دانشگاه شهید بهشتی
۱۳۳۸

Title:

Language Modeling for Persian Tweets

Department of Computer and Data Sciences

Advisor:
Dr. Kheradpisheh

Student:
Fatemeh Saberi Khomami

Student Number:
400422114

Table of Contents

1. Introduction

Language modeling is a gateway into many deep learning applications, like speech recognition, image captioning and machine translation. Language modeling is the process of assigning probabilities to sequences of words. So, language modeling could analyze a sequence of words and predict which word is most probable to come next.

In this project, we are going to build such a system based on Persian tweets to predict the next word given a token.

2. Getting to know the Data

We have 249,981 tweets with different length in the dataset. At the first step, we cleaned the data by replacing redundant characters like half-spaces, next-lines with a space.

Then we omitted the missing records from the dataset. In the below figure, you can see the distribution of words in tweets.

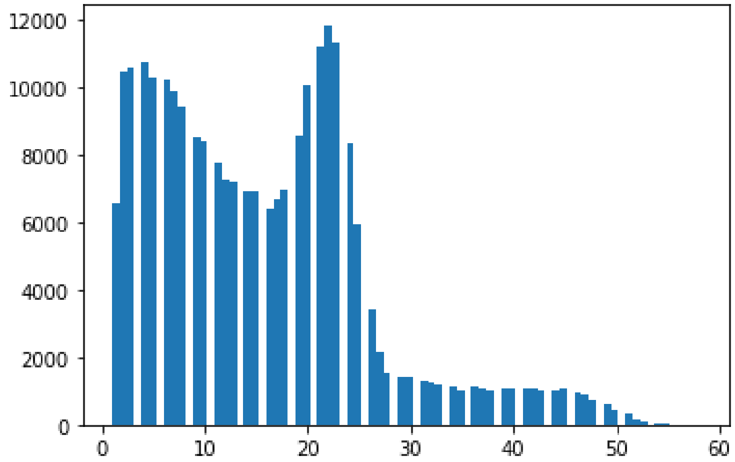

Figure (1): Distribution of number of words in tweets

3. Data Cleaning

To prepare our texts for the network, we must tokenize them so that each sentence has a numeric representation. We decided to continue with only the first 1,000 longest tweets due to the computation limitations. In our dictionary, we used <oov> to represent the unknown words.

Total number of words in our dictionary is 15,477. We turned our tokenized tweets into n-gram sequences to include all the subsequences. For example, the tweet: "I love my cat" will turn into "I love my cat", "I love my", and "I love". To make our tokenized representations even, we used zero padding to make the size of each sentence-representation equaled to the longest sentence.

At the last step, we took the last element of all sequences as the label and left the rest for the training data.

We are ready to start training!

4. Begin Training

The first model is using a Bidirectional LSTM with 32 units, optimized by Adam algorithm with learning algorithm 0.01. The training was 40 epochs long. This architecture has 1,517,909 parameters, and here is the result of the last epoch: loss: 1.7223 - accuracy: 0.6313

You can see the performance of the model in the following figure.

