lab_3

April 2, 2021

1 RNN

1.1 Do a bit of research on similar problems such as named entity recognition, find a dataset, train a model, and report your results.

[1]:
```
#load the packages
from __future__ import unicode_literals, print_function, division
from io import open
import os, string, random, time, math
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.model_selection import train_test_split

import torch
import torch.nn as nn
import torch.optim as optim

from io import open
import glob
import os
import unicodedata
import string
import pandas as pd
device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")

all_letters = string.ascii_letters + " .,;'"
n_letters = len(all_letters)
```

2 load data

[2]:
```
all_letters = string.ascii_letters + " .,;'"
n_letters = len(all_letters)
```

```
df = pd.read_csv(r'/root/labs/RTML/lab10/lecture/data/new/ner_dataset.csv',␣
 ↪encoding= 'unicode_escape')
df = df.fillna(method="ffill")
df = df.drop(['Tag'], axis =1)
df = df.drop(['Sentence #'], axis =1)
```

[3]:
```
print(df)
```

```
                 Word  POS
0           Thousands  NNS
1                  of   IN
2       demonstrators  NNS
3                have  VBP
4             marched  VBN
…                 …  …
1048570           they  PRP
1048571      responded  VBD
1048572             to   TO
1048573            the   DT
1048574         attack   NN

[1048575 rows x 2 columns]
```

[4]:
```
cat_words = {}
X = []
y = []
tags = list(set(df["POS"].values))

for i in tags:
    cat_words[i] = []

for i in range(len(df)):
    cat_words[df['POS'].iloc[i]].append(df['Word'].iloc[i])
    y.append(df['POS'].iloc[i])
    X.append(df['Word'].iloc[i])
```

3 Train test split

[5]:
```
#split the data

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,␣
 ↪random_state = 123)
```

[6]:
```
print("The number of observations in the training data: ", len(X_train))
print("The number of observations in the test data: ", len(X_test))
```

```
The number of observations in the training data:  838860
The number of observations in the test data:  209715
```

[7]:
```python
#create a batched name rep

def batched_name_rep(names, max_word_size):
    rep = torch.zeros(max_word_size, len(names), n_letters)
    for name_index, name in enumerate(names):
        for letter_index, letter in enumerate(name):
            pos = all_letters.find(letter)
            rep[letter_index][name_index][pos] = 1
    return rep

#function to print the output
def print_char(name_reps):
    name_reps = name_reps.view((-1, name_reps.size()[-1]))
    for t in name_reps:
        if torch.sum(t) == 0:
            print('')
        else:
            index = t.argmax()
            print(all_letters[index])

def batched_lang_rep(langs):
    rep = torch.zeros([len(langs)], dtype=torch.long)
    for index, lang in enumerate(langs):
        rep[index] = tags.index(lang)
    return rep

#create dataloader
def batched_dataloader(npoints, X_, y_, verbose=False, device = 'cpu'):
    names = []
    langs = []
    X_lengths = []

    for i in range(npoints):
        index_ = np.random.randint(len(X_))
        name, lang = X_[index_], y_[index_]
        X_lengths.append(len(name))
        names.append(name)
        langs.append(lang)
    max_length = max(X_lengths)

    names_rep = batched_name_rep(names, max_length).to(device)
    langs_rep = batched_lang_rep(langs).to(device)
```

```
    padded_names_rep = torch.nn.utils.rnn.pack_padded_sequence(names_rep,␣
↪X_lengths, enforce_sorted = False)

    if verbose:
        print(names_rep.shape, padded_names_rep.data.shape)
        print('--')

    if verbose:
        print(names)
        print_char(names_rep)
        print('--')

    if verbose:
        print_char(padded_names_rep.data)
        print('Lang Rep', langs_rep.data)
        print('Batch sizes', padded_names_rep.batch_sizes)


    return padded_names_rep.to(device), langs_rep
```

[8]:
```
n_points = 100
batch_input, batch_groundtruth = batched_dataloader(n_points, X_train, y_train,␣
↪False)
```

4 RNN Model

[9]:
```
#create simple rnn network
class RNN(nn.Module):
    #Create a constructor
    def __init__(self, input_size, hidden_size, output_size):
        super(RNN, self).__init__()
        self.hidden_size = hidden_size
        self.rnn_cell = nn.RNN(input_size, hidden_size)
        self.h20 = nn.Linear(hidden_size, output_size)
        self.softmax = nn.LogSoftmax(dim = 1)

    #create a forward pass function
    def forward(self, input_, hidden = None, batch_size = 1):
        out, hidden = self.rnn_cell(input_, hidden)
        output = self.h20(hidden.view(-1, self.hidden_size))
        output = self.softmax(output)
        return output, hidden

    def init_hidden(self, batch_size = 1):
        #function to init the hidden layers
```

```
        return torch.zeros(1, batch_size, self.hidden_size)
```

5 Dataloader

[10]:
```
def dataloader(npoints, X_, y_):
    """Function to load the data"""
    to_ret = []
    for i in range(npoints):
        index_ = np.random.randint(len(X_))
        name, lang = X_[index_], y_[index_] #subset the data
        to_ret.append((name, lang, name_rep(name), lang_rep(lang)))

    return to_ret
#function to create representation of the name
def name_rep(name):
    rep = torch.zeros(len(name), 1, n_letters) #Create a zeros tensor
    #iterate through all the characters in the name
    for index, letter in enumerate(name):
        pos = all_letters.find(letter)
        rep[index][0][pos] = 1 #Assign a value for each pos value
    return rep
#function to create vec representation of the language
def lang_rep(lang):
    return torch.tensor([tags.index(lang)], dtype = torch.long)
```

