Language-independent Emotion Recognition from Speech

Ayhan Kaplan
University of Stuttgart
Institute for Natural Language Processing
ayhan.kaplan@ims.uni-stuttgart.de

Abstract

Automated recognition of emotions from spontaneous speech is a challenging but engaging task of Natural Language Processing. In this work, we conduct experiments using a Convolutional Neural Network to predict language-independent emotion on a dataset of English and French speech. We researched on how different activation functions, optimizers and the usage of language-alternating batches effect the overall model performance of the neural network. This paper will particularly focus on evaluation and discussion of different optimization algorithms and their effect on the overall model performance.

1 Introduction

Among Natural Language Processing tasks like Sentiment Analysis and Machine Translation, Emotion Recognition on Natural Speech has reached a point where it is no longer a task of the future but one that will have an impact in the usability of todays modern technologies and intelligent systems. Reacting to prosodic features in speech like recognizing emotion or being context-aware in general represents a desirable feature for many commercial products to immerse the users interaction which makes it an interesting topic especially in the field of ubiquitous computing in human-computer interaction. There has been a variety of related work to this domain of Deep Learning in which the researchers applied models to recognize emotion from Natural Speech using acoustic, prosodic and cepstral features. (JLCW15) (LT15) Following up to these, there are several approaches on cross-language and multilingual emotion recognition (FS15)(ZWR11)(EBSS03). In this work, we focus on defining a model for language-independent emotion recognition from natural speech. We distinguish between multilingual, cross-language and monolingual training. After comparing the results of these 3 tasks, we conduct experiments on this working baseline to evaluate the effect of different hyper-parameters to the networks overall performance such as activation functions, optimizers and language alternating batches.

2 Data

We use two corpora of emotional speech on English and French in order to train and evaluate our model. For English speech we use the Interactive Emotional Dyadic Motion Capture database (BBL$^+$08) in which scripted and spontaneous spoken communication scenarios were recorded and provided in $10,039$ samples. For French speech we use the database for Remote Collaborative and Affective Interactions (Recola) which provides $1,308$ samples of speech (RSSL13). Both corpora measure emotion on two dimensions: *arousal* and *valence*. The annotation scheme we use in this project consists of a binary classification of arousal (low/high) and valence (negative/positive) as described by Neumann and Thang Vu (NT18). Thus, we distinguish between 4 emotion classes for the representation of each arousal/valence combination in the annotation (Table 1).

Arousal \ Valence	positive	negative
high	joy	anger
low	pleasure	sadness

Table 1: A coarse-grained binary classification of valence and arousal for the audio samples in our corpus.

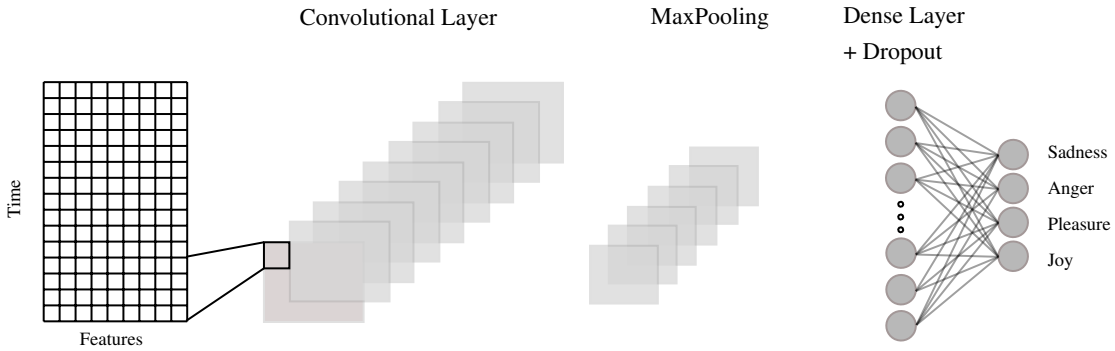

Figure 1: Architecture of our CNN consisting of input matrix, convolutional layer, maxpooling layer, dense layer with dropout regularization and finally the 4 output classes.