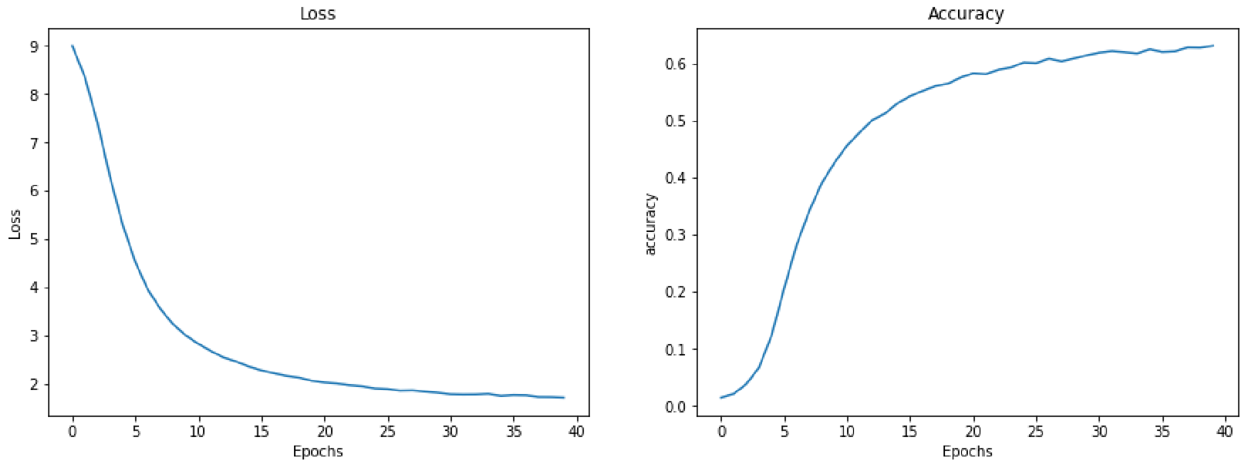

Figure (2): Loss and accuracy of the first model

When we gave the token "پس" to the model, it predicted:

" پس منتظرچه هستیدوقتی"

Let's proceed to the second model.

Here, we used a Bidirectional GRU with 30 units, optimized by Adam algorithm with learning algorithm 0.01. The training was 40 epochs long.

This architecture has 2,515,557 parameters, and here is the result of the last epoch:

loss: 1.1266 - accuracy: 0.7260

You can see the performance of the model in the following figure.

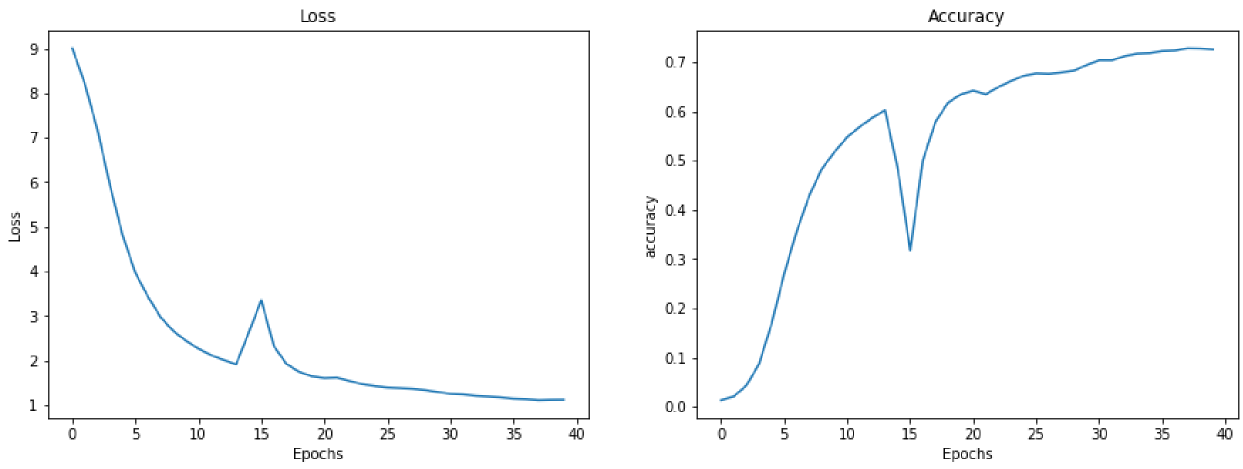

Figure (3): Loss and accuracy of the second model

When we gave the token "شاید" to the model, it predicted:

" شاید دیده که "

Let's proceed to the third model.

We used the same architecture and configuration as before, but increased the number of epochs to 50. The last epoch had the following results:

loss: 1.0603 - accuracy: 0.7441

You can see the performance of the model in the following figure.

