rnn_c

May 1, 2019

Reference: Tensorflow tutorials(https://www.tensorflow.org/tutorials/sequences/text_generation)

```
[1]: import tensorflow as tf
    tf.enable_eager_execution()

import numpy as np
    import os
    import time
    import matplotlib.pyplot as plt

[2]: # Read, then decode for py2 compat.
    text = open('huge_c.txt', 'rb').read().decode(encoding='utf-8')
    # length of text is the number of characters in it
    print ('Length of text: {} characters'.format(len(text)))
```

Length of text: 26346 characters

```
[3]: # The unique characters in the file
vocab = sorted(set(text))
print ('{} unique characters'.format(len(vocab)))
```

85 unique characters

```
[4]: # Creating a mapping from unique characters to indices
    char2idx = {u:i for i, u in enumerate(vocab)}
    idx2char = np.array(vocab)

    text_as_int = np.array([char2idx[c] for c in text])

[5]: # The maximum length sentence we want for a single input in characters
    seq_length = 100
    examples_per_epoch = len(text)//seq_length

# Create training examples / targets
    char_dataset = tf.data.Dataset.from_tensor_slices(text_as_int)
    sequences = char_dataset.batch(seq_length+1, drop_remainder = True)

for item in sequences.take(5):
```

```
print(repr(''.join(idx2char[item.numpy()])))
   '// SPDX-License-Identifier: GPL-2.0\n#include "audit.h"\n#include
   <linux/fsnotify_backend.h>\n#include <'</pre>
   'linux/namei.h>\n#include <linux/mount.h>\n#include <linux/kthread.h>\n#include
   <linux/refcount.h>\n#inclu'
   'de <linux/slab.h>\n\nstruct audit_tree;\nstruct audit_chunk;\n\nstruct
   audit_tree {\n\trefcount_t count;\n\tin'
   't goner; \n\tstruct audit_chunk *root; \n\tstruct list_head chunks; \n\tstruct
   list head rules; \n\tstruct list h'
   'ead list;\n\tstruct list_head same_root;\n\tstruct rcu_head head;\n\tchar
   pathname[];\n};\n\nstruct audit_chun'
[6]: def split_input_target(chunk):
        input_text = chunk[:-1]
        target_text = chunk[1:]
        return input_text, target_text
   dataset = sequences.map(split_input_target)
[7]: for input_example, target_example in dataset.take(1):
        print ('Input data: ', repr(''.join(idx2char[input_example.numpy()])))
        print ('Target data:', repr(''.join(idx2char[target_example.numpy()])))
   Input data: '// SPDX-License-Identifier: GPL-2.0\n#include "audit.h"\n#include
   <linux/fsnotify_backend.h>\n#include '
   Target data: '/ SPDX-License-Identifier: GPL-2.0\n#include "audit.h"\n#include
   <linux/fsnotify_backend.h>\n#include <'</pre>
[8]: for i, (input_idx, target_idx) in enumerate(zip(input_example[:5],_
     →target_example[:5])):
        print("Step {:4d}".format(i))
        print(" input: {} ({:s})".format(input_idx, repr(idx2char[input_idx])))
        print(" expected output: {} ({:s})".format(target_idx,__
     →repr(idx2char[target_idx])))
   Step
           0
     input: 16 ('/')
     expected output: 16 ('/')
   Step
     input: 16 ('/')
     expected output: 2 (' ')
   Step
           2
     input: 2 (' ')
     expected output: 44 ('S')
   Step
     input: 44 ('S')
```

```
Step
     input: 42 ('P')
     expected output: 31 ('D')
 [9]: # Batch size
    BATCH SIZE = 64
    steps_per_epoch = examples_per_epoch//BATCH_SIZE
    # Buffer size to shuffle the dataset
    # (TF data is designed to work with possibly infinite sequences,
    # so it doesn't attempt to shuffle the entire sequence in memory. Instead,
    # it maintains a buffer in which it shuffles elements).
    BUFFER_SIZE = 10000
    dataset = dataset.shuffle(BUFFER_SIZE).batch(BATCH_SIZE, drop_remainder=True)
    dataset
[9]: <BatchDataset shapes: ((64, 100), (64, 100)), types: (tf.int64, tf.int64)>
[10]: # Length of the vocabulary in chars
    vocab_size = len(vocab)
    # The embedding dimension
    embedding_dim = 256
    # Number of RNN units
    rnn units = 1024
[11]: model = tf.keras.Sequential()
    model.add(tf.keras.layers.Embedding(len(vocab), embedding_dim,
                               batch_input_shape=[BATCH_SIZE, None]))
    model.add(tf.keras.layers.CuDNNGRU(rnn_units,
           return_sequences=True,
           recurrent_initializer='glorot_uniform',
           stateful=True))
    model.add(tf.keras.layers.Dense(len(vocab)))
    print(model.summary())
   Layer (type)
                             Output Shape
                                                    Param #
   ______
                             (64, None, 256)
   embedding (Embedding)
                                                     21760
   cu_dnngru (CuDNNGRU)
                             (64, None, 1024)
                                                  3938304
    ______
```

expected output: 42 ('P')

dense (Dense)

