loss: 1.8151, Perplexity: 6.14145
Epoch [2/3], Step [39700/41412], Loss: 1.9246, Perplexity: 6.85245
Epoch [2/3], Step [39800/41412], Loss: 2.6645, Perplexity: 14.3606
Epoch [2/3], Step [39900/41412], Loss: 2.4453, Perplexity: 11.53382
Epoch [2/3], Step [40000/41412], Loss: 1.9504, Perplexity: 7.03122
Epoch [2/3], Step [40100/41412], Loss: 2.4626, Perplexity: 11.7355
Epoch [2/3], Step [40200/41412], Loss: 2.0618, Perplexity: 7.860136
Epoch [2/3], Step [40300/41412], Loss: 1.9098, Perplexity: 6.75208
Epoch [2/3], Step [40400/41412], Loss: 2.1828, Perplexity: 8.87095
Epoch [2/3], Step [40500/41412], Loss: 2.8893, Perplexity: 17.9800
Epoch [2/3], Step [40600/41412], Loss: 2.3941, Perplexity: 10.9578
Epoch [2/3], Step [40700/41412], Loss: 2.4932, Perplexity: 12.0999
Epoch [2/3], Step [40800/41412], Loss: 3.2089, Perplexity: 24.7530
Epoch [2/3], Step [40900/41412], Loss: 2.2315, Perplexity: 9.31402
Epoch [2/3], Step [41000/41412], Loss: 2.1536, Perplexity: 8.61562
Epoch [2/3], Step [41100/41412], Loss: 2.4342, Perplexity: 11.4065
Epoch [2/3], Step [41200/41412], Loss: 2.0137, Perplexity: 7.49075
Epoch [2/3], Step [41300/41412], Loss: 3.0329, Perplexity: 20.7569
Epoch [2/3], Step [41400/41412], Loss: 2.3778, Perplexity: 10.7815
Epoch [2/3], Step [41412/41412], Loss: 2.4192, Perplexity: 11.2369
Epoch [2/3] completed in 4838.67 seconds.
Epoch [3/3], Step [100/41412], Loss: 2.5606, Perplexity: 12.9430
Epoch [3/3], Step [200/41412], Loss: 2.0960, Perplexity: 8.13350
Epoch [3/3], Step [300/41412], Loss: 1.9794, Perplexity: 7.238567
Epoch [3/3], Step [400/41412], Loss: 2.5852, Perplexity: 13.2661
Epoch [3/3], Step [500/41412], Loss: 1.6866, Perplexity: 5.40092
Epoch [3/3], Step [600/41412], Loss: 2.4976, Perplexity: 12.1529
Epoch [3/3], Step [700/41412], Loss: 2.4770, Perplexity: 11.9054
Epoch [3/3], Step [800/41412], Loss: 2.1338, Perplexity: 8.44719
Epoch [3/3], Step [900/41412], Loss: 2.0899, Perplexity: 8.08422
Epoch [3/3], Step [1000/41412], Loss: 2.1424, Perplexity: 8.5197
Epoch [3/3], Step [1100/41412], Loss: 2.2023, Perplexity: 9.04627
Epoch [3/3], Step [1200/41412], Loss: 2.9527, Perplexity: 19.1574
Epoch [3/3], Step [1300/41412], Loss: 2.4481, Perplexity: 11.5664
Epoch [3/3], Step [1400/41412], Loss: 2.0733, Perplexity: 7.95069
Epoch [3/3], Step [1500/41412], Loss: 2.4608, Perplexity: 11.7141
Epoch [3/3], Step [1600/41412], Loss: 2.5895, Perplexity: 13.3233
Epoch [3/3], Step [1700/41412], Loss: 4.2359, Perplexity: 69.1264
Epoch [3/3], Step [1800/41412], Loss: 2.1871, Perplexity: 8.90911
Epoch [3/3], Step [1900/41412], Loss: 2.1734, Perplexity: 8.78775
Epoch [3/3], Step [2000/41412], Loss: 2.6569, Perplexity: 14.2525
Epoch [3/3], Step [2100/41412], Loss: 1.7005, Perplexity: 5.47691
Epoch [3/3], Step [2200/41412], Loss: 2.4392, Perplexity: 11.4643
Epoch [3/3], Step [2300/41412], Loss: 2.4306, Perplexity: 11.3656
Epoch [3/3], Step [2400/41412], Loss: 2.9140, Perplexity: 18.4304
Epoch [3/3], Step [2500/41412], Loss: 1.9966, Perplexity: 7.36436
Epoch [3/3], Step [2600/41412], Loss: 2.8066, Perplexity: 16.5535
Epoch [3/3], Step [2700/41412], Loss: 2.3896, Perplexity: 10.9091
Epoch [3/3], Step [2800/41412], Loss: 2.9063, Perplexity: 18.2884
Epoch [3/3], Step [2900/41412], Loss: 1.6068, Perplexity: 4.98679
Epoch [3/3], Step [3000/41412], Loss: 2.2707, Perplexity: 9.68570
Epoch [3/3], Step [3100/41412], Loss: 1.8676, Perplexity: 6.47242
Epoch [3/3], Step [3200/41412], Loss: 2.6409, Perplexity: 14.0262
Epoch [3/3], Step [3300/41412], Loss: 1.7903, Perplexity: 5.99157
Epoch [3/3], Step [3400/41412], Loss: 2.1587, Perplexity: 8.66034
Epoch [3/3], Step [3500/41412], Loss: 2.4069, Perplexity: 11.0989
Epoch [3/3], Step [3600/41412], Loss: 2.1863, Perplexity: 8.90197
```

```
Epoch [3/3], Step [3700/41412], Loss: 2.3350, Perplexity: 10.3290
Epoch [3/3], Step [3800/41412], Loss: 2.1666, Perplexity: 8.72891
