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

H.M. Fayek and M. Lech
School of Engineering
RMIT University
Melbourne, Victoria 3001, Australia
haytham.fayek@ieee.org, margaret.lech@rmit.edu.au

L. Cavedon
School of Science
RMIT University
Melbourne, Victoria 3001, Australia
lawrence.cavedon@rmit.edu.au

***Abstract*—Ground truth labels obtained by averaging or majority voting are commonly used to train automatic emotion classifiers. However, ground truth labels fail to encapsulate inter-annotator variability and ignore the subjectivity of emotions. In this paper, we propose two viable approaches to model the subjectiveness of emotions by incorporating inter-annotator variability, which are soft labels and model ensembling, where each model represents an annotator. Using a deep neural network that recognizes emotions in real-time from one second windows of speech spectrograms, we demonstrate that both approaches lead to consistent improvement over using ground truth labels. It is empirically shown that the performance gain of the ensemble over the baseline model could be achieved using soft labels generated from multiple annotators.**

I. Introduction

Automatic Emotion Recognition (AER) is key in the endeavor to improve human-computer interaction, which has been highlighted by recent interest from academia and industry alike. There are numerous out-of-the-box applications of AER such as emotion monitoring, augmenting human judgment, emotional convergence as well as production [1]. Moreover, the integration of AER into other systems may be of great benefit; for example, the accuracy of an Automatic Speech Recognition (ASR) system can drop significantly when dealing with emotional speech [2], which could be addressed using an integrated AER.

An agreed-upon formal definition of emotions remains elusive to psychologists [3]. In fact, many fundamental issues in emotion theory are still debatable, such as what constitutes an emotion, the dependency of emotions on language and culture, and how can we model emotions [4]. However, it is generally agreed that emotions are subjective, in that several studies indicate that humans understand and perceive emotions varyingly [5].

Emotions are most commonly represented using categorical labels or dimensional descriptors [6]. Either way, these representations tend to completely ignore the subjectivity of emotions and fail to represent mixed emotions by assigning a ground truth label, usually by majority voting or averaging between multiple annotators [7]. By doing so, ground truth labels discard inter-annotator variability, which not only ignores the subjective attribute of emotions but may also lead to the omission of valuable information that could be beneficial in training the models. For instance, if an annotator labeled a particular utterance as *depressed* while the remaining four annotators labeled the same utterance as *neutral*, then a majority vote would completely discard the annotation of *depressed*. It is therefore sensible to incorporate information collected from all annotators, thus yielding a richer representation of the underlying emotions.

One approach to incorporating knowledge collected from all annotators is to train an *ensemble* of models, such that each model effectively represents an annotator, and combine the individual outputs of the ensemble. Another more efficient and arguably more interesting approach is to train only one model using *soft labels*, generated from all annotators. Both approaches are explored in this paper using an AER system that uses speech as its input modality. Our hypothesis is that both approaches should yield similar results.

Most of the reported work on Speech Emotion Recognition (SER) follows one of two pipelines [8]; *turn-based processing* or *frame-based processing*, with the former being more popular recently [9]. Turn-based processing aims at recognizing emotions from a complete utterance. It comprises extraction of many frame-based acoustic parameters, also called Low-Level Descriptors (LLDs) such as fundamental frequency (F_0), Mel-Frequency Cepstral Coefficients (MFCCs) and Teager Energy Operator (TEO), which are then fed into statistical functionals to compute statistics over the entire utterance. A major drawback of turn-based processing is that it loses intra-utterance timing and contour information. A long sentence may contain one or more transitions from one emotion to another, which may not be detected by, and in fact degrade, an SER relying on turn-based processing. Moreover, such amounts of preprocessing may impose limitations on a real-time SER system and hinder its deployment to real-world applications.