87125

(64, None, 85)

```
Total params: 4,047,189
    Trainable params: 4,047,189
    Non-trainable params: 0
    None
[12]: for input_example_batch, target_example_batch in dataset.take(1):
        example_batch_predictions = model(input_example_batch)
        print(example_batch_predictions.shape, "# (batch_size, sequence_length,u
     →vocab_size)")
    (64, 100, 85) # (batch_size, sequence_length, vocab_size)
[13]: def loss(labels, logits):
        return tf.nn.sparse_softmax_cross_entropy_with_logits(labels=labels,_
     →logits=logits)
    model.compile(
        optimizer = tf.train.AdamOptimizer(),
        loss = loss)
[14]: # Directory where the checkpoints will be saved
    checkpoint_dir = './training_checkpoints'
     # Name of the checkpoint files
    checkpoint_prefix = os.path.join(checkpoint_dir, "ckpt_c_{epoch}")
    checkpoint_callback=tf.keras.callbacks.ModelCheckpoint(
        filepath=checkpoint_prefix,
        save_weights_only=True)
[15]: EPOCHS=200
     # history = model.fit(dataset.repeat(), epochs=EPOCHS,_
     →steps_per_epoch=steps_per_epoch, callbacks=[checkpoint_callback])
    history = model.fit(dataset.repeat(), epochs=EPOCHS,__
     →steps_per_epoch=steps_per_epoch, callbacks=[checkpoint_callback])
    Epoch 1/200
    Epoch 2/200
    4/4 [============== ] - Os 74ms/step - loss: 4.8581
    Epoch 3/200
    4/4 [=============== ] - Os 77ms/step - loss: 4.1511
    Epoch 4/200
    4/4 [============== ] - Os 83ms/step - loss: 3.9667
    Epoch 5/200
```

```
Epoch 6/200
Epoch 7/200
Epoch 8/200
Epoch 9/200
4/4 [============== ] - Os 79ms/step - loss: 2.9910
Epoch 10/200
Epoch 11/200
Epoch 12/200
Epoch 13/200
Epoch 14/200
Epoch 15/200
Epoch 16/200
Epoch 17/200
Epoch 18/200
Epoch 19/200
Epoch 20/200
Epoch 21/200
Epoch 22/200
Epoch 23/200
Epoch 24/200
4/4 [============== ] - Os 75ms/step - loss: 1.8763
Epoch 25/200
4/4 [=============== ] - Os 73ms/step - loss: 1.8223
Epoch 26/200
4/4 [============= ] - Os 81ms/step - loss: 1.7843
Epoch 27/200
Epoch 28/200
Epoch 29/200
```

```
Epoch 30/200
Epoch 31/200
4/4 [=================== ] - Os 81ms/step - loss: 1.5687
Epoch 32/200
Epoch 33/200
4/4 [============== ] - Os 81ms/step - loss: 1.4916
Epoch 34/200
Epoch 35/200
Epoch 36/200
4/4 [============== ] - Os 74ms/step - loss: 1.3637
Epoch 37/200
Epoch 38/200
4/4 [============== ] - Os 76ms/step - loss: 1.2990
Epoch 39/200
Epoch 40/200
Epoch 41/200
Epoch 42/200
4/4 [=============== ] - 1s 154ms/step - loss: 1.1496
Epoch 43/200
Epoch 44/200
Epoch 45/200
Epoch 46/200
Epoch 47/200
Epoch 48/200
4/4 [=============== ] - Os 80ms/step - loss: 0.9671
Epoch 49/200
4/4 [============== ] - Os 76ms/step - loss: 0.9359
Epoch 50/200
Epoch 51/200
Epoch 52/200
Epoch 53/200
```

```
Epoch 54/200
Epoch 55/200
4/4 [==================== ] - Os 74ms/step - loss: 0.7647
Epoch 56/200
Epoch 57/200
4/4 [=============== ] - 0s 89ms/step - loss: 0.7218
Epoch 58/200
Epoch 59/200
Epoch 60/200
4/4 [============== ] - Os 73ms/step - loss: 0.6509
Epoch 61/200
Epoch 62/200
Epoch 63/200
Epoch 64/200
Epoch 65/200
Epoch 66/200
Epoch 67/200
Epoch 68/200
Epoch 69/200
Epoch 70/200
Epoch 71/200
Epoch 72/200
4/4 [============== ] - Os 78ms/step - loss: 0.4293
Epoch 73/200
4/4 [============== ] - Os 86ms/step - loss: 0.4130
Epoch 74/200
Epoch 75/200
Epoch 76/200
Epoch 77/200
```