Epoch [3/3], Step [3900/41412], Loss: 1.8151, Perplexity: 6.14187
Epoch [3/3], Step [4000/41412], Loss: 1.8465, Perplexity: 6.33750
Epoch [3/3], Step [4100/41412], Loss: 2.8067, Perplexity: 16.5551
Epoch [3/3], Step [4200/41412], Loss: 2.1870, Perplexity: 8.90850
Epoch [3/3], Step [4300/41412], Loss: 2.1744, Perplexity: 8.79662
Epoch [3/3], Step [4400/41412], Loss: 2.2580, Perplexity: 9.56431
Epoch [3/3], Step [4500/41412], Loss: 1.8986, Perplexity: 6.676300
Epoch [3/3], Step [4600/41412], Loss: 2.2132, Perplexity: 9.14512
Epoch [3/3], Step [4700/41412], Loss: 2.1315, Perplexity: 8.42785
Epoch [3/3], Step [4800/41412], Loss: 2.4020, Perplexity: 11.0457
Epoch [3/3], Step [4900/41412], Loss: 1.8653, Perplexity: 6.45769
Epoch [3/3], Step [5000/41412], Loss: 2.3947, Perplexity: 10.9650
Epoch [3/3], Step [5100/41412], Loss: 2.1585, Perplexity: 8.65858
Epoch [3/3], Step [5200/41412], Loss: 2.1728, Perplexity: 8.78266
Epoch [3/3], Step [5300/41412], Loss: 2.0494, Perplexity: 7.76313
Epoch [3/3], Step [5400/41412], Loss: 2.0626, Perplexity: 7.86624
Epoch [3/3], Step [5500/41412], Loss: 2.7658, Perplexity: 15.8925
Epoch [3/3], Step [5600/41412], Loss: 2.6150, Perplexity: 13.6676
Epoch [3/3], Step [5700/41412], Loss: 2.3299, Perplexity: 10.2768
Epoch [3/3], Step [5800/41412], Loss: 2.0616, Perplexity: 7.85848
Epoch [3/3], Step [5900/41412], Loss: 2.5579, Perplexity: 12.9081
Epoch [3/3], Step [6000/41412], Loss: 2.0572, Perplexity: 7.82432
Epoch [3/3], Step [6100/41412], Loss: 2.6703, Perplexity: 14.4446
Epoch [3/3], Step [6200/41412], Loss: 2.0362, Perplexity: 7.66175
Epoch [3/3], Step [6300/41412], Loss: 2.3585, Perplexity: 10.5747
Epoch [3/3], Step [6400/41412], Loss: 2.0808, Perplexity: 8.01102
Epoch [3/3], Step [6500/41412], Loss: 1.7689, Perplexity: 5.86418
Epoch [3/3], Step [6600/41412], Loss: 2.9668, Perplexity: 19.42888
Epoch [3/3], Step [6700/41412], Loss: 1.4123, Perplexity: 4.10547
Epoch [3/3], Step [6800/41412], Loss: 2.7955, Perplexity: 16.3708
Epoch [3/3], Step [6900/41412], Loss: 2.7842, Perplexity: 16.1876
Epoch [3/3], Step [7000/41412], Loss: 1.9026, Perplexity: 6.70368
Epoch [3/3], Step [7100/41412], Loss: 2.7553, Perplexity: 15.7264
Epoch [3/3], Step [7200/41412], Loss: 2.3489, Perplexity: 10.4735
Epoch [3/3], Step [7300/41412], Loss: 2.3222, Perplexity: 10.1983
Epoch [3/3], Step [7400/41412], Loss: 2.4573, Perplexity: 11.6737
Epoch [3/3], Step [7500/41412], Loss: 2.3728, Perplexity: 10.7269
Epoch [3/3], Step [7600/41412], Loss: 1.7544, Perplexity: 5.78005
Epoch [3/3], Step [7700/41412], Loss: 2.2553, Perplexity: 9.53778
Epoch [3/3], Step [7800/41412], Loss: 2.0753, Perplexity: 7.96707
Epoch [3/3], Step [7900/41412], Loss: 2.4420, Perplexity: 11.4960
Epoch [3/3], Step [8000/41412], Loss: 1.7244, Perplexity: 5.60933
Epoch [3/3], Step [8100/41412], Loss: 2.3399, Perplexity: 10.3801
Epoch [3/3], Step [8200/41412], Loss: 2.4989, Perplexity: 12.1693
Epoch [3/3], Step [8300/41412], Loss: 2.1152, Perplexity: 8.29137
Epoch [3/3], Step [8400/41412], Loss: 2.3450, Perplexity: 10.4335
Epoch [3/3], Step [8500/41412], Loss: 2.8710, Perplexity: 17.65500
Epoch [3/3], Step [8600/41412], Loss: 1.9383, Perplexity: 6.94729
Epoch [3/3], Step [8700/41412], Loss: 2.3333, Perplexity: 10.3115
Epoch [3/3], Step [8800/41412], Loss: 2.1026, Perplexity: 8.18749
Epoch [3/3], Step [8900/41412], Loss: 2.6882, Perplexity: 14.7049
Epoch [3/3], Step [9000/41412], Loss: 2.6957, Perplexity: 14.8157
Epoch [3/3], Step [9100/41412], Loss: 2.4481, Perplexity: 11.5664
Epoch [3/3], Step [9200/41412], Loss: 2.5637, Perplexity: 12.9843
Epoch [3/3], Step [9300/41412], Loss: 2.5157, Perplexity: 12.3753
Epoch [3/3], Step [9400/41412], Loss: 2.5955, Perplexity: 13.4039
Epoch [3/3], Step [9500/41412], Loss: 2.0467, Perplexity: 7.74252
Epoch [3/3], Step [9600/41412], Loss: 1.9931, Perplexity: 7.33846
