Emotion Label Encoding using Word Embeddings for Speech Emotion Recognition

Eimear Stanley, Eric DeMattos, Anita Klementiev, Piotr Ozimek, Georgia Clarke, Michael Berger and Dimitri Palaz

Speech Graphics Ltd, Edinburgh, United Kingdom
{eimear.stanley, eric.demattos, anita.klementiev, piotr.ozimek, g.clarke, berger, dpalaz}@speech-graphics.com

Abstract

Speech Emotion Recognition (SER) is an important and challenging task for human-computer interaction. Human emotions are complex and human emotions are difficult to represent. The standard representations of emotions, categorical or continuous, tend to oversimplify the problem. Recently, the label encoding approach has been proposed, where vectors are used to represent the emotion space. In this paper, we hypothesise that using a pre-existing vector space that encodes semantic information about emotion is beneficial for the task. To this aim, we propose using word embeddings obtained from a Language Model (LM) as labels for SER. We evaluate the performance of the proposed approach on the IEMOCAP corpus and show that it yields better performance than a standard baseline. We also present a method to combine free text labels, which are unusable in conventional approaches, and by doing so we show that the model can learn more nuanced representations of emotions. A major novelty of the proposed approach is its ability to take into account different labels among annotators. Our hypothesis is that this approach could be the key to encoding emotions efficiently and bringing more nuanced emotion information to the model.

The key contributions of the paper are: (1) we present a novel approach where the model is trained to predict the word embeddings of the emotion labels, as opposed to the label itself. Using the IEMOCAP corpus [5], different loss functions are compared along with a baseline using the same model architecture trained with the standard approach. We show that the proposed approach yields better performance than the baseline, and yields performance on par with recent literature. A major novelty of the proposed approach is its ability to use any free text labels without mapping, which is impossible with conventional approaches. Based on this property, we propose an approach to combine individual annotations from multiple annotators and free text comments to investigate bringing more nuanced emotion information to the model. We also present a study on leveraging natural language to analyse the trained models and show that the models learn more nuanced emotion representations.

The key contributions of the paper are: (1) we present a novel approach to use word embeddings as labels for SER and show its benefits, and (2) we propose a novel approach to combine emotion labels, including free text labels and show its potential to learn better emotion representations.

1. Introduction

Speech Emotion Recognition (SER) uses machine learning models to classify emotion from audio, which rely on accurate labels. However, human emotions are highly complex and emotion annotation is a difficult and subjective task [1]. The two common representations of emotion are categorical and continuous. The categorical approach uses a small set of discrete labels (angry, happy, sad, etc.) and relies on majority voting across a number of annotators to establish one ground truth label. However, this approach may oversimplify the task as it overlooks the complexity and natural subjectivity of emotion perception. Additionally relying on majority voting may exclude relevant information in the form of individual annotations, as well as data without agreement. The continuous approach looks to represent emotions in a continuous space, most commonly using arousal, valence, and dominance as dimensions. This has the same problems as the categorical approach, as well as potentially being more challenging to annotate.

Recently, using categorical labels in the framework of label encoding has been investigated. In this framework categorical labels are used to generate a vector, which is then used as a target for training the model. Two approaches have been proposed for SER: soft labels and metric learning. In the soft labels approach [2], the vectors are probability distributions computed from the individual annotations from each annotator. The metric learning approach [3] proposes to learn the vector space using a triplet loss function. Both approaches have shown some improvements on the SER task and have made first steps towards a better representation of the emotion space.

2. Related Work

In this section we present the related work for word embeddings and for label encoding approaches for SER. For more details on other approaches in SER see [6, 7].

Word embeddings are vectors representing words in a space which preserve syntactic and semantic properties to some extent, which can then be used in different NLP tasks. Word embeddings are often computed using neural networks-based models [8] and are typically learned within a LM. The most popular language models are word2vec [4], GloVe [9], BERT [10] and GPT-3, [11] as used in the popular application ChatGPT.
Word embeddings have recently been used to represent emotions in the context of speech synthesis. They were originally introduced in [12] as style tags for the synthesised speech, and the approach was then extended to use word embeddings computed from emotion labels [13].