2.1 Processing the dataset

Using the openSMILE (EWGS13) toolkit, we split each audio sample into overlapping frame segments of (where one frame has length of $10ms$) and extracted the first 13 MFCC features (TC02) for each frame. Since the audio samples differ in duration, we computed the arithmetic mean of the sum over all durations and normalized them to the resulting average length, that is, cutting off the last frames of utterances that were longer and filling up the shorter ones with zero-frames (appending zero-vectors to the end of them). Having the data normalized in respect to their average frame count f, we processed each utterance into a matrix representation: $W \in \mathbb{R}^{f \times d}$ (d being the dimension of the feature vectors with 13 MFCC features). Lastly, we split the corpus into following sets:

- 80% Training
- 10% Validation
- 10% Test

Furthermore, the validation and test data are divided into 4 class-separated sets and written out into binaries. This allows us to evaluate the accuracy for each emotional class. We did not distinguish if the speaker of the utterances was male or female. The resulting corpus is unbalanced since there are big differences in the number of utterances for each class.

3 Model

We applied a model for multi class prediction with respect to emotion categories from spoken language. This model consists of a convolutional neural network (CNN) with one convolutional layer followed by a max pooling layer. Following, the output is fed into a softmax layer after applying dropout regularization (Hin14) to prevent our network from overfitting. The input to our CNN is a matrix representation of an audio sample (as introduced in section 2)

4 Experimental Setup

We trained our model for binary classification on arousal/valence for the following tasks:

Monolingual Train and evaluate the model on the same language

Multilingual Train and evaluate the model on both languages

Cross-language Train the model on one language and evaluate it on another

In the case of multilingual training the mean for the average frame count of each corpus is computed by the network and the input dimensions are normalized analog to the process described in subsection 2.1.

Hyper-parameters

In order to implement the CNN model, we used Google's open source machine learning framework Tensorflow (ABB+15). Since we have sparse input data, we apply an adaptive learning rate method for training (KB15). The system consists of 50 Kernels with filter dimension $10 \times d$ in the convolutional layer (convolving over all $d = 13$ MFCC filter-banks), a mini-batch size of 50 and max pooling size of 30 with stride 3. We further apply dropout regularization to the last hidden layer with dropout rate 0.5 to prevent our

network from overfitting. We run 50 epochs for each experiment whereby the training data is shuffled after every epoch to avoid sessions that could bias the optimization algorithm by curriculum-like training.

5 Experimental Results

As mentioned in section 4 we evaluated our model for three tasks. Table 3 can be considered for fine-grained results of our individual experiments. We depicted the accuracy for each emotion class (defined in Table 1) and the micro-average per task which has been observed to be the best measure to represent the overall accuracy of a dataset that is not balanced with respect to the number of samples per class. As shown in the table, we generally observe the highest accuracy for the emotional class *Joy*. This is most likely due to the fact, that Joy has full activation (arousal) and positive valence. It seems that our model performs fine on recognizing positive valence and high arousal. That brings us to the realization that higher activation like laughter or excited speech can be captured with lower resolution but more stationary emotions in speech (Pleasure, Sadness) need much higher attendance to be recognized as such. Furthermore, it is noticeable that the accuracy distribution is much more balanced for testing on the French corpus than it is on the English one. Since the French corpus is smaller there are lower absolute differences in the number of samples per class which can possibly yield a better distribution over all emotion classes. However, this also means that the test sets for the individual emotion classes are very small for Recola which can lead to issues like the invariant accuracy for *Anger* in the French test set. The comparison of the three baseline tasks shows that monolingual and multilingual training yield approximately similar results for micro-average performance even though there are still large margins in accuracy between the individual emotion classes. Overall, testing on French data yields notably higher micro-average in performance for monolingual and multilingual tasks. This accuracy drops below the English testing results for cross-language training can again be lead back to the high difference in dataset size for our corpora. Lastly, the fact that our model performs well for Joy in all tasks shows the effectiveness of language-independent emotion recognition with respect to arousal activation. On the other hand, these results validate the hardness of valence prediction.