Epoch [3/3], Step [9700/41412], Loss: 2.1597, Perplexity: 8.66860
Epoch [3/3], Step [9800/41412], Loss: 1.7217, Perplexity: 5.59387
```

```
Epoch [3/3], Step [9900/41412], Loss: 2.3929, Perplexity: 10.94495
Epoch [3/3], Step [10000/41412], Loss: 2.2513, Perplexity: 9.4997
Epoch [3/3], Step [10100/41412], Loss: 2.4555, Perplexity: 11.6525
Epoch [3/3], Step [10200/41412], Loss: 2.1177, Perplexity: 8.31176
Epoch [3/3], Step [10300/41412], Loss: 2.1371, Perplexity: 8.47496
Epoch [3/3], Step [10400/41412], Loss: 2.5800, Perplexity: 13.1974
Epoch [3/3], Step [10500/41412], Loss: 1.9739, Perplexity: 7.19865
Epoch [3/3], Step [10600/41412], Loss: 2.4441, Perplexity: 11.51994
Epoch [3/3], Step [10700/41412], Loss: 2.3602, Perplexity: 10.5926
Epoch [3/3], Step [10800/41412], Loss: 1.8696, Perplexity: 6.48599
Epoch [3/3], Step [10900/41412], Loss: 1.8045, Perplexity: 6.07695
Epoch [3/3], Step [11000/41412], Loss: 2.1906, Perplexity: 8.94027
Epoch [3/3], Step [11100/41412], Loss: 1.7606, Perplexity: 5.81606
Epoch [3/3], Step [11200/41412], Loss: 2.2586, Perplexity: 9.57013
Epoch [3/3], Step [11300/41412], Loss: 2.2896, Perplexity: 9.87088
Epoch [3/3], Step [11400/41412], Loss: 2.8645, Perplexity: 17.5409
Epoch [3/3], Step [11500/41412], Loss: 2.3416, Perplexity: 10.3976
Epoch [3/3], Step [11600/41412], Loss: 2.3677, Perplexity: 10.6729
Epoch [3/3], Step [11700/41412], Loss: 2.3294, Perplexity: 10.2722
Epoch [3/3], Step [11800/41412], Loss: 2.3718, Perplexity: 10.7172
Epoch [3/3], Step [11900/41412], Loss: 2.0900, Perplexity: 8.08474
Epoch [3/3], Step [12000/41412], Loss: 2.8967, Perplexity: 18.1138
Epoch [3/3], Step [12100/41412], Loss: 1.8279, Perplexity: 6.22095
Epoch [3/3], Step [12200/41412], Loss: 2.0089, Perplexity: 7.45486
Epoch [3/3], Step [12300/41412], Loss: 2.4659, Perplexity: 11.7739
Epoch [3/3], Step [12400/41412], Loss: 1.9766, Perplexity: 7.21804
Epoch [3/3], Step [12500/41412], Loss: 2.5391, Perplexity: 12.6677
Epoch [3/3], Step [12600/41412], Loss: 1.5469, Perplexity: 4.69715
Epoch [3/3], Step [12700/41412], Loss: 2.6539, Perplexity: 14.2089
Epoch [3/3], Step [12800/41412], Loss: 2.8086, Perplexity: 16.5862
Epoch [3/3], Step [12900/41412], Loss: 2.0242, Perplexity: 7.56975
Epoch [3/3], Step [13000/41412], Loss: 2.0618, Perplexity: 7.85981
Epoch [3/3], Step [13100/41412], Loss: 4.1192, Perplexity: 61.5079
Epoch [3/3], Step [13200/41412], Loss: 2.2790, Perplexity: 9.76735
Epoch [3/3], Step [13300/41412], Loss: 2.1024, Perplexity: 8.18581
Epoch [3/3], Step [13400/41412], Loss: 2.6418, Perplexity: 14.0391
Epoch [3/3], Step [13500/41412], Loss: 2.6044, Perplexity: 13.5231
Epoch [3/3], Step [13600/41412], Loss: 2.6126, Perplexity: 13.63410
Epoch [3/3], Step [13700/41412], Loss: 1.7811, Perplexity: 5.93655
Epoch [3/3], Step [13800/41412], Loss: 1.8386, Perplexity: 6.28784
Epoch [3/3], Step [13900/41412], Loss: 2.6242, Perplexity: 13.7941
Epoch [3/3], Step [14000/41412], Loss: 2.4109, Perplexity: 11.1440
Epoch [3/3], Step [14100/41412], Loss: 2.2300, Perplexity: 9.29975
Epoch [3/3], Step [14200/41412], Loss: 2.3884, Perplexity: 10.8956
Epoch [3/3], Step [14300/41412], Loss: 2.3203, Perplexity: 10.1792
Epoch [3/3], Step [14400/41412], Loss: 1.7187, Perplexity: 5.57716
Epoch [3/3], Step [14500/41412], Loss: 2.0022, Perplexity: 7.40546
Epoch [3/3], Step [14600/41412], Loss: 2.6614, Perplexity: 14.3158
Epoch [3/3], Step [14700/41412], Loss: 1.8744, Perplexity: 6.51720
Epoch [3/3], Step [14800/41412], Loss: 2.2489, Perplexity: 9.47737
Epoch [3/3], Step [14900/41412], Loss: 1.8746, Perplexity: 6.51802
Epoch [3/3], Step [15000/41412], Loss: 2.2042, Perplexity: 9.06278
Epoch [3/3], Step [15100/41412], Loss: 2.0820, Perplexity: 8.02050
Epoch [3/3], Step [15200/41412], Loss: 2.2478, Perplexity: 9.466798
Epoch [3/3], Step [15300/41412], Loss: 2.1358, Perplexity: 8.46367