In the context of SER, one recent approach to representing emotion is metric learning, where a measure of distance is learned using similarity between classes. Most commonly the triplet loss function is used, where the distance between an anchor, a positive and a negative sample is optimised, which has been shown to improve SER model performance [3, 14, 15]. Recent work [16] shows further improvements when integrating an additional sentiment constraint within the triplet loss function. A denoising autoencoder using a continuous metric loss based on either activation or valence labels was proposed in [17] to improve generalisation across languages.

To the same aim, recent approaches have sought to account for annotations from multiple annotators and model the subjectiveness of the SER task. The most common approach uses soft labels, which typically converts the traditional one-hot emotion label attained through majority voting to a probability distribution taking into account the individual annotations. Numerous previous studies have successfully used soft labels on the task of SER [2, 18, 19, 20]. More recently, [21] leverages the co-occurrence of emotion labels to create a weighted matrix, while [22] investigates modeling the uncertainty of emotion labels with Dirichlet priors.

### 3. Methodology

#### 3.1. Label encoding using word embeddings

In the label encoding approach, the model is trained to predict vectors corresponding to the label, rather than the label index \( y \). In this paper, vectors are word embeddings from a dictionary obtained from a pre-trained LM. This dictionary \( W \) has an entry for all words in the English language. The vectors \( w \) used for training are obtained by getting the entry of the dictionary corresponding to the name \( y \): \( w = W(y) \).

The model \( f(\cdot) \) is trained by minimising the distance between its output \( f(x) \) for input utterance \( x \) and the target word embedding \( w \). The loss functions \( \mathcal{L}(x, w) \) investigated in this paper are the Mean Square Error (MSE), Mean Absolute Error (MAE) and Log Cos (LC).

\[
\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_n (f(x_n) - w_n)^2 \quad (1)
\]

\[
\mathcal{L}_{\text{MAE}} = \frac{1}{N} \sum_n |f(x_n) - w_n| \quad (2)
\]

\[
\mathcal{L}_{\text{LC}} = \frac{1}{N} \sum_n \log(\cosh(f(x_n) - w_n)) \quad (3)
\]

#### 3.2. Combining encoded labels

Due to the complexity and subjectivity of emotion, it is common for a given utterance to have more than one label. In an attempt to represent emotion in a more fine-grained and subtle manner, we use one of the most useful properties of word embeddings: their ability to capture semantic linear substructures, which enable simple arithmetical operations. For instance, word2vec [4] reports that \( W(\text{Rome}) = W(\text{Paris}) - W(\text{France}) + W(\text{Italy}) \). Taking inspiration from this principle, we investigate an approach to combine encoded labels using vector geometry. The combined word embedding \( w_{\text{new}} \) is computed as the average of the different word embeddings \( w_i \) obtained from the different labels \( y_i, i = 1, ..., L \):

\[
w_{\text{new}} = \frac{1}{L} \sum_{i=1}^{L} w_i = \frac{1}{L} \sum_{i=1}^{L} W(y_i). \quad (4)
\]

For instance, if a given utterance has three labels \{anger, sadness, sadness\}, the new word embedding will be on the line between the word embedding for sadness and the word embedding for anger, at \( 1/3 \) of the distance from sadness and \( 2/3 \) from anger.

#### 3.3. Evaluation

In order to evaluate the output of the models trained on word embeddings, a metric measuring the distance between embeddings is first needed. Following GloVe [9], we use the cosine similarity defined as the cosine of the angle between two embeddings.

In the SER literature, two metrics are commonly used: the Weighted Accuracy (WA), and the Unweighted Accuracy (UA), which is the average of the per-class accuracies. To compute these metrics with the proposed models using word embeddings, their output needs to be mapped to the index labels. The prediction \( \hat{y} \) is therefore computed by finding the closest embedding to the output \( w \) from the embedding dictionary \( W_Y \) using the cosine similarity \( s_{\text{cos}} \).

\[
\hat{y} = \arg\min_{s} s_{\text{cos}}(w, W_Y(s)), \quad (5)
\]

where \( W_Y \) is a subset of the word embedding dictionary \( W \) containing only entries corresponding to the label set (for example: \{anger, happiness, neutral, sadness\}).

