Machine learning project: Speech recognition challenge

Massimiliano Conte
massimiliano.conte.2@studenti.unipd.it

Pierpaolo D'Odorico
pierpaolo.dodorico@studenti.unipd.it

1. Introduction

The faced problem is the speech recognition challenge, where the goal is to build a system that can automatically recognize spoken words. In particular, the spoken words are given as input in form of audio recordings of one second length, while the output of the system is the text of the recognized words. All the recordings are waveform of one of the following 8 words:

- Down;
- Go;
- Left;
- Off;
- On;
- Right;
- Stop;
- Up.

This task is a key component in many artificial intelligence services, such as virtual assistants, and more generally speech recognition is part of the natural language processing domain.
Our work begin after the features extraction provided us by the professors, in form of log mel-spectrogram, that is a bidimensional representation of the recordings, involving frequency and time. After that, the images of the spectrograms are first resized and than reshaped in 1024-dimensional vectors.
We have faced the problem by applying several machine learning methods, from the easier ones to the more complicated ones, and choosing the best performing method based on accuracy. This led us to a system capable of classifying spoken words with an accuracy of more that 95%.

2. Dataset

The used dataset is a reduced version of TensorFlow Speech Commands Dataset v0.0.1 [2]. The data provided us is alredy divided in training and validation set. The training set is composed of 1600 samples, 200 for each of the 8 classes, and the validation set is composed of 109 samples. Every recording has a sampling frequency of 16 kHz. The hop length used for constructing the log Mel-spectrogram is 512, while it is not specified the window size. The feature extraction mimics the humany auditory system, providing a representation of the audio taking into account the fact that humans perceve both the frequencies and the amplitude of the sound logarithmically. The whole preocess of extracting the features for each recording can be summarized in:

- **Create the windows**, by sampling the recording and making hops of size 512;
- **Compute the discrete Fuorier transform** for each window, using the fast Fourier transform algorithm;
- **Convert to the Mel scale**: changing the representation from of the frequencies from Hz to the Mel scale (the scale of pitches judged by listeners to be equal in distance one from another, a kind of human perceving frequency scale);
- **Create the features**, by computing the log Mel-spectrogram (that is an image), resizing and reshaping it to a 1024 dimensional vector.

The dataset provided us alredy contains the extracted features and the correct class of the samples. The dataset doesn't have unbalanced classes problem.

3. Method

The methods used in this project are machine learning tecniques that are well suited for multiclass classification tasks. We tried the performance of various methods under different configurations, starting from the basic models and ending with the more complicated ones. Many models we tried are based on binary classification, the way we used such models are using the one-vs-all strategy, i. e. training one binary classifier per class and than predict the instances by loocking at the classifier that maximizes the confidence score.

3.1. List of used methods

Every method can be tuned by changing some hyperparameters. In the following table we reported the tecniques, which of those hyperparameters we taked into account for selecting the best configuration, and how that method handle multiclass classification. The tecniques we tried are:

Method	Hyperparameters	handling
Linear Classification	C : Regularization	One-vs-all
Logistic Regression	C : Regularization	One-vs-all
K-Neighbors Classifier	K : number of neighbors	Majority vote of the K-neighbors
Classification tree	Max depth; Min samples leaf; Min impurity decrease	Naturally handle multiclass
Random forest	Number of trees	Naturally handle multiclass
Support Vector Machine	C : Regularization; Type of kernel	One-vs-all
Neural network		Softmax activation on the output layer

Table 1. Results. Ours is better.

- Introduction (20%): describe the problem you are working on, why it's important, what are your goals, and provide also an overview of your main results.
- Dataset (20%): describe the data you are working with for your project. What type of data is it? Where did it come from? How much data are you working with? Did you have to do any preprocessing, filtering, etc., and why?
- Method (30%): discuss your approach for solving the problems that you set up in the introduction. Why is your approach the right thing to do? Did you consider alternative approaches? It may be helpful to include figures, diagrams, or tables to describe your method or compare it with others.
- Experiments (30%): discuss the experiments that you performed. The exact experiments will vary depending on the project, but you might compare with prior work, perform an ablation study to determine the impact of various components of your system, experiment with different hyperparameters or architectural choices. You should include graphs, tables, or other figures to illustrate your experimental results.

4. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

4.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high. Page numbers should be in footer with page numbers, centered and .75 inches from the bottom of the page and make it start at the correct page number rather than the 4321 in the example. To do this fine the line (around line 23)

```
%\ifcvprfinal\pagestyle{empty}\fi
\setcounter{page}{4321}
```

where the number 4321 is your assigned starting page.

Make sure the first page is numbered by commenting out the first page being empty on line 46

```
%\thispagestyle{empty}
```

4.2. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Table 2. Short captions should be centred.
Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 2. Results. Ours is better.

4.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

4.4. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books.

4.5. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

References

[1] Authors. *The frobnicatable foo filter*. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`. 2014.

[2] Martın Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `https://www.tensorflow.org/`.

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.