Modeling Subjectiveness in Emotion Recognition with Deep Neural Networks: Ensembles vs Soft Labels

Haytham M. Fayek, Margaret Lech, Lawrence Cavedon
{School of Engineering, School of Science}, RMIT University, Melbourne, VIC 3001, Australia

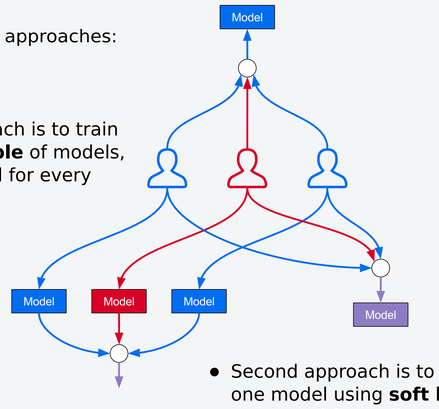
Highlights

- We propose two approaches for modeling subjectiveness in subjective classification tasks and demonstrate that both approaches achieve improved results in a speech emotion recognition system.
- It was recently shown that soft labels obtained by training large ensembles can be used to distill knowledge from large ensembles to a single small model. We propose a method to generate soft labels directly from multiple annotations that leads to a similar performance gain without having to train large ensembles.

Summary

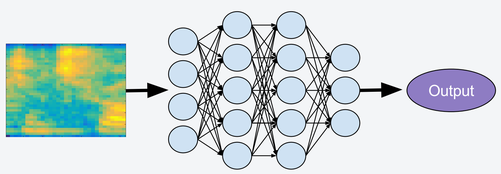
Subjective classification tasks (e.g. emotion recognition) lack ground truth labels. Labels are usually obtained by majority voting or averaging between multiple annotators, which does not reflect the inter-annotator variability and may lead to the omission of valuable information.

We propose two approaches:



- First approach is to train an **ensemble** of models, i.e. a model for every annotator.
- Second approach is to train only one model using **soft labels**, generated from all annotators.

We also propose a deep neural network that recognizes emotions in real-time directly from a one-second window of raw speech spectrograms achieving a speech emotion recognition system with a simple pipeline and low latency.



Methodology

Data: IEMOCAP database. Four classes: anger, happiness, sadness and neutral. Utterances were labeled by three annotators. Ground truths (hard) labels were obtained by majority voting, where 74.6% of the utterances were agreed upon by at least two annotators.

Preprocessing: Speech was analyzed using a 25 ms Hamming window with a stride of 15 ms. Spectrograms were generated using 41 log Fourier-transform based filter-banks on a linear scale. Every 64 consecutive frames in an utterance were concatenated to form a one-second window. No speaker dependent operations were performed.

Deep Neural Network: The architecture of the network had 7 feed-forward fully- connected layers. All hidden layers had 1024 Rectified Linear Units (ReLUs). The output layer was a four-way softmax layer as in (1), producing the posterior class probabilities:

$$y^{(L)} = \frac{\exp(z^{(L)})}{\sum_{k=1}^{K} \exp(z^{(L)})} \quad (1) \qquad z^{(L)} = y^{(L-1)} W^{(L)} + b^{(L)} \quad (2)$$

A cross-entropy cost function was used as in (3) that has a simple derivative as in (4), making it straightforward to use either hard labels or soft labels.

$$C = -\sum_{k=1}^{K} Y_k \log(y_k^{(L)}) \quad (3) \qquad \frac{\partial C}{\partial y^{(L)}} = y^{(L)} - Y \quad (4)$$

Model Ensembling: To model inter-annotator variability, an ensemble of deep neural networks was trained such that each network represented an annotator. Two ensemble combination rules: [a] the geometric mean of the posterior probabilities as in (5) and [b] unweighted majority vote as in (6).

$$y_k = (\prod_{n=1}^{N} y_{n,k})^{1/N} \quad (5) \qquad t = \underset{k=1,\cdots,K}{\operatorname{argmax}} \sum_{n=1}^{N} d_{n,k} \quad (6)$$

Label Encoding: Using one-of-K (one-hot) encoding, soft labels were generated from multiple one-of-K encoded labels as in (7):

$$s = \frac{\sum_{n=1}^{N} h^{(n)}}{\sum_{k=1}^{K} \sum_{n=1}^{N} h^{(n)}} \quad (7)$$

Table 1: Label Encoding Scheme for Three Annotators

Annotation	Hard Label	Soft Label
[ang][ang][ang]	[1,0,0,0]	[1, 0, 0, 0]
[hap][neu][neu]	[0,0,1,0]	[0, 0.33, 0.66, 0]
[sad][sad][sad;neu]	[0,0,0,1]	[0, 0, 0.25, 0.75]

Results

Table 2: Test Performance using Hard Labels (Baseline)

Test Set	Average Recall	Average F-Score
Prototypical	57.55%	53.55%
Full Test Set	47.62%	46.66%

Table 3: Test Performance using an Ensemble of Networks

Geometric Mean		
Test Set	Average Recall	Average F-Score
Prototypical	58.91%	53.07%
Full Test Set	49.53%	48.47%
Majority Voting		
Test Set	Average Recall	Average F-Score
Prototypical	58.64%	53.36%
Full Test Set	49.01%	48.10%

Table 4: Test Performance using Soft Labels

Test Set	Average Recall	Average F-Score
Prototypical	58.85%	53.20%
Full Test Set	49.18%	48.01%

Conclusion

A deep neural network that maps one-second windows of speech spectrograms into emotions in real-time was presented with promising results.

Data prototypicality has a significant effect on the model's performance.

Model ensembles and soft labels were proposed to model subjectiveness. Both approaches outperformed the baseline model with ground truth labels.

Empirical results showed that we were able to retain the performance gain of the ensemble over the baseline model using soft labels generated directly from multiple annotators.

Acknowledgment: This research was funded by the Vice-Chancellor's PhD Scholarship (VCPS) from RMIT University. We gratefully acknowledge the support of NVIDIA Corporation for the donation of a Tesla K40 GPU.

haytham.fayek@ieee.org

2016 International Joint Conference on Neural Networks, Vancouver, Canada