### 4. Experimental Setup

#### 4.1. Database

IEMOCAP is an audio-visual corpus totalling 12 hours of data from 10 speakers [5]. Each utterance was labeled by three annotators choosing from a discrete set of categorical emotion labels. Annotators were permitted to use one or more labels per utterance and could leave free text comments if they felt none of the labels were adequate. For the purpose of this study and in line with previous works [23], majority voting is used and only utterances with anger, excitement, happiness, sadness, and neutral ground truth labels are used, with excitement merged with happiness.

#### 4.2. Encoded label sets

The following three label sets are used to train the models using word embeddings. All sets use the same audio data, which is selected as described in section 4.1. We use the noun form of all emotion labels.

- **Baseline set** \( S_1 \) is created by extracting the word embeddings from the ground truth labels. We refer to this as the 4-label set.
- **Annotator set** \( S_2 \) is created by extracting the individual annotations from multiple annotators and combining them as presented in Section 3.2, leading to 129 different embeddings targets. This means that labels outside of the four label sets are included, for example frustration. The other label is discarded as its meaning is unclear and could confuse the model.
5. Results

5.1. Recognition studies

5.1.1. Evaluation on the 4-label set

In this study, the outputs of the three models are mapped to the 4-label set as described in Section 3.3. We compare them with the baseline and related work. Table 1 presents the results of the models using word embeddings, trained with the three losses presented in Section 3.1, along with the baseline. All three models outperform the baseline, showing the benefits of using word embeddings as labels. The $M_1$ model yields the best performance, which is expected as it is the only model trained exclusively on the 4-label set. The performance gap of $M_2$ and $M_3$ compared to $M_1$ is very small, which shows another benefit of the proposed approach as training models on hundreds of different labels tends to be challenging when using categorical labels. The three loss functions yield similar performances: the LC function being slightly better for $M_1$ and the MSE loss for $M_2$ and $M_3$. For fair comparison across all three models, going forward, we use the same loss function. We select the LC as it yields the highest singular performance.

As seen in Table 2, $M_1$ yields performance on par with recent work but not reaching state-of-the-art. This is expected as this work is an exploratory study and did not include an extensive hyper-parameter search or fine-tuning of the wav2vec features.

5.1.2. Evaluation on combined labels

In order to further evaluate all three models, we compute the average cosine similarity between the predicted and target embeddings for the three label sets. The results are presented in Table 3. $M_2$ and $M_3$, trained on combined labels, have a higher similarity on $S_3$ and $S_3$, showing that the models have learned a more fine-grained representation of the word embedding space. Contrary to expectation, the best model on $S_3$ is not $M_2$ but $M_3$, which shows that adding the comments is beneficial. This highlights one of the key benefits of our approach, as free text comments are very difficult to use in standard approaches and so are typically discarded.

Table 3: Average cosine similarity across the label sets

<table>
<thead>
<tr>
<th>Model</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>0.789</td>
<td>0.764</td>
<td>0.759</td>
</tr>
<tr>
<td>$M_2$</td>
<td>0.756</td>
<td>0.834</td>
<td>0.831</td>
</tr>
<tr>
<td>$M_3$</td>
<td>0.756</td>
<td>0.841</td>
<td>0.839</td>
</tr>
</tbody>
</table>
5.2. Exploratory studies using natural language

One of the major appeals of the proposed approach is that the word embedding space not only contains the target labels but the entire vocabulary from the word embedding dictionary. In this section, we use natural language to further assess our hypothesis that the combined label sets can enable the models to learn more nuanced and subtle representations of emotion. More specifically, we use nearest neighbours in both qualitative and quantitative analysis.

5.2.1. Qualitative analysis

To illustrate the benefit of using a semantic space for label encoding, we look at the closest embeddings to the target embedding. Table 4 presents the 5 nearest neighbours to a small number of example labels from $S_1$, $S_2$ and $S_3$. The combined sets, $S_2$ and $S_3$, encapsulate the meanings from multiple labels and better model entangled emotions. To evaluate if the models are able to capture this, we perform the same analysis of finding the closest embedding, but on the predicted embeddings. Table 5 presents the top 5 nearest neighbours to the predicted embeddings for each of the four emotion labels in $S_1$. First, this shows that the model is able to learn the word embedding space. On closer inspection, we can see that the nearest neighbours vary across the three models, for example, the neighbours of neutral in $M_1$ are more syntactically related while in $M_2$ and $M_3$ they are more semantically related. Also of note, is that sadness is a neighbour to happiness in $M_1$, and not in $M_2$ and $M_3$. This suggests that the models trained on the combined sets capture a more informed representation.