6 Evaluation

[18]:
```
#create an evaluation function

def eval(net, n_points, topk, X_, y_, device = device):
    "Evaluation function"

    net = net.eval().to(device)
    data_ = dataloader(n_points, X_, y_)
    correct = 0

    #iterate
    for name, language, name_ohe, lang_rep in data_:
        name_ohe = name_ohe.to(device)
        lang_rep = lang_rep.to(device)
        #get the output
        output = infer(net, name, device)
        val, indices = output.topk(topk) #get the top k values
```

```
        indices = indices.to(device) #convert to devices

        if lang_rep in indices:
            correct += 1

    accuracy = correct/n_points
    return accuracy
```

7 Training

[19]:
```
#basic train function

def train(net, opt, criterion, n_points):

    opt.zero_grad()
    total_loss = 0

    data_ = dataloader(n_points, X_train, y_train)

    total_loss = 0

    for name, language, name_ohe, lang_rep in data_:

        hidden = net.init_hidden()

        for i in range(name_ohe.size()[0]):
            output, hidden = net(name_ohe[i:i+1], hidden)
        loss = criterion(output, lang_rep)
        loss.backward(retain_graph=True)

        total_loss += loss

    opt.step()
    return total_loss/n_points


def train_batch(net, opt, criterion, n_points, device):
    net.train().to(device)
    opt.zero_grad()

    batch_input, batch_groundtruth = batched_dataloader(n_points, X, y, False,␣
 ↪device)
    batch_input = batch_input.to(device)
    batch_groundtruth = batch_groundtruth.to(device)
```

```
    output, hidden = net(batch_input)
    loss = criterion(output, batch_groundtruth)
    loss.backward()
    opt.step()

    return loss
```

[20]:
```
from tqdm.contrib.discord import tqdm, trange
def train_setup(net, lr = 0.01, n_batches = 100, batch_size = 10, momentum = 0.
↪9, display_freq=5, device='cpu'):
    net = net.to(device)
    criterion = nn.NLLLoss()
    opt = optim.SGD(net.parameters(), lr=lr, momentum=momentum)

    loss_arr = np.zeros(n_batches + 1)

    for i in tqdm(range(n_batches), token='ODI3MTkyMDYyNzU0ODE2MDIy.YGXcpA.
↪py4WQmZsv0HPu5eh2se117TNRzk', channel_id='827196085738536971'):
        loss_arr[i+1] = (loss_arr[i]*i + train_batch(net, opt, criterion,␣
↪batch_size, device))/(i + 1)

        if i%display_freq == display_freq-1:
            clear_output(wait=True)

            print('Iteration', i, 'Loss', loss_arr[i])
            # print('Top-1:', eval(net, len(X_test), 1, X_test, y_test), 'Top-2:
↪', eval(net, len(X_test), 2, X_test, y_test))
            plt.figure()
            plt.plot(loss_arr[1:i], '-*')
            plt.xlabel('Iteration')
            plt.ylabel('Loss')
            plt.show()
            print('\n\n')

    print('Top-1 Accuracy:', eval(net, len(X_test), 1, X_test, y_test, device),␣
↪'Top-2 Accuracy:', eval(net, len(X_test), 2, X_test, y_test, device))
```

[21]:
```
n_hidden = 128
#device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
n_categories = len(y)

#number of tags
n_tags= len(cat_words.keys())
print("Total number of tags present: ", n_tags)

net = RNN(n_letters, n_hidden, n_tags)
criterion = nn.NLLLoss()
```

```
opt = optim.SGD(net.parameters(), lr=0.01, momentum=0.9)
```

```
Total number of tags present:  42
```

7.1 Inference

[22]:
```
def infer(net, name, device = "cpu"):
    name_ohe = name_rep(name).to(device)

    #get the output
    output, hidden = net(name_ohe)

    if type(hidden) is tuple: #for lSTM
        hidden = hidden[0]
    index = torch.argmax(hidden)

    return output
```

[23]:
```
from IPython.display import clear_output
train_setup(net, lr=0.15, n_batches=300, batch_size = 512,␣
 ↪display_freq=500,device=device)
```

```
  0%|          | 0/300 [00:00<?, ?it/s]
```

```
Top-1 Accuracy: 0.8453997091290562 Top-2 Accuracy: 0.9325370145197053
```

8 Result

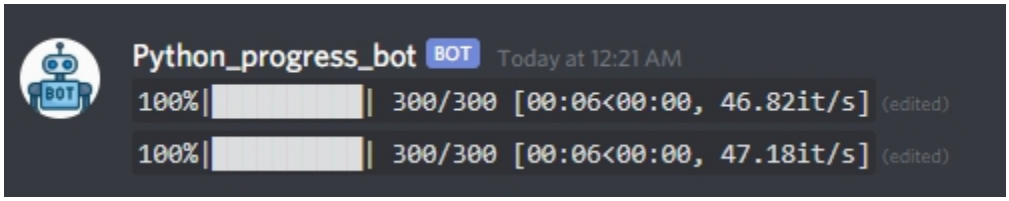
Result with test set showed that top-1 accuracy is 0.84 and top-2 accuracy is 0.93253

8.1 Fancy thing

I create a bot to show a progress of training data:)

[24]:
```
from IPython.display import Image
Image(filename='bot.jpg')
```

[24]:



[ ]: