CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation

Théo Deschamps-Berger

I. Introduction

With the advancement of technology our understanding of emotions are advancing, there is a growing need for automatic emotion recognition systems. This paper aim to recognize emotions in speech, using Convolutional Neural Network for extracting high-level features from raw spectrograms and recurrent ones for aggregating long-term dependencies on the IEMOCAP dataset [2].

II. Corpus

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted, multimodal and multispeaker database. The corpus contains 12 hours of improvisations and scenarios performed by 10 professional actors (five women and five men) along 5 sessions of dialogues between two actors of different genders. Each sample of the audio set is an utterance assigned with an emotion label. An utterance means a spoken word. The dataset includes video, speech, motion capture of face and text transcriptions. IEMOCAP database is annotated by six students of USC into categorical labels, such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance.

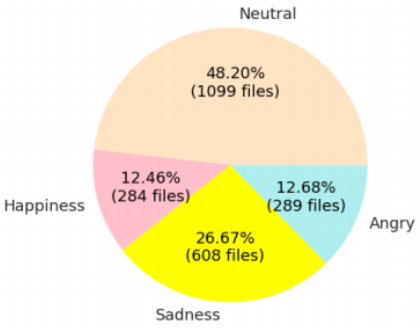

Fig. 1: Class distribution of the utterances in the improvised part of the IEMOCAP dataset

III. Data augmentation

One of the main difficulties encountered with the IEMOCAP is class imbalance. A dataset is unbalanced if the classification categories, or classes, are not represented approximately equally. The difficulty lies in the fact that the model only learns about the majority class if we do not adapt the distribution of the number of files per class, or even adapt the cost function, resulting in a learning bias. And another drawback is that IEMOCAP dataset is too small. The data augmentation is used to alleviate the overfitting and oversampling the less represented classes to obtain a balanced training corpus.
The authors chose to use the VTLP method (Vocal Track Length Peturbartion), which consists of rescaling the original spectrograms along the frequency axis with a scaling factor α taking values between 0.9 and 1.1, as follows:

$$G(f) = \begin{cases} \alpha f & 0 \leq f \leq f_0 \\ \frac{f_{max} - \alpha f_0}{f_{max} - f_0}(f - f_0) + \alpha f_0 & f_0 \leq f \leq f_{max} \end{cases}$$

Where f_{max} is the upper cut-off frequency and f_0 is defined to be larger than the highest significant formants.

A. Improvement proposal

VLTP method helped the author to find a viable solution to the IEMOCAP drawbacks. But other methods could be used like Generative adversarial networks.
GANs are powerful generative models that try to approximate the data distribution by training simultaneously two competing networks, a generator and a discriminator. A lot of research has focused on improving the quality of generated samples and stabilizing GAN training. Recently, the GAN ability to generate realistic in-distribution samples has been leveraged for data augmentation.

IV. The model

Instead of using state of the art machine learning architectures with hand-crafted acoustic features (MFCC, pitch, energy, ZCR...), the authors decided to work with deep learning methods and work on raw data. They transformed the raw data in Log-spectrograms which is then used as an input to convolutional layers. The convolutional layers extract the high-level features then feed the Recurrent layers (BI-LSTM). The authors chose the BI-LSTM model because it helps to alleviate the problem of vanishing gradient and it is efficient to model long-distance dependencies. The idea of LSTMs is to allow the network to ”forget” or disregard certain past observations in order to give weight to important information in the current prediction. A bidirectional LSTM is a combination of two LSTMs — one runs forward from “right to left” and one runs backward from “left to right”.

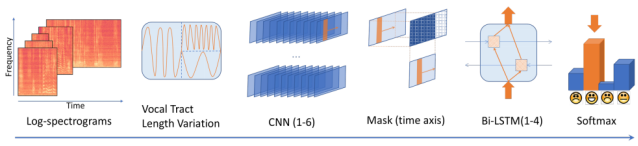

Fig. 2: The model used in the paper

A. Improvement proposal

Instead of using a BI-LSTM model it would be interesting to use a Transformer. Despite BiLSTM's recent success in the SER task, it has to compute utterance representations one by one, which massively hinders full exploitation of GPU's parallelism.
The Transformer is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. "The Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution." [1]. Here, "transduction" means the conversion of input sequences into output sequences. The idea behind Transformer is to handle the dependencies between input and output with attention and recurrence completely.
Moreover BI-LSTM is well-known to be a model complex to optimize.

V. Evaluation metrics

The authors used two metrics the Weighted accuracy (WA) and Unweighted accuracy (UA).
Weighted accuracy is computed by taking the average, over all the classes, of the fraction of correct predictions in this class (i.e. the number of correctly predicted instances in that class, divided by the total number of instances in that class). Unweighted accuracy is the fraction of instances predicted correctly (i.e. total correct predictions, divided by total instances). The distinction between these two measures is useful especially if there exist classes that are under-represented by the samples. Unweighted accuracy gives the same weight to each class, regardless of how many samples of that class the dataset contains. Weighted accuracy weighs each class according to the number of samples that belong to that class in the dataset.

A. Critics

This way to evaluate the data shows that data preparation can indeed influence the overall performance of a classification system.

	Baseline			Best model
Augmentation during training	-	-	+	+
Oversampling ($\times 2$) of happiness and anger	-	+	+	+
Frequency range (kHz)	4	4	4	8
Weighted accuracy	66.4	63.5	64.2	64.5
Unweighted accuracy	57.7	59.8	60.9	**61.7**

Fig. 3: 10-cross validation scores depending on the techniques applied (for each experiment we present the results corresponding to its best run)

Due diligence is essential to ensure that data used is of the highest possible quality. Changing the way you prepare your data may lead you to re-think your original question, which may ultimately help you develop a more successful model.

References

[1] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. *Attention Is All You Need, In Advances in Neural Information Processing Systems*. 2017.

[2] Caroline Etienne et al. *CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation, Workshop on Speech, Music and Mind 2018*. 2018.