Epoch [3/3], Step [15400/41412], Loss: 2.1038, Perplexity: 8.19717
Epoch [3/3], Step [15500/41412], Loss: 2.0289, Perplexity: 7.60575
Epoch [3/3], Step [15600/41412], Loss: 2.0881, Perplexity: 8.06936
Epoch [3/3], Step [15700/41412], Loss: 2.5523, Perplexity: 12.8365
Epoch [3/3], Step [15800/41412], Loss: 2.4214, Perplexity: 11.2620
Epoch [3/3], Step [15900/41412], Loss: 2.4822, Perplexity: 11.9676
Epoch [3/3], Step [16000/41412], Loss: 2.3103, Perplexity: 10.0770
```

```
Epoch [3/3], Step [16100/41412], Loss: 2.2263, Perplexity: 9.26551
Epoch [3/3], Step [16200/41412], Loss: 1.9883, Perplexity: 7.30345
Epoch [3/3], Step [16300/41412], Loss: 2.0122, Perplexity: 7.48003
Epoch [3/3], Step [16400/41412], Loss: 1.8993, Perplexity: 6.68146
Epoch [3/3], Step [16500/41412], Loss: 2.2302, Perplexity: 9.30202
Epoch [3/3], Step [16600/41412], Loss: 2.5345, Perplexity: 12.6095
Epoch [3/3], Step [16700/41412], Loss: 2.2110, Perplexity: 9.12451
Epoch [3/3], Step [16800/41412], Loss: 2.4306, Perplexity: 11.3657
Epoch [3/3], Step [16900/41412], Loss: 2.5221, Perplexity: 12.4543
Epoch [3/3], Step [17000/41412], Loss: 2.4861, Perplexity: 12.0148
Epoch [3/3], Step [17100/41412], Loss: 2.4083, Perplexity: 11.1146
Epoch [3/3], Step [17200/41412], Loss: 1.9557, Perplexity: 7.06925
Epoch [3/3], Step [17300/41412], Loss: 2.2627, Perplexity: 9.60881
Epoch [3/3], Step [17400/41412], Loss: 2.3665, Perplexity: 10.6599
Epoch [3/3], Step [17500/41412], Loss: 1.8461, Perplexity: 6.33508
Epoch [3/3], Step [17600/41412], Loss: 2.3218, Perplexity: 10.1940
Epoch [3/3], Step [17700/41412], Loss: 2.5382, Perplexity: 12.6573
Epoch [3/3], Step [17800/41412], Loss: 2.2946, Perplexity: 9.92003
Epoch [3/3], Step [17900/41412], Loss: 2.0522, Perplexity: 7.78525
Epoch [3/3], Step [18000/41412], Loss: 2.5651, Perplexity: 13.0015
Epoch [3/3], Step [18100/41412], Loss: 2.1586, Perplexity: 8.65881
Epoch [3/3], Step [18200/41412], Loss: 2.2411, Perplexity: 9.40353
Epoch [3/3], Step [18300/41412], Loss: 2.9505, Perplexity: 19.1162
Epoch [3/3], Step [18400/41412], Loss: 2.0823, Perplexity: 8.02295
Epoch [3/3], Step [18500/41412], Loss: 2.0885, Perplexity: 8.07304
Epoch [3/3], Step [18600/41412], Loss: 1.9209, Perplexity: 6.82734
Epoch [3/3], Step [18700/41412], Loss: 2.9155, Perplexity: 18.4590
Epoch [3/3], Step [18800/41412], Loss: 2.4847, Perplexity: 11.9974
Epoch [3/3], Step [18900/41412], Loss: 2.6759, Perplexity: 14.5254
Epoch [3/3], Step [19000/41412], Loss: 1.6924, Perplexity: 5.43263
Epoch [3/3], Step [19100/41412], Loss: 2.5067, Perplexity: 12.2641
Epoch [3/3], Step [19200/41412], Loss: 2.0832, Perplexity: 8.03029
Epoch [3/3], Step [19300/41412], Loss: 1.9251, Perplexity: 6.85608
Epoch [3/3], Step [19400/41412], Loss: 2.4125, Perplexity: 11.1619
Epoch [3/3], Step [19500/41412], Loss: 1.8190, Perplexity: 6.16548
Epoch [3/3], Step [19600/41412], Loss: 1.6928, Perplexity: 5.43478
Epoch [3/3], Step [19700/41412], Loss: 2.4700, Perplexity: 11.8229
Epoch [3/3], Step [19800/41412], Loss: 2.0471, Perplexity: 7.74560
Epoch [3/3], Step [19900/41412], Loss: 2.1474, Perplexity: 8.56297
Epoch [3/3], Step [20000/41412], Loss: 2.3383, Perplexity: 10.3640
Epoch [3/3], Step [20100/41412], Loss: 2.3235, Perplexity: 10.2114
Epoch [3/3], Step [20200/41412], Loss: 2.0816, Perplexity: 8.01740
Epoch [3/3], Step [20300/41412], Loss: 2.1913, Perplexity: 8.94729
Epoch [3/3], Step [20400/41412], Loss: 2.0243, Perplexity: 7.57092
Epoch [3/3], Step [20500/41412], Loss: 1.8740, Perplexity: 6.51433
Epoch [3/3], Step [20600/41412], Loss: 2.3782, Perplexity: 10.7854
Epoch [3/3], Step [20700/41412], Loss: 2.2857, Perplexity: 9.83210
Epoch [3/3], Step [20800/41412], Loss: 2.1700, Perplexity: 8.75832
Epoch [3/3], Step [20900/41412], Loss: 2.0795, Perplexity: 8.00018
Epoch [3/3], Step [21000/41412], Loss: 2.3524, Perplexity: 10.5104
Epoch [3/3], Step [21100/41412], Loss: 2.1877, Perplexity: 8.91432
