INTERNSHIP REPORT

Monalisa Maity

Introduction

I completed my training for the **Data Science Internship** by working on a project based on a **Translation App.** During the training, I learnt how to build translation models and train them using tensorflow. I was assigned the internship tasks in which I had to use the skills that I learnt during this training.

Background and Learning Objectives

In the project, 4 models for language translation were built. The first three models were trained to translate English to French and the last model was used for English to Spanish Translation.
The types of models that were used are:

1. Simple RNN
2. Bidirectional RNN
3. Embedding RNN
4. Transformer Model

Tensorflow was the framework that was used to build and train the models and Tkinter was used to make a graphic user interface to take input sentences from the users for translation.

The learning objectives include:

1. Preprocessing the text data i.e., tokenization and padding.
2. Splitting the dataset for training and testing the dataset.
3. Building the model and then training it.
4. Saving the trained model and tokenizer
5. Loading the saved model and tokenizer and then using them for translation
6. Building the UI using Tkinter

Activities and Tasks

1. Load a pre-trained LSTM-based NMT model and use it to translate a sentence from one language to another.
2. Implement beam search decoding for an NMT model to improve translation quality.

3. Create a feature to translate the language with a combination of two languages at the same time . We should be able to convert the 2 different languages at the same time . translate English to French and Hindi at the same time . This model should work only for 10 letter English words . If we enter below 10 letters or above 10 letters it should not work.
4. Create a feature to translate the audio into Hindi . The system will listen the english audio from user and it will convert into Hindi word. If the system does not understand the audio it will ask repeat one more time to make it better.. The audio should be in English word only . This translation feature work on only after 6 PM IST timing and before that it should show message like please try after 6 PM IST as well as it should not translate any english which is start with M and O apart from that it should translate all other words
5. Create a feature to translate the English word to Hindi and it should not translate if the English starts with vowels and other words it should convert . If we enter a English word starts with Vowels it should show an error message as This word starts with Vowels provide some other words and this model should be able to convert english word starts with vowels around 9 PM to 10 PM

Skills and Competencies

By working on the tasks assigned to me, I gained the following new skills:

1. Loading a pre-trained model and using it for translation
2. Loading a HuggingFace transformer model and tokenizer for translation
3. Taking audio input through microphone in python
4. Running code in python according to a given time
5. Using Google speech recognizer and translator in Python applications

Feedback and Evidence

I had a positive experience during the training phase as well as while working on the internship tasks. I learnt many important skills in the field of data science which will help me in achieving my future goals and aspirations in the field of Machine Learning and Data Science.

I have attached all the code files for the tasks assigned to me. I used the following resources for completing my internship tasks:

1. Kaggle
2. Stackoverflow
3. towardsdatascience.com
4. medium.com
5. HuggingFace Transformers library

6. pypi.org
7. Tensorflow Documentation

Challenges and Solutions

Task-1:

For the first task, I had to use a pre-trained LSTM based model for translation. However, most of the pre-trained open source models available were not LSTM based. There was only one LSTM based model: German-English - 2-layer BiLSTM available on OpenNMT but there was no proper documentation to use this model. Therefore I searched for pre-trained models on kaggle but the tokenizers were not available for them. Since I was unable to find a pre-trained LSTM based model online, I trained a simple LSTM based model and saved its tokenizer as well. I loaded this model and tokenizer to translate French to English.

Architecture of the pre-trained model that I used:

```
def LSTM_model(in_vocab, out_vocab, in_timesteps, out_timesteps, units):
    model = Sequential()
    model.add(Embedding(in_vocab, units, input_length=in_timesteps, mask_zero=True))
    model.add(LSTM(units))
    model.add(RepeatVector(out_timesteps))
    model.add(LSTM(units, return_sequences=True))
    model.add(Dense(out_vocab, activation='softmax'))
    return model
```

Task-2:

I was not familiar with the concept of Beam search decoding so I went through some articles, codes and stackoverflow to understand the concept and get some help in the implementation.

I implemented a basic beam search decoder with beam width 3 and tried using it for one of the English to French translation models from the training project. It was able to give the 3 best candidates for the first token however, I was not able to use it to decode the entire sentence.

TASK-3, TASK-4, TASK-5 :

I trained the embedding RNN model with a new English-Hindi dataset for English to Hindi translation.The training was successful but the model was not able to translate the words properly. I was unable to identify the error so I wrote the code with this trained model for tasks - 3, 4 and 5 along with alternative codes using open-source models from Huggingface and google trans for performing the same tasks.
I have added both codes for each task as well as the code for training the English-Hindi Translation model.

Task-3:

I used **Helinksi Opus**, an open source model on huggingface for translating English to Hindi and French. I created a function to take word, model and tokenizer as input and perform translation and then return the output. The second function calls the previous function twice by passing the two models and the word to be translated and prints the translated words. If the word to be translated exceeds 10 letters, it doesn't translate the word.

Task-4:

I used googletrans and speech_recognition libraries in python for translating and recognizing the audio respectively. I created a function to translate the word by checking if it starts with 'M', 'm', 'O' or 'o'. I created another function to get the IST and current time using the pytz and datetime library.

Task-5:

I used googletrans library for translation, pytz and datetime libraries for getting the current time according to the IST. I created a function to translate the word to hindi while checking if it starts with a vowel or not.

Outcomes and Impact

The internship tasks helped me in learning new skills as well as testing the skills that I learned from the training program. These tasks provided me with the knowledge of training and building my own translation models as well as using the understanding and working of different types of open-source translation models available and how to use them in my own programs.

Conclusion

Overall this internship helped me learn a lot more about translation models and improved data science and programming skills. I have submitted my work for all the tasks that I was assigned.