Hyper parameter	Model
Activation	Rectified Linear Units
Loss	Softmax Cross Entropy
Optimizer	ADAM
Mini-Batch	50
Convolution Filter	10×13
Kernels	50
Maxpooling	30×1
Dropout	0.5
Initial Learning Rate	0.001
Epochs	50

Table 2: Main parameters of the CNN model for language-independent emotion recognition

		Accuracy in %		
Tested on	Emotion	Mono	Multi	Cross
English	Sadness	00.00	01.52	01.51
	Anger	01.93	01.45	01.45
	Pleasure	04.37	01.10	12.02
	Joy	94.22	98.55	86.42
	micro	42.14	43.27	40.52
French	Sadness	07.70	00.00	23.08
	Anger	20.00	20.00	20.00
	Pleasure	35.71	03.57	35.71
	Joy	75.44	91.23	40.35
	micro	53.40	52.43	35.92

Table 3: Experiment results on the model introduced in section 3 for monolingual-, multilingual- and cross-language training. The accuracy is given for testing on English (top) and French (bottom) test set.

6 The effect of different optimizers to the network

As we implemented our model using the Tensorflow library, concepts like optimization algorithms and activation functions are method-calls and pretty much a black box to the developer. In order to create transparency over the distinction of optimization algorithms we used in the following experiments we discuss the significant parts that they differ in.

Gradient Descent

The general approach to find a local minimum of a objective function Φ with parameter set θ is called Gradient Descent (or: Batch Gradient Descent). After computing the loss, Gradient Descent tunes the network parameters for the whole set:

$$\theta = \theta - \eta \cdot \nabla_\theta \Phi(\theta) \tag{1}$$

In our experiments, we used Mini-Batch Gradient Descent to get an update for every mini-batch instead of back-propagating once for the whole dataset. However, there are many optimizations for Gradient Descent from which we will present a selection of promising algorithms to improve the model performance.

Adagrad

Adagrad (DHS11) is an gradient-based optimization that adapts the learning rate to the parameters. Instead of using the same learning rate on every parameter, Adagrad uses an individual learning rate η for each parameter θ_i at every time step t with objective function Φ:

$$g_{t,i} = \nabla_{\theta_t} \Phi(\theta_{t,i}) \tag{2}$$

The modification of the learning rate based on the past gradients of the current parameter (4) by stochastic gradient descent update (3) will be computed with diagonal matrix $G_t \in \mathbb{R}^{d \times d}$:

$$\theta_{t+1,i} = \theta_{t,i} - \eta \cdot g_{t,i} \tag{3}$$

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i} \tag{4}$$

Adadelta

While Adagrad stores the past squared gradients to update the learning rate at each time step, Matthew D. Zeiler introduced a learning rate method called Adadelta (Zei12). This optimization method provides a per-dimension learning rate with no need of hyper parameter tuning. This is established by defining the sum of gradients recursively as a decaying average of all past squared gradients with Momentum (Qia99):

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1 - \gamma) g_t^2 \tag{5}$$

For brevity, we take Equation 4 and replace the diagonal matrix G_t with the decaying average over past squared gradients:

$$\Delta\theta_t = -\frac{\eta}{E[g^2]_t + \epsilon} g_t \tag{6}$$

Adam

The Adaptive Moment Estimation (KB15) is a method for stochastic optimization that only requires first-order gradients and yields high memory efficiency. Adam extends Adadelta by storing the exponentially decaying average of past gradients, whereas m_t and v_t are estimates of the first (mean) and the second moment (variance) of the gradients. To prevent zero-biasing, the first and second estimates are computed in a bias-corrected manner:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{7}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{8}$$

Finally the bias-corrected estimates replace the gradient and decaying average of Adadelta:

$$\Delta\theta_t = -\frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t \tag{9}$$