We therefore use frame-based processing, which aims at recognizing emotions at the frame level, where several frames

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

are typically concatenated to form a window to make use of utterance dynamics and intra-utterance contour information. The notion of a short-time fixed-length window is essential to achieving a practical real-time SER as opposed to turn-based processing that relies on segmenting utterances by detecting speaker pauses and other features, which may be difficult in a practical noisy environment. Another advantage of using a fixed-length window of speech is the ability to rely on raw speech spectrograms and a deep multi-layered model, avoiding the need to compute many acoustic parameters, statistical functionals and other preprocessing operations [10]. In doing so, we aim to achieve an SER with a simple pipeline and low latency that learns features in an adaptive hierarchical manner.

In this paper, we address the issues highlighted in the previous paragraphs. Concretely, we propose a Deep Neural Network (DNN) that recognizes emotions in real-time directly from a one-second window of raw speech spectrograms. We investigate using soft labels as a richer representation of emotions, as well as using an ensemble of networks to represent multiple annotators. We validate our hypothesis that both approaches yield similar results and compare the resulting models against a baseline model that uses ground truth labels. Results are reported on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [11].

The contributions of this paper are twofold. We propose two viable approaches to modeling the subjectiveness of emotions by incorporating inter-annotator variability, which are soft labels and model ensembling. We empirically demonstrate that the performance gain by using an ensemble of models can be reproduced by a single model trained using soft labels derived directly from ground truth hard labels. The approaches proposed in this paper can be generalized in a straightforward manner to other subjective classification problems.

The remainder of this paper is structured into four sections. In the next section, the relationship between the current work and selected prior work is established. Subsequently, in Section III, the methodology is presented, describing the data, its preprocessing, the DNN architecture, the training recipe used, as well as model ensembling and label encoding techniques. Results are presented and discussed in Section IV. The final section concludes the paper and depicts future work.

II. Relation to prior work

In [12], soft labels obtained by training large models or ensembles of models on ground truth labels were used to train smaller models. The objective of the work was to distill the knowledge in large models into smaller models while maintaining the performance, allowing faster run-time and less resources. A similar idea was pioneered in [13], where model compression techniques were devised to compress large complex ensembles into smaller faster models without much loss in performance. The work presented in [12], [13] motivated the comparison carried out in this paper between model ensembles and genuine soft targets generated by incorporating knowledge from multiple annotators.

In [14], soft labels were used to measure the unanimity of annotators in emotion recognition. It was reported that while ground truth hard labels performed better than soft labels, soft labels had a more similar entropy to human annotators. In [15], the inter-annotator standard deviation was used to model the variability between multiple annotators in a multi-task learning emotion recognition framework.

In [7], a comprehensive treatment of the problem of data prototypicality in emotion recognition was presented. Concretely, data labeled by multiple annotators was used to study the effect of varying the degree of data prototypicality on an emotion classifier's performance. Results reported demonstrated a strong correlation between the classifier's performance and data prototypicality.

Direct comparison with prior SER on the IEMOCAP database is difficult due to differences in data subsets and modalities considered, which is a common problem in the field of AER [16]. We only mention studies that use speech as a modality, since it is the primary focus of our study. In [17], an average recall of 57.39% was reported when speech was used as a modality and a higher average recall was attained by incorporating other modalities. In [18], an unweighted accuracy and a weighted accuracy of 48.5% and 54.3% respectively were reported using speech and a hybrid of a deep neural network and an extreme learning machine. Finally, in [19], an accuracy of 53.99% and average recall of 50.64% were reported when speech was used as the only modality; however, it was shown that taking other modalities into consideration yielded performance improvements.

III. Methodology

A. Data

In this paper, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [11] was used since labels from multiple annotators were available. The database comprised 12 hours of audio-visual recordings divided into five sessions. Each session was composed of two actors, a male and a female, performing emotional scripts as well as improvised scenarios. In total, the database contained 10039 utterances with an average duration of 4.5 seconds. Only audio was used in our experiments. The database predominantly focused on five emotions, namely, *anger*, *happiness*, *sadness*, *neutral* and *frustration*; however, annotators were not limited to these emotions during annotation. To be consistent with other studies on this database [17], [19], [20], we considered only four emotions: *anger*, *happiness*, *sadness* and *neutral*, with *excitement* considered as *happiness*.