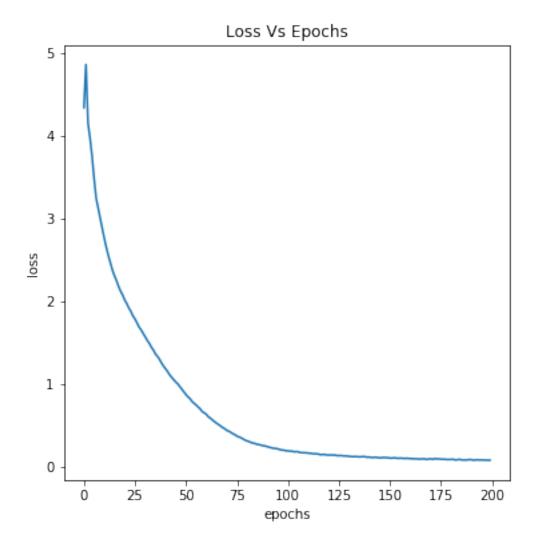
```
Epoch 78/200
4/4 [================ ] - Os 118ms/step - loss: 0.3493
Epoch 79/200
Epoch 80/200
Epoch 81/200
4/4 [=============== ] - Os 114ms/step - loss: 0.3112
Epoch 82/200
Epoch 83/200
Epoch 84/200
Epoch 85/200
Epoch 86/200
4/4 [============== ] - Os 84ms/step - loss: 0.2692
Epoch 87/200
4/4 [============== ] - Os 78ms/step - loss: 0.2660
Epoch 88/200
Epoch 89/200
Epoch 90/200
Epoch 91/200
Epoch 92/200
Epoch 93/200
Epoch 94/200
Epoch 95/200
Epoch 96/200
Epoch 97/200
4/4 [============== ] - Os 71ms/step - loss: 0.2070
Epoch 98/200
Epoch 99/200
Epoch 100/200
Epoch 101/200
```

```
Epoch 102/200
Epoch 103/200
Epoch 104/200
Epoch 105/200
4/4 [============== ] - 0s 76ms/step - loss: 0.1827
Epoch 106/200
Epoch 107/200
Epoch 108/200
4/4 [============== ] - Os 75ms/step - loss: 0.1700
Epoch 109/200
Epoch 110/200
Epoch 111/200
Epoch 112/200
Epoch 113/200
Epoch 114/200
Epoch 115/200
Epoch 116/200
Epoch 117/200
Epoch 118/200
Epoch 119/200
Epoch 120/200
4/4 [============== ] - Os 87ms/step - loss: 0.1414
Epoch 121/200
4/4 [============== ] - Os 84ms/step - loss: 0.1400
Epoch 122/200
Epoch 123/200
Epoch 124/200
Epoch 125/200
```

```
4/4 [================ ] - Os 100ms/step - loss: 0.1347
Epoch 126/200
Epoch 127/200
Epoch 128/200
Epoch 129/200
Epoch 130/200
Epoch 131/200
Epoch 132/200
Epoch 133/200
Epoch 134/200
4/4 [============== ] - Os 71ms/step - loss: 0.1234
Epoch 135/200
Epoch 136/200
Epoch 137/200
Epoch 138/200
Epoch 139/200
Epoch 140/200
Epoch 141/200
Epoch 142/200
Epoch 143/200
Epoch 144/200
Epoch 145/200
4/4 [============== ] - Os 84ms/step - loss: 0.1091
Epoch 146/200
4/4 [================ ] - Os 116ms/step - loss: 0.1077
Epoch 147/200
Epoch 148/200
Epoch 149/200
```

```
Epoch 150/200
Epoch 151/200
Epoch 152/200
Epoch 153/200
4/4 [============== ] - 0s 86ms/step - loss: 0.1084
Epoch 154/200
Epoch 155/200
Epoch 156/200
Epoch 157/200
Epoch 158/200
4/4 [============== ] - Os 83ms/step - loss: 0.0998
Epoch 159/200
4/4 [============== ] - Os 83ms/step - loss: 0.1020
Epoch 160/200
Epoch 161/200
Epoch 162/200
4/4 [=============== ] - Os 101ms/step - loss: 0.0980
Epoch 163/200
Epoch 164/200
Epoch 165/200
Epoch 166/200
Epoch 167/200
Epoch 168/200
4/4 [============== ] - Os 77ms/step - loss: 0.0944
Epoch 169/200
Epoch 170/200
4/4 [================ ] - Os 116ms/step - loss: 0.0945
Epoch 171/200
Epoch 172/200
Epoch 173/200
```

```
Epoch 174/200
Epoch 175/200
Epoch 176/200
Epoch 177/200
4/4 [============== ] - 0s 76ms/step - loss: 0.0899
Epoch 178/200
Epoch 179/200
Epoch 180/200
4/4 [============== ] - Os 73ms/step - loss: 0.0871
Epoch 181/200
Epoch 182/200
Epoch 183/200
4/4 [============== ] - Os 73ms/step - loss: 0.0816
Epoch 184/200
Epoch 185/200
Epoch 186/200
Epoch 187/200
Epoch 188/200
Epoch 189/200
Epoch 190/200
Epoch 191/200
Epoch 192/200
Epoch 193/200
4/4 [============== ] - Os 80ms/step - loss: 0.0834
Epoch 194/200
Epoch 195/200
Epoch 196/200
Epoch 197/200
```



```
Layer (type)
      Output Shape
______
embedding_2 (Embedding) (1, None, 256)
                       21760
______
cu_dnngru_2 (CuDNNGRU) (1, None, 1024)