Follow The Regularized Leader

For the sake of variety, we modified our network with an approach that stands out from the group of adaptive gradient descent algorithms. The FTRL Proximal (MHS$^+$13) is an Mirror Descent Algorithm (MDA) and comes with L_1- and L_2-Regularization for Logistic Regression which performs great in Online Learning. MDAs are in fact a variant of Gradient Descent (Beck et al (BT03)) where the square norm is replaced by the Bregman Distance (Bre67)

6.1 Experimental Results for alternative Optimizers

We trained our model as described in section 3 for each of the aforementioned optimizers. Table 4 presents averaged results across 4 runs of multilingual training with 50 epochs each. As seen in the experiment on our baseline model architecture the best performance is in general on the emotion class Joy. Merely the AdaDelta optimizer seems to be an outlier in all runs since our model achieves 100% accuracy for Anger on the English test set through all runs with this optimization method. None of the optimizers have worked well at low arousal and negative valence (Sadness) for both datasets. Only AdaDelta, again, is an exception at this point with an accuracy of over 68% for Sadness on the French corpus. The combination of high arousal without positive valence (Anger)

Activation on ReLU						
		Accuracy in %				
Tested on	Emotion	Mini-batch SGD	AdaGrad	AdaDelta	Adam	FTRLProximal
English	Sadness	00.00	00.00	00.00	00.07	00.00
	Anger	00.97	12.32	100.0	00.97	22.95
	Pleasure	23.77	10.93	00.00	00.55	05.74
	Joy	64.02	84.68	00.00	98.27	81.65
	micro	33.29	42.21	15.33	42.83	42.46
French	Sadness	00.00	03.85	68.88	03.85	00.00
	Anger	00.00	00.00	01.69	20.00	00.00
	Pleasure	00.00	07.14	02.19	03.57	26.79
	Joy	93.86	92.11	31.67	92.11	86.84
	micro	51.94	53.40	22.51	53.40	55.34

Table 4: Experimental results on the model introduced in section 3 for differing optimizers. The accuracy is given for testing on English and French dataset.

seems to be hard to predict as the highest recorded accuracy across all experiments was 22.95% (if we ignore AdaDelta at this point). We observed that the accuracy on the emotion class Pleasure happens to be an arbitrary distribution over the researched optimization methods with the highest performance on FTRL for French (26.79%) and on mini-batch Gradient Descent for English (23.77%). Regarding micro-averaged accuracy, we found that AdaGrad, Adam and FTRL perform similar at multilanguage classification tasks in context of emotion recognition. It was to be expected that AdaGrad and Adam perform alike since Adam is a refinement of AdaGrad. It is still surprising that AdaDelta had such a dissimilar performance on the same task as it is also closely related to Adam and AdaGrad. To clear any doubt aside, we started the experiment again with 4 runs for AdaDelta and achieved similar results. Overall, we observed that most optimizers perform very well in prediction of high arousal and positive valence and sometimes high arousal also results in a higher accuracy in combination with negative valence. Yet, the results for all optimizer functions were too similar as we could draw any meaningful conclusions from them.

6.2 Extension of the Experiment

Since we were not able to draw firm conclusions from the findings in subsection 6.1 we modified the baseline architecture by using Leaky ReLU instead of ReLU to get a higher range of accuracy distribution and strive to avoid hard-zero sparse representation through "dying" neurons for emotion classes that are harder to predict. We trained our model similar as described in subsection 6.1 with averaged results across 4 runs of multilingual training with 50 epochs which are presented in Table 5. It is noticeable that AdaDelta has now gained a higher distribution over all classes but seems to be an outlier again regarding the emotional class Anger tested on the French validation set. That fact that the higher accuracies are trending in the direction of Joy consolidates our former hypothesis for the best overall performance for high arousal and positive valence. Furthermore, similar to our first experiment the accuracy over individual emotion classes are constantly lower for negative valence and lower arousal (Sadness/Anger). Yet, it cannot be unseen that the Adam optimizer is now way more balanced over the classes and not too overfitted to Joy as it was in Table 4 without forfeiting performance in the micro-averaged results. Overall, this extension of our experiment lead to the conclusion that all of the optimization algorithms, especially those which provide adaptive learning rate for standard SGD, as well as the FTRL Proximal (which is actually Online Learning SGD with integrated regularization) perform more or less similar on multilingual training. Solely, AdaDelta seems not to be the optimizer-of-choice in our findings.