Utterances were labeled by three annotators using categorical labels as well as continuous descriptors on the valence, activation and dominance axes. Only categorical labels were used in this work. Each annotator was allowed to choose more than one categorical label if they felt the necessity, or simply select 'other' if the provided labels were not adequate. Ground truths (hard) labels were obtained by majority voting, where 74.6% of the utterances were agreed upon by at least two annotators. Utterances that were labeled differently by all three annotators were excluded from our study, since there would be no ground truth labels. To analyze the inter-annotator agreement, Fleiss' *kappa* statistic was calculated for the entire database and utterances that were agreed upon by at least two annotators; this was found to be $\kappa = 0.35$ and $\kappa = 0.48$ respectively, after considering *fear*, *disgust* and *surprise* as 'other', and *happiness* and *excitement* as *happiness*. For more details, see [11].

B. Preprocessing

The audio was downsampled to 8 KHz and a voice activity detector was applied to remove silent fragments. Subsequently, the audio 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.

A four-fold Leave-One-Speaker-Group-Out (LOSGO) cross-validation scheme [16] was employed in all our experiments using the first four sessions, where each two actors from the same session were considered a fold. The fifth session was used to cross-validate the hyperparameters of our model and to apply early-stopping during training. Therefore it was not included in the cross-validation folds so as not to bias the results [21].

The data was normalized to have zero mean and unit variance. The mean and variance were computed for each fold separately using only the training subset. No speaker dependent operations were performed.

C. Deep Neural Network

A Deep Neural Network (DNN) [22] that recognizes emotions from a one-second window of speech spectrograms was used as a baseline in our study, similar to [10].

The architecture of the DNN had 7 feed-forward fully-connected layers. The input layer's dimensionality was 2624 (64 frames × 41 coefficients per frame). All hidden layers had 1024 Rectified Linear Units (ReLUs) [23]. The output layer was a four-way softmax layer as in (1), producing the posterior class probabilities:

$$y^{(L)} = \frac{\exp(z^{(L)})}{\sum_{k=1}^{K} \exp(z^{(L)})} \tag{1}$$

$$z^{(L)} = y^{(L-1)} W^{(L)} + b^{(L)} \tag{2}$$

where $y^{(L)}$ is a vector of normalized class probabilities, $z^{(L)}$ is the input to layer L computed as in (2), $y^{(L-1)}$ is the output of the final hidden layer, $L = 7$ is the total number of layers, $W^{(L)}$ and $b^{(L)}$ are a matrix of weights and a vector of biases respectively, and $K = 4$ is the number of output classes.

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

$$C = -\sum_{k=1}^{K} Y_k \log(y_k^{(L)}) \tag{3}$$

$$\frac{\partial C}{\partial y^{(L)}} = y^{(L)} - Y \tag{4}$$

where Y is a vector of either hard labels or soft labels, and k denotes the k^{th} class.

Mini-batch stochastic gradient descent with a batch size of 128 and RMSProp [24], [25] per-parameter adaptive learning rate were used to optimize the DNN parameters. The base learning rate was set to 2×10^{-4} and a decay of 0.99.

The DNN was regularized using dropout [26] with a retention probability of 0.5 applied to all hidden layers and a retention probability of 0.9 applied to the input layer. A max-norm constraint was imposed on all layers, such that the l_2 norm of the incoming weight vector w of each layer was constrained to have an upper bound, $\|w\|_2 \leq c$, where c was chosen to be 3. If the constraint was violated, w was projected back onto a circle of radius c.

The parameters of each layer in the DNN were initialized from a Gaussian distribution with zero mean and $\sqrt{2/n}$ standard deviation, where n is the number of inputs to the layer, as recommended in [27].

The model architecture and all hyperparameters were tuned to achieve the best average recall using the validation set (the excluded fifth session). The same session was also used to perform early stopping during training, such that the training halts if the average recall ceases to improve in 20 epochs and the model with the best average recall on the validation set was returned.

Training was carried out using an NVIDIA Tesla K40 Graphics Processing Unit (GPU) to accelerate the process.