5.2.2. Quantitative analysis

This experiment uses the top-10 accuracy, which measures the accuracy of a given word being present in the 10 nearest neighbours within the combined label respectively. We compute the top-10 accuracy of $l_1$ and $l_2$ for each prediction on $S_2$ and $S_3$ across the three models. The results are presented in Table 6. The top-10 accuracy on $l_1$ shows that training on the combined labels does not hurt the ability of the model to predict the majority label, conversely the accuracy improves for $M_2$ and $M_3$. Additionally, the top-10 accuracy on $l_2$ is higher for $M_2$ and $M_3$, which shows that this approach helps the models predict a secondary label. This indicates that $M_2$ and $M_3$ are able to better capture the nuance of emotions. Overall, these two findings indicate that the predictions from $M_2$ and $M_3$ are better situated in the word embedding space, showing the benefit of the combined label approach.

Table 4: Top 5 nearest neighbours of the target labels. Those in bold match an emotion from the target label.

<table>
<thead>
<tr>
<th>Set</th>
<th>Label Description</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>happiness, happiness, happiness, happiness</td>
<td>happiness</td>
<td>joy</td>
<td>contentment</td>
<td>bliss</td>
<td>prosperity</td>
</tr>
<tr>
<td>$S_2$</td>
<td>amazed, enthusiastic, happiness, happiness, happiness, happiness, impressed</td>
<td>happiness</td>
<td>joy</td>
<td>contentment</td>
<td>excitement</td>
<td>bliss</td>
</tr>
<tr>
<td>$S_3$</td>
<td>bantering, excitement, happiness, happiness, joking</td>
<td>happiness</td>
<td>laughter</td>
<td>joy</td>
<td>excitement</td>
<td>contentment</td>
</tr>
</tbody>
</table>

Table 5: Top 5 nearest neighbours of the predictions

<table>
<thead>
<tr>
<th>Model</th>
<th>Emotion label</th>
<th>anger</th>
<th>neutral</th>
<th>happiness</th>
<th>sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>anger, frustration, resentment, grief, rage</td>
<td>neutral</td>
<td>opportunity</td>
<td>happiness</td>
<td>sad</td>
</tr>
<tr>
<td></td>
<td></td>
<td>neutral</td>
<td>neat</td>
<td>sad</td>
<td>sorrow</td>
</tr>
<tr>
<td>$M_2$</td>
<td>frustration, anger, resentment, fear, rage</td>
<td>neutral</td>
<td>opportunity</td>
<td>joy</td>
<td>sad</td>
</tr>
<tr>
<td></td>
<td></td>
<td>neutral</td>
<td>sense</td>
<td>happiness</td>
<td>sad</td>
</tr>
<tr>
<td>$M_3$</td>
<td>frustration, anger, resentment, fear, rage</td>
<td>neutral</td>
<td>opportunity</td>
<td>excitement</td>
<td>sad</td>
</tr>
<tr>
<td></td>
<td></td>
<td>neutral</td>
<td>feeling</td>
<td>happiness</td>
<td>sad</td>
</tr>
</tbody>
</table>

6. Conclusion

In this paper we presented an exploratory study using word embeddings as labels for speech emotion recognition. This included a novel approach to combine multiple emotion labels, including free text comments. We showed that the proposed approach yields better performance than the standard approach using the same model architecture, highlighting the benefit of using a label encoding space which already encodes semantic relations between emotions. Additionally, we presented a study using natural language, which indicates that the model trained on the combined label sets learned a more informed and nuanced representation of emotion, and that the proposed approach can leverage semantic relationships between emotions. These are novel insights which could pave the way for new developments towards improving SER models and their interpretability. For future work, we will investigate applying this approach to cross-corpus training as it could address one of its main issues: the mismatch between label sets. We will also investigate filtering the word embedding space.

In terms of limitations, the approach is restricted by both the data and word embeddings, which are readily available in English, but are less available for a wider range of languages. Additionally, using word embeddings as label encodings introduces new biases to the SER task, which otherwise would not exist, as language models are known to manifest potentially harmful social biases [31].
7. References