Epoch [3/3], Step [21200/41412], Loss: 2.0363, Perplexity: 7.66219
Epoch [3/3], Step [21300/41412], Loss: 2.1174, Perplexity: 8.30977
Epoch [3/3], Step [21400/41412], Loss: 1.9751, Perplexity: 7.20750
Epoch [3/3], Step [21500/41412], Loss: 2.4191, Perplexity: 11.2360
Epoch [3/3], Step [21600/41412], Loss: 2.7032, Perplexity: 14.9277
Epoch [3/3], Step [21700/41412], Loss: 2.0375, Perplexity: 7.67168
Epoch [3/3], Step [21800/41412], Loss: 2.4007, Perplexity: 11.0306
Epoch [3/3], Step [21900/41412], Loss: 2.7377, Perplexity: 15.4511
Epoch [3/3], Step [22000/41412], Loss: 2.6135, Perplexity: 13.6471
Epoch [3/3], Step [22100/41412], Loss: 2.3681, Perplexity: 10.6772
Epoch [3/3], Step [22200/41412], Loss: 2.0647, Perplexity: 7.88277
```

```
Epoch [3/3], Step [22300/41412], Loss: 1.9750, Perplexity: 7.20665
Epoch [3/3], Step [22400/41412], Loss: 2.2368, Perplexity: 9.36336
Epoch [3/3], Step [22500/41412], Loss: 2.4354, Perplexity: 11.4207
Epoch [3/3], Step [22600/41412], Loss: 2.6901, Perplexity: 14.7331
Epoch [3/3], Step [22700/41412], Loss: 2.3328, Perplexity: 10.3065
Epoch [3/3], Step [22800/41412], Loss: 2.1861, Perplexity: 8.90018
Epoch [3/3], Step [22900/41412], Loss: 2.1884, Perplexity: 8.92088
Epoch [3/3], Step [23000/41412], Loss: 2.2225, Perplexity: 9.23089
Epoch [3/3], Step [23100/41412], Loss: 2.3360, Perplexity: 10.3399
Epoch [3/3], Step [23200/41412], Loss: 1.6679, Perplexity: 5.300949
Epoch [3/3], Step [23300/41412], Loss: 2.4860, Perplexity: 12.0132
Epoch [3/3], Step [23400/41412], Loss: 2.4743, Perplexity: 11.8736
Epoch [3/3], Step [23500/41412], Loss: 2.0994, Perplexity: 8.16136
Epoch [3/3], Step [23600/41412], Loss: 2.6829, Perplexity: 14.6281
Epoch [3/3], Step [23700/41412], Loss: 2.2870, Perplexity: 9.84493
Epoch [3/3], Step [23800/41412], Loss: 2.2632, Perplexity: 9.61391
Epoch [3/3], Step [23900/41412], Loss: 2.1641, Perplexity: 8.70696
Epoch [3/3], Step [24000/41412], Loss: 2.3658, Perplexity: 10.6530
Epoch [3/3], Step [24100/41412], Loss: 2.0293, Perplexity: 7.60869
Epoch [3/3], Step [24200/41412], Loss: 2.1723, Perplexity: 8.77890
Epoch [3/3], Step [24300/41412], Loss: 1.8250, Perplexity: 6.20298
Epoch [3/3], Step [24400/41412], Loss: 2.6557, Perplexity: 14.2345
Epoch [3/3], Step [24500/41412], Loss: 2.3177, Perplexity: 10.1524
Epoch [3/3], Step [24600/41412], Loss: 1.8701, Perplexity: 6.48874
Epoch [3/3], Step [24700/41412], Loss: 2.0450, Perplexity: 7.72884
Epoch [3/3], Step [24800/41412], Loss: 2.2898, Perplexity: 9.87349
Epoch [3/3], Step [24900/41412], Loss: 2.2597, Perplexity: 9.57974
Epoch [3/3], Step [25000/41412], Loss: 2.0547, Perplexity: 7.80466
Epoch [3/3], Step [25100/41412], Loss: 2.0455, Perplexity: 7.73288
Epoch [3/3], Step [25200/41412], Loss: 2.4121, Perplexity: 11.1570
Epoch [3/3], Step [25300/41412], Loss: 2.1522, Perplexity: 8.60422
Epoch [3/3], Step [25400/41412], Loss: 2.3604, Perplexity: 10.5956
Epoch [3/3], Step [25500/41412], Loss: 1.6135, Perplexity: 5.02054
Epoch [3/3], Step [25600/41412], Loss: 2.5803, Perplexity: 13.2008
Epoch [3/3], Step [25700/41412], Loss: 2.1428, Perplexity: 8.52303
Epoch [3/3], Step [25800/41412], Loss: 2.1785, Perplexity: 8.83312
Epoch [3/3], Step [25900/41412], Loss: 2.2367, Perplexity: 9.36265
Epoch [3/3], Step [26000/41412], Loss: 2.2742, Perplexity: 9.72007
Epoch [3/3], Step [26100/41412], Loss: 2.3233, Perplexity: 10.2095
Epoch [3/3], Step [26200/41412], Loss: 2.3573, Perplexity: 10.5619
Epoch [3/3], Step [26300/41412], Loss: 2.1220, Perplexity: 8.34770
Epoch [3/3], Step [26400/41412], Loss: 2.0571, Perplexity: 7.82329
Epoch [3/3], Step [26500/41412], Loss: 2.0444, Perplexity: 7.72457
Epoch [3/3], Step [26600/41412], Loss: 2.2533, Perplexity: 9.51864
Epoch [3/3], Step [26700/41412], Loss: 2.4487, Perplexity: 11.5737
Epoch [3/3], Step [26800/41412], Loss: 2.6555, Perplexity: 14.2321
Epoch [3/3], Step [26900/41412], Loss: 2.2181, Perplexity: 9.19029
Epoch [3/3], Step [27000/41412], Loss: 2.1285, Perplexity: 8.40213
Epoch [3/3], Step [27100/41412], Loss: 2.7884, Perplexity: 16.2549
Epoch [3/3], Step [27200/41412], Loss: 2.0827, Perplexity: 8.02608
Epoch [3/3], Step [27300/41412], Loss: 1.7944, Perplexity: 6.01593
Epoch [3/3], Step [27400/41412], Loss: 2.4425, Perplexity: 11.5013
Epoch [3/3], Step [27500/41412], Loss: 2.4190, Perplexity: 11.2348
Epoch [3/3], Step [27600/41412], Loss: 2.1904, Perplexity: 8.93894
Epoch [3/3], Step [27700/41412], Loss: 1.8819, Perplexity: 6.56595
Epoch [3/3], Step [27800/41412], Loss: 2.6056, Perplexity: 13.5395
Epoch [3/3], Step [27900/41412], Loss: 2.7082, Perplexity: 15.0018
Epoch [3/3], Step [28000/41412], Loss: 2.1273, Perplexity: 8.39195
Epoch [3/3], Step [28100/41412], Loss: 2.4740, Perplexity: 11.8704
Epoch [3/3], Step [28200/41412], Loss: 2.3557, Perplexity: 10.5452
Epoch [3/3], Step [28300/41412], Loss: 3.0210, Perplexity: 20.5120
Epoch [3/3], Step [28400/41412], Loss: 1.8966, Perplexity: 6.66347
```

```
Epoch [3/3], Step [28500/41412], Loss: 2.2416, Perplexity: 9.40834
Epoch [3/3], Step [28600/41412], Loss: 2.5418, Perplexity: 12.7020
Epoch [3/3], Step [28700/41412], Loss: 2.5421, Perplexity: 12.7059
Epoch [3/3], Step [28800/41412], Loss: 2.3044, Perplexity: 10.0177
Epoch [3/3], Step [28900/41412], Loss: 2.3090, Perplexity: 10.0641
Epoch [3/3], Step [29000/41412], Loss: 2.6760, Perplexity: 14.5272
Epoch [3/3], Step [29100/41412], Loss: 2.3128, Perplexity: 10.1024
Epoch [3/3], Step [29200/41412], Loss: 2.0356, Perplexity: 7.65687
Epoch [3/3], Step [29300/41412], Loss: 2.4775, Perplexity: 11.9119
Epoch [3/3], Step [29400/41412], Loss: 2.1332, Perplexity: 8.44165
Epoch [3/3], Step [29500/41412], Loss: 2.1846, Perplexity: 8.88673
Epoch [3/3], Step [29600/41412], Loss: 2.4276, Perplexity: 11.3318
Epoch [3/3], Step [29700/41412], Loss: 2.1080, Perplexity: 8.23162
Epoch [3/3], Step [29800/41412], Loss: 2.1312, Perplexity: 8.42520
Epoch [3/3], Step [29900/41412], Loss: 2.5918, Perplexity: 13.3533
Epoch [3/3], Step [30000/41412], Loss: 1.8574, Perplexity: 6.40729
Epoch [3/3], Step [30100/41412], Loss: 2.2044, Perplexity: 9.06457
Epoch [3/3], Step [30200/41412], Loss: 2.0598, Perplexity: 7.84405
Epoch [3/3], Step [30300/41412], Loss: 1.9862, Perplexity: 7.28811
Epoch [3/3], Step [30400/41412], Loss: 2.6248, Perplexity: 13.8016
Epoch [3/3], Step [30500/41412], Loss: 2.7042, Perplexity: 14.9428
Epoch [3/3], Step [30600/41412], Loss: 2.3447, Perplexity: 10.4301
Epoch [3/3], Step [30700/41412], Loss: 2.1157, Perplexity: 8.29542
Epoch [3/3], Step [30800/41412], Loss: 2.1993, Perplexity: 9.01908
Epoch [3/3], Step [30900/41412], Loss: 2.6402, Perplexity: 14.0161
Epoch [3/3], Step [31000/41412], Loss: 2.2801, Perplexity: 9.77788
Epoch [3/3], Step [31100/41412], Loss: 2.0090, Perplexity: 7.45613
Epoch [3/3], Step [31200/41412], Loss: 2.7840, Perplexity: 16.1837
Epoch [3/3], Step [31300/41412], Loss: 2.7238, Perplexity: 15.2382
Epoch [3/3], Step [31400/41412], Loss: 1.7897, Perplexity: 5.987953
Epoch [3/3], Step [31500/41412], Loss: 2.7372, Perplexity: 15.4442
Epoch [3/3], Step [31600/41412], Loss: 2.3849, Perplexity: 10.8584
Epoch [3/3], Step [31700/41412], Loss: 2.2952, Perplexity: 9.92655
Epoch [3/3], Step [31800/41412], Loss: 2.0335, Perplexity: 7.64111
Epoch [3/3], Step [31900/41412], Loss: 2.2730, Perplexity: 9.70847
Epoch [3/3], Step [32000/41412], Loss: 2.2110, Perplexity: 9.12514
Epoch [3/3], Step [32100/41412], Loss: 2.7685, Perplexity: 15.9339
Epoch [3/3], Step [32200/41412], Loss: 2.0966, Perplexity: 8.13815
Epoch [3/3], Step [32300/41412], Loss: 1.9339, Perplexity: 6.91658
Epoch [3/3], Step [32400/41412], Loss: 2.2343, Perplexity: 9.33966
Epoch [3/3], Step [32500/41412], Loss: 2.5791, Perplexity: 13.1856