                    3938304
_____
dense_2 (Dense) (1, None, 85) 87125
______
Total params: 4,047,189
Trainable params: 4,047,189
Non-trainable params: 0
_____
None
```

```
[21]: def generate_text(model, start_string):
    # Evaluation step (generating text using the learned model)

# Number of characters to generate
    num_generate = 1000

# Converting our start string to numbers (vectorizing)
    input_eval = [char2idx[s] for s in start_string]
    input_eval = tf.expand_dims(input_eval, 0)

# Empty string to store our results
    text_generated = []

# Low temperatures results in more predictable text.
```

```
# Higher temperatures results in more surprising text.
         # Experiment to find the best setting.
         temperature = 1.0
         # Here batch size == 1
         model.reset_states()
         for i in range(num_generate):
             predictions = model(input_eval)
             # remove the batch dimension
             predictions = tf.squeeze(predictions, 0)
             # using a multinomial distribution to predict the word returned by the
      \rightarrowmodel
             predictions = predictions / temperature
             predicted_id = tf.multinomial(predictions, num_samples=1)[-1,0].numpy()
             # We pass the predicted word as the next input to the model
             # along with the previous hidden state
             input_eval = tf.expand_dims([predicted_id], 0)
             text_generated.append(idx2char[predicted_id])
         return (start_string + ''.join(text_generated))
[22]: print(generate_text(model, start_string=u"struct"))
    struct audit_tree *tree)
    {
            put_tree *tree)
    {
            return tree->pathname, 0, &path);
            if (IS_ERR(mark))
                     return;
            out_mutex:
            mutex_unlock(&audit_tree_group->mark_mutex);
                     return -ENOMEM;
            }
            mark = ale;
                     fsnotify_free_mark(mark);
                     fsnotify_put_mark(mark);
                     kfree(chunk);
                     return 0;
            }
            replace_mark_chunk(old->mark, new);
            /*
                     struct audit_chunk *chunk)
    {
```

```
struct audit_chunk *old;
        assert_spin_locted->chunks, list) {
                        list_move(&owner->list, &prune_list);
                        need_prune = 1;
                } else {
                struct path path1, path2;
        struct vfsmount *tagged;
        int err;
        err = kern_path(new, 0, &path2);
                if (!err) {
                        good_one = canlic inline struct audit_tree_mark
*audit_mark(struct fsnotify_mark *mark,
                        goto skip_it;
                root_mnt = collect_mounts(&path);
                path_put(&path);
                if (IS_ERR(rootagner) /* reorder */
        for (p = tree->chunks.next; p != &tree->chunks; p = q) {
                struct node *node = list_entry(prune_list.next,
                                        struct audit_chunk *chunk =
from_chunk(mark) != chunk)
                goto o
```