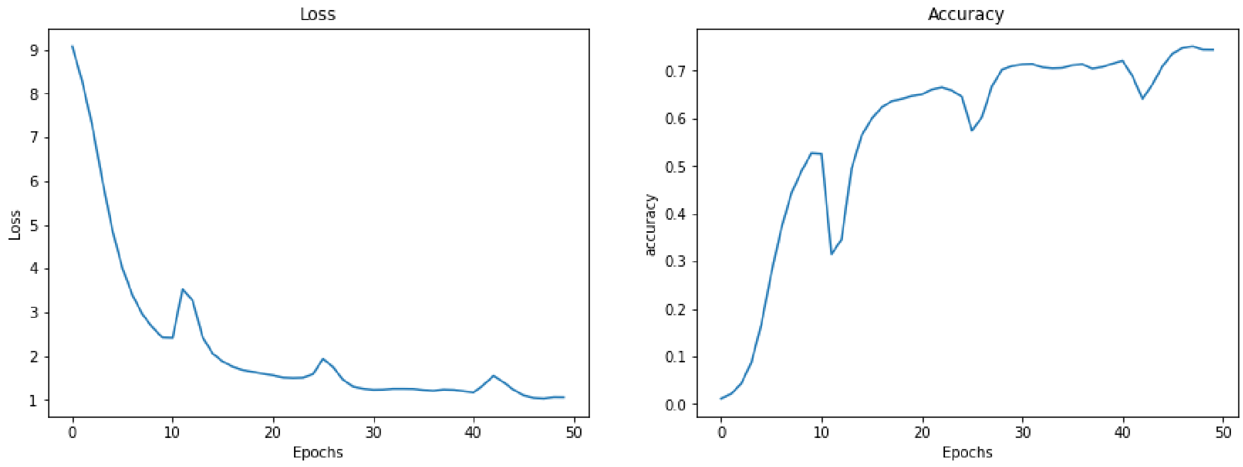

Figure (4): Loss and accuracy of the third model

When we gave the token "روزی" to the model, it predicted:

" روزی فرا میرسدکه "

Let's proceed to the last model.

We used a Bidirectional LSTM with 100 units, optimized by Adam algorithm with learning algorithm 0.01. The training was 40 epochs long. This architecture has 2,523,237 parameters, and here is the result of the last epoch:

loss: 0.9671 - accuracy: 0.7892

You can see the performance of the model in the following figure.

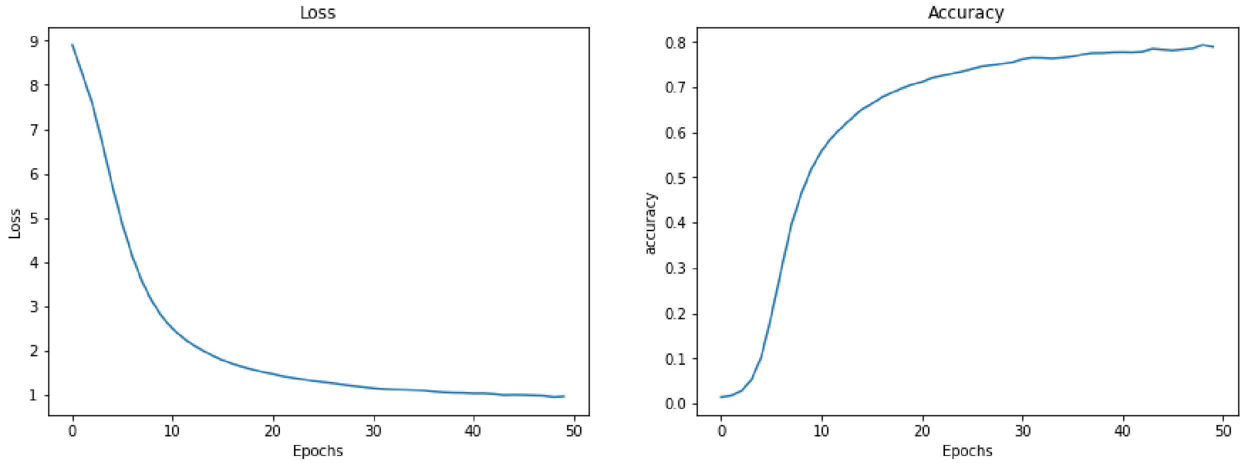

Figure (5): Loss and accuracy of the fourth model

When we gave the token "کاش" to the model, it predicted:

" کاش شما میدانستید "