Epoch [3/3], Step [32600/41412], Loss: 2.1492, Perplexity: 8.57845
Epoch [3/3], Step [32700/41412], Loss: 2.8062, Perplexity: 16.5470
Epoch [3/3], Step [32800/41412], Loss: 2.5276, Perplexity: 12.5238
Epoch [3/3], Step [32900/41412], Loss: 2.4936, Perplexity: 12.1042
Epoch [3/3], Step [33000/41412], Loss: 2.5318, Perplexity: 12.5756
Epoch [3/3], Step [33100/41412], Loss: 2.1046, Perplexity: 8.20407
Epoch [3/3], Step [33200/41412], Loss: 1.9601, Perplexity: 7.10020
Epoch [3/3], Step [33300/41412], Loss: 1.7325, Perplexity: 5.65475
Epoch [3/3], Step [33400/41412], Loss: 2.4366, Perplexity: 11.4335
Epoch [3/3], Step [33500/41412], Loss: 1.9565, Perplexity: 7.07437
Epoch [3/3], Step [33600/41412], Loss: 2.9997, Perplexity: 20.0802
Epoch [3/3], Step [33700/41412], Loss: 1.8064, Perplexity: 6.08878
Epoch [3/3], Step [33800/41412], Loss: 2.1561, Perplexity: 8.63726
Epoch [3/3], Step [33900/41412], Loss: 2.2016, Perplexity: 9.03985
Epoch [3/3], Step [34000/41412], Loss: 2.1947, Perplexity: 8.97734
Epoch [3/3], Step [34100/41412], Loss: 2.3876, Perplexity: 10.8877
Epoch [3/3], Step [34200/41412], Loss: 2.1736, Perplexity: 8.79028
Epoch [3/3], Step [34300/41412], Loss: 2.1602, Perplexity: 8.67258
Epoch [3/3], Step [34400/41412], Loss: 2.7578, Perplexity: 15.7652
Epoch [3/3], Step [34500/41412], Loss: 2.2743, Perplexity: 9.72134
Epoch [3/3], Step [34600/41412], Loss: 3.1433, Perplexity: 23.1809
```

```
Epoch [3/3], Step [34700/41412], Loss: 2.2409, Perplexity: 9.40224
Epoch [3/3], Step [34800/41412], Loss: 2.1038, Perplexity: 8.19703
Epoch [3/3], Step [34900/41412], Loss: 2.4004, Perplexity: 11.0281
Epoch [3/3], Step [35000/41412], Loss: 1.9282, Perplexity: 6.87697
Epoch [3/3], Step [35100/41412], Loss: 2.1165, Perplexity: 8.30221
Epoch [3/3], Step [35200/41412], Loss: 2.4953, Perplexity: 12.1249
Epoch [3/3], Step [35300/41412], Loss: 2.3808, Perplexity: 10.8140
Epoch [3/3], Step [35400/41412], Loss: 1.8526, Perplexity: 6.37655
Epoch [3/3], Step [35500/41412], Loss: 2.2898, Perplexity: 9.87288
Epoch [3/3], Step [35600/41412], Loss: 2.2128, Perplexity: 9.14127
Epoch [3/3], Step [35700/41412], Loss: 2.2388, Perplexity: 9.38164
Epoch [3/3], Step [35800/41412], Loss: 2.4066, Perplexity: 11.0959
Epoch [3/3], Step [35900/41412], Loss: 1.7914, Perplexity: 5.99817
Epoch [3/3], Step [36000/41412], Loss: 2.1537, Perplexity: 8.61659
Epoch [3/3], Step [36100/41412], Loss: 2.3943, Perplexity: 10.9606
Epoch [3/3], Step [36200/41412], Loss: 1.8120, Perplexity: 6.12280
Epoch [3/3], Step [36300/41412], Loss: 2.0058, Perplexity: 7.43200
Epoch [3/3], Step [36400/41412], Loss: 1.8112, Perplexity: 6.11798
Epoch [3/3], Step [36500/41412], Loss: 2.0912, Perplexity: 8.09454
Epoch [3/3], Step [36600/41412], Loss: 2.0258, Perplexity: 7.58268
Epoch [3/3], Step [36700/41412], Loss: 1.9287, Perplexity: 6.88089
Epoch [3/3], Step [36800/41412], Loss: 2.1153, Perplexity: 8.29216
Epoch [3/3], Step [36900/41412], Loss: 2.7525, Perplexity: 15.6820
Epoch [3/3], Step [37000/41412], Loss: 1.9414, Perplexity: 6.96877
Epoch [3/3], Step [37100/41412], Loss: 1.9971, Perplexity: 7.36756
Epoch [3/3], Step [37200/41412], Loss: 2.2963, Perplexity: 9.93745
Epoch [3/3], Step [37300/41412], Loss: 2.1301, Perplexity: 8.41605
Epoch [3/3], Step [37400/41412], Loss: 2.6840, Perplexity: 14.6431
Epoch [3/3], Step [37500/41412], Loss: 2.2836, Perplexity: 9.81183
Epoch [3/3], Step [37600/41412], Loss: 2.4083, Perplexity: 11.1153
Epoch [3/3], Step [37700/41412], Loss: 2.0824, Perplexity: 8.02359
Epoch [3/3], Step [37800/41412], Loss: 2.2014, Perplexity: 9.03739
Epoch [3/3], Step [37900/41412], Loss: 1.7492, Perplexity: 5.75000
Epoch [3/3], Step [38000/41412], Loss: 1.8727, Perplexity: 6.50586
Epoch [3/3], Step [38100/41412], Loss: 1.9844, Perplexity: 7.27448
Epoch [3/3], Step [38200/41412], Loss: 1.9922, Perplexity: 7.33151
