Project Report:

Speech Based Sentiment Detection

Authors: Aniket Yadav

Table of Contents

1. Introduction

Project Overview

This project report documents our exploration of sentiment analysis on the HarperValleyBank (HVB) Spoken Dialog Corpus. The Project focuses on training a sentiment detection model using a pre-existing CRDNN (Convolutional Recurrent Deep Neural Network) model pretrained on the LibriSpeech dataset. We aim to adapt this pretrained model for sentiment analysis on the HVB dataset.

Dataset Description

The HVB corpus is a publicly-available spoken dialog corpus containing transcriptions, audio segments, intent annotations, sentiment/emotion labels, and dialog actions. The dataset includes conversations between agents and callers, and each utterance is associated with sentiment probabilities (positive, neutral, negative).

Project Goals

Our main objectives for this project were:

- Preprocess the dataset to create train and dev splits, and extract relevant metadata and audio segments.
- Adapt a pretrained CRDNN model for sentiment analysis by fine-tuning and adding a linear layer for sentiment classification.
- Train and evaluate the model's performance on sentiment prediction, comparing it against a baseline.
- Conduct error analysis to identify common failure modes and gain insights into model performance.
- Fine-tune the model on the HVB data and compare its performance with the pretrained model.

2. Acknowledgements

Sources and Credits

We acknowledge the following sources and credits for this project:

- The SpeechBrain toolkit for providing pretrained ASR models and utilities.
- Stanford CS224S Assignment 4 for code snippets and ASR-related functionality.
- The SpeechBrain tutorial for guidance on building an ASR system using the SpeechBrain toolkit.

Tools and Libraries

- Python programming language
- PyTorch and SpeechBrain libraries
- Jupyter Notebook environment for code development
- Google Colab for training and experimentation

3. Data Preprocessing

Obtaining and Organizing Data

We obtained the HVB dataset, which includes transcripts and audio segments, along with annotations for intent, sentiment/emotion, and dialog actions. The dataset was organized into transcript and audio folders.

Transcript and Audio Processing

We processed the transcripts to extract relevant metadata such as sentiment labels, speaker roles, start times, and durations. The audio segments were extracted based on the start times and durations from the transcripts.

Train-Dev Split

We created train and dev splits using train.txt and dev.txt files. These splits were used to create corresponding datasets for training and evaluation.

4. Model Architecture and Training

Loading Pretrained CRDNN Model

We loaded a pretrained CRDNN model from the SpeechBrain toolkit, which was originally trained on the LibriSpeech dataset. We extracted the encoder layers from the model for our sentiment analysis task.

Adding Linear Layer for Sentiment Classification

We added a linear layer on top of the pretrained encoder layers to perform sentiment classification. The linear layer was initialized with Xavier uniform weights and trained alongside the encoder layers.

Training Procedure

We used a cross-entropy loss function and the Adam optimizer for training. The model was trained on the training dataset, and we monitored its performance on the dev dataset.

5. Sentiment Analysis Results

Word Error Rate (WER)

The Word Error Rate (WER) for the ASR model is calculated as 75.84%. The following are some notable error examples:

> Example at Index 21066: Words with similar pronunciation were wrongly transcribed.
> Example at Index 8211: No penalty on word insertion resulted in an excessive number of words in the transcription.
> Example at Index 11540: Noisy audio led to transcription errors.
> Example at Index 17273: The transcript was not available, resulting in an error.

Example at Index 5729: Homophones were wrongly transcribed in the absence of a language model.

Training Progress and Loss

The training progress for ASR is as follows:

- Epoch 1: Train Loss - 1.78, Valid Loss - 1.67, Valid Character Error Rate (CER) - 1.04e+02, Valid WER - 99.98 (Learning Rate: 1.00e+00)
- Epoch 2: Train Loss - 1.43, Valid Loss - 1.47, Valid CER - 42.27, Valid WER - 41.58 (Learning Rate: 1.00e+00)

Comparison of ASR Models

The ASR models' performance after fine-tuning and pre-training is as follows:

- Fine-tuned Model: Word Error Rate (WER) - 33.08
- Pre-trained Model: Word Error Rate (WER) - 71.07

Sentiment Analysis Results

- Train Accuracy: 77.46%
- Dev Accuracy: 70.37%

These results provide insights into the performance of the ASR and Sentiment Analysis models, demonstrating their accuracy and areas of improvement.

6. Error Analysis

Investigating Common Error Modes

We conducted error analysis by examining examples where the model's predictions were incorrect. Common error modes were identified, such as misinterpretation of emotional cues and challenges posed by overlapping speech.

Identifying Challenging Cases

Challenging cases were identified based on the model's poor performance. These included cases with unclear sentiment cues, low audio quality, and ambiguous context.

Hypotheses for Model Failures

Based on our error analysis, we hypothesized that the model's failures were partly due to the differences between the LibriSpeech and HVB datasets, particularly in terms of speaker characteristics and emotional expression.

7. Fine-Tuning and Comparison

Fine-Tuning on HVB Data

We fine-tuned the sentiment analysis model using the HVB training data to adapt it to the specific domain and characteristics of the HVB dataset.

Evaluation on Dev Set

The fine-tuned model was evaluated on the dev dataset, and its performance was compared against the initial pretrained model.

Comparison with Pretrained Model

We compared the sentiment analysis performance of the fine-tuned model with the initial pretrained CRDNN model to determine the effectiveness of domain adaptation.

8. Conclusion

Summary of Findings

In this project, we successfully adapted a pretrained CRDNN model for sentiment analysis on the HarperValleyBank Spoken Dialog Corpus. We explored the challenges posed by the dataset's unique characteristics, such as emotional variability and overlapping speech.

Lessons Learned

Through error analysis and fine-tuning, we gained insights into the limitations and potential improvements of the model. We learned the importance of domain adaptation and the value of analyzing failure modes.

Future Work

Future work could involve further fine-tuning the model, exploring advanced architectures, and incorporating additional features (e.g., prosody) to improve sentiment analysis performance. Additionally, investigating multi-modal approaches could enhance the model's ability to capture emotional cues.