Abstract

The two aims of this thesis are introducing the concepts of speech recognition and deep learning (focusing on recurrent neural networks and long short-term memory neural networks) and then building a Romanian speech-to-text application powered by a long short-term memory neural network, which was to be trained and tested on the SWARA Speech Corpus, a dataset containing around 20 hours of recorded Romanian speech.

The thesis is divided into three chapters. The first one defines the problem of speech recognition and its solutions throughout time. The term *end-to-end speech recognition system* is also introduced, together with a useful mechanism and scoring function: *connectionist temporal classification* (CTC). The second chapter introduces deep learning concepts such as: artificial neural networks, supervised learning, testing, and training. From there it builds up to more specific architectures: recurrent neural networks (RNNs) and long short-term memory neural networks (LSTMs). Finally, the last chapter focuses on the end-to-end speech-to-text application I have developed, tackling both the minimal user interface and the underlying model generating the transcripts, along with the frameworks I have used: PyQt5 and Keras, respectively.

Although the obtained model is not accurate enough to be usable in a real-life situation, the process lead me to some valuable observations about how the performance of the model changes with the change of certain parameters and hyperparameters. The most performant model I have managed to obtain consists of 5 LSTM layers, with 33 neurons each and a BatchNormalization and a TimeDistributed layer. It achieved a CTC loss of 48.

With regards to future work, one useful addition to the deep learning model would be a language model. In addition to that, when larger Romanian speech corpora will appear, it would be interesting to train the model obtained in this application and see how much better it performs when dealing with bigger datasets. Another suggestion for the future is to replace the LSTM units with gated recurrent units (GRUs) −another recurrent neural network architecture.

This work is the result of my own activity. I have neither given nor received unauthorized assistance on this work.