Epoch [3/3], Step [38300/41412], Loss: 2.9182, Perplexity: 18.50791
Epoch [3/3], Step [38400/41412], Loss: 2.8293, Perplexity: 16.9335
Epoch [3/3], Step [38500/41412], Loss: 2.9641, Perplexity: 19.3781
Epoch [3/3], Step [38600/41412], Loss: 2.7414, Perplexity: 15.5092
Epoch [3/3], Step [38700/41412], Loss: 1.6834, Perplexity: 5.38400
Epoch [3/3], Step [38800/41412], Loss: 2.6884, Perplexity: 14.7075
Epoch [3/3], Step [38900/41412], Loss: 2.0583, Perplexity: 7.83240
Epoch [3/3], Step [39000/41412], Loss: 1.9284, Perplexity: 6.87833
Epoch [3/3], Step [39100/41412], Loss: 2.4339, Perplexity: 11.4034
Epoch [3/3], Step [39200/41412], Loss: 2.3530, Perplexity: 10.5168
Epoch [3/3], Step [39300/41412], Loss: 2.9211, Perplexity: 18.5614
Epoch [3/3], Step [39400/41412], Loss: 2.2074, Perplexity: 9.09194
Epoch [3/3], Step [39500/41412], Loss: 1.9297, Perplexity: 6.88740
Epoch [3/3], Step [39600/41412], Loss: 2.9692, Perplexity: 19.4771
Epoch [3/3], Step [39700/41412], Loss: 2.0721, Perplexity: 7.94172
Epoch [3/3], Step [39800/41412], Loss: 1.9719, Perplexity: 7.18419
Epoch [3/3], Step [39900/41412], Loss: 2.5787, Perplexity: 13.1805
Epoch [3/3], Step [40000/41412], Loss: 2.3423, Perplexity: 10.4048
Epoch [3/3], Step [40100/41412], Loss: 2.1347, Perplexity: 8.45473
Epoch [3/3], Step [40200/41412], Loss: 1.7183, Perplexity: 5.57533
Epoch [3/3], Step [40300/41412], Loss: 2.4107, Perplexity: 11.14152
Epoch [3/3], Step [40400/41412], Loss: 1.6757, Perplexity: 5.34260
Epoch [3/3], Step [40500/41412], Loss: 1.9819, Perplexity: 7.25661
Epoch [3/3], Step [40600/41412], Loss: 2.3183, Perplexity: 10.1581
Epoch [3/3], Step [40700/41412], Loss: 2.3983, Perplexity: 11.0043
Epoch [3/3], Step [40800/41412], Loss: 2.4523, Perplexity: 11.6151
```

```
Epoch [3/3], Step [40900/41412], Loss: 2.5623, Perplexity: 12.9653
Epoch [3/3], Step [41000/41412], Loss: 1.8341, Perplexity: 6.25938
Epoch [3/3], Step [41100/41412], Loss: 1.9880, Perplexity: 7.30098
Epoch [3/3], Step [41200/41412], Loss: 2.0242, Perplexity: 7.57039
Epoch [3/3], Step [41300/41412], Loss: 2.0607, Perplexity: 7.85116
Epoch [3/3], Step [41400/41412], Loss: 2.3088, Perplexity: 10.0627
Epoch [3/3], Step [41412/41412], Loss: 1.9903, Perplexity: 7.31796
Epoch [3/3] completed in 4860.76 seconds.
```

Step 3: (Optional) Validate your Model

To assess potential overfitting, one approach is to assess performance on a validation set. If you decide to do this **optional** task, you are required to first complete all of the steps in the next notebook in the sequence (**3_Inference.ipynb**); as part of that notebook, you will write and test code (specifically, the `sample` method in the `DecoderRNN` class) that uses your RNN decoder to generate captions. That code will prove incredibly useful here.

If you decide to validate your model, please do not edit the data loader in **data_loader.py**. Instead, create a new file named **data_loader_val.py** containing the code for obtaining the data loader for the validation data. You can access:

- the validation images at filepath `'/opt/cocoapi/images/train2014/'`, and
- the validation image caption annotation file at filepath `'/opt/cocoapi/annotations/captions_val2014.json'`.

The suggested approach to validating your model involves creating a json file such as this one containing your model's predicted captions for the validation images. Then, you can write your own script or use one that you find online to calculate the BLEU score of your model. You can read more about the BLEU score, along with other evaluation metrics (such as TEOR and Cider) in section 4.1 of this paper. For more information about how to use the annotation file, check out the website for the COCO dataset.

In [8]:
```
import torch
print(torch.cuda.is_available())
```

```
True
```

In [9]:
```
print(torch.cuda.get_device_name(0))
```

```
NVIDIA GeForce RTX 3080
```