Activation on Leaky ReLU						
		Accuracy in %				
Tested on	Emotion	Mini-batch SGD	AdaGrad	AdaDelta	Adam	FTRLProximal
English	Sadness	03.03	01.52	00.00	07.58	00.00
	Anger	30.43	18.84	42.03	39.89	15.22
	Pleasure	21.31	07.65	06.01	44.54	04.10
	Joy	62.72	80.20	54.48	52.17	89.88
	micro	40.02	41.33	35.72	43.58	43.64
French	Sadness	00.00	00.00	07.69	15.38	00.00
	Anger	00.00	00.00	50.00	10.00	00.00
	Pleasure	28.57	28.57	16.07	32.14	17.86
	Joy	74.56	85.09	37.72	71.05	91.23
	micro	49.03	54.85	28.64	50.49	55.34

Table 5: Experimental results on the modified network for differing optimizers and activation on Leaky ReLU. The accuracy is given for testing on English and French dataset.

7 Conclusion

We presented a model for language-independent emotion recognition from speech and conducted experiments based on multilingual, monolingual and cross-language training. We further have shown that the results correlate to the speech features arousal and valence. According to the baseline results it became clear that multilingual classification can be achieved with an architecture that works well on monolingual tasks. Based on this model, we conducted experiments on how different optimizers perform with ReLU and Leaky ReLU activation. Our overall findings emphasize the thesis of the hardness in prediction of emotion with respect to valence in given speech. In conclusion, we found that high arousal is leading to good performance for language-independent emotion recognition on a CNN with adaptive learning rate.

References

[ABB+15] Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Vi, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *ArXiv e-prints*, 2015.

[BBL+08] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[Bre67] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics, Volume 7, Issue 3, 1967, Pages 200-217*, 1967.

[BT03] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. 31:167–175, 05 2003.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research 12 (2011) 2121-2159*, 12:2121–2159, 2011.

[EBSS03] Florian Eyben, Anton Batliner, Dino Seppi, and Stefan Steidl. Cross-Corpus Classification of Realistic Emotions – Some Pilot Experiments. In *Proceedings 3rd International Workshop on EMOTION: Corpora for Research on Emotion and Affect, satellite of LREC 2010*, 2003.

[EWGS13] Florian Eyben, Felix Weninger, Florian Gross, and Bjoern Schuller. Recent developments in

opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 835–838, 2013.

[FS15] Silvia Monica Feraru and Dagmar Schuller. Cross-Language Acoustic Emotion Recognition : An Overview and Some Tendencies. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII) Cross-Language*, pages 125–131, 2015.

[Hin14] Geoffrey Hinton. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. 15:1929–1958, 2014.

[JLCW15] Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu. Speech Emotion Recognition with Acoustic and Lexical Features. pages 4749–4753, 2015.

[KB15] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *ICLR 2015*, pages 1–15, 2015.

[LT15] Jinkyu Lee and Ivan Tashev. High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition. *INTERSPEECH 2015*, pages 1537–1540, 2015.

[MHS+13] H Brendan Mcmahan, Gary Holt, D Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. Ad Click Prediction : a View from the Trenches Categories and Subject Descriptors. 2013.

[NT18] M. Neumann and N. Thang Vu. Cross-lingual and Multilingual Speech Emotion Recognition on English and French. *ArXiv e-prints*, March 2018.

[Qia99] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks Volume 12, Issue 1, January 1999, Pages 145-151*, 12(1):145–151, January 1999.

[RSSL13] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (i), 2013.

[TC02] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[Zei12] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[ZWR11] Zixing Zhang, Felix Weninger, and Gerhard Rigoll. Selecting Training Data for Cross-Corpus Speech Emotion Recognition : Prototypicality vs . Generalization. *Conference: Proceedings 2011 Speech Processing Conference*, 2011.