D. Model Ensembling

To model inter-annotator variability, an ensemble of DNNs was trained such that each DNN represented an annotator. Concretely, three DNNs were trained, each using ground truth hard labels from only one annotator, and their individual outputs were combined.

Two ensemble combination rules were explored and are described. The first rule was the geometric mean of the posterior class probabilities as in (5):

$$y_k = (\prod_{n=1}^{N} y_{n,k})^{1/N} \tag{5}$$

where y_k is the geometric mean of the k^{th} class probabilities, $y_{n,k}$ is the k^{th} class probability of the n^{th} DNN, and $N = 3$ is the total number of DNNs in the ensemble.

The second rule was an unweighted majority vote as in (6), which is similar to the rule that was used to generate the original ground truth labels in the IEMOCAP database:

$$t = \underset{k=1,\cdots,K}{\operatorname{argmax}} \sum_{n=1}^{N} d_{n,k} \tag{6}$$

where t is the output class, $d_{n,k} \in \{0, 1\}$ is the decision of the n^{th} DNN in the ensemble on the k^{th} class.

TABLE I
LABEL ENCODING SCHEME BASED ON THREE ANNOTATORS

Annotation[a]	Hard Label[b]	Soft Label[b]
[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]	$[0, 0, 0, 1]$	$[0, 0, 0.25, 0.75]$

[a] In the form of: [Annotator 1][Annotator 2][Annotator 3].
[b] In the form of: [anger, happiness, neutral, sadness].

E. Label Encoding

Using the one-of-K (one-hot) encoding scheme [28], soft labels were generated from multiple one-of-K encoded labels as in (7):

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

where s is a K-dimensional vector of soft labels, $h^{(n)}$ is a K-dimensional vector of one-of-K hard labels encoded from the n^{th} annotator, and N is the number of annotators. Table I illustrates several annotation examples from the IEMOCAP database labeled by three annotators and their corresponding labels to alleviate ambiguity.

IV. RESULTS AND DISCUSSION

To demonstrate the effect of data prototypicality on the performance of the proposed SER, two types of test sets were used in our experiments. The first is a prototypical set, where all annotators were in complete agreement over the label; the second set is the full test set which includes prototypical and non-prototypical cases. The training set in both cases was identical containing prototypical and non-prototypical data. Note that only the training labels were different (i.e. hard labels or soft labels). In both cases the output of the classifier during evaluation, was the class with the highest posterior probability computed from (1).

As is standard practice in the field of AER [16], results are reported using the unweighted average recall and average F-score to reflect imbalanced classes. These metrics were averaged over the four-fold LOSGO cross-validation scheme described in Section III-A.

Using the DNN described in Section III-C, the baseline model which was trained using ground truth hard labels was evaluated. Table II depicts the average recall and F-score of the baseline model on the prototypical and full test sets.

In agreement with [7], [20], it is conspicuous that data prototypicality has a significant effect on the classifier's performance. The presence of non-prototypical data led to a significant degradation in the evaluation metrics, which highlights the difficultly of correctly classifying such data. Nevertheless, given that the only input is a one-second window of speech spectrograms as opposed to a complete utterance, the performance on both test sets is very promising and compares favorably with previous results on the same database.

TABLE II
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 III
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%

Next, using an ensemble of the same DNNs and the same training recipe for each DNN, the performance of the ensemble of three models, each representing an annotator (trained using ground truth labels from one annotator), was evaluated. Table III demonstrates the average recall and F-score of the ensemble on the prototypical and full test sets using the geometric mean and majority voting rules.

It can be observed that the ensemble had a slightly lower F-score compared to the baseline model in Table II on the prototypical set, but that is of little interest since emotions are rarely prototypical [5]. On the other hand, a notable improvement in the average recall and F-score is evident on the full test set compared to the baseline model. A relative improvement of 4.01% and 3.88% in average recall and F-score respectively over the baseline was achieved using the ensemble with the geometric mean rule and similarly a relative improvement of 2.92% and 3.09% in average recall and F-score respectively using the majority voting rule. This suggests that useful information was obtained by incorporating knowledge from all annotators. However, this improvement was at the expense of training and evaluating N models, where N is the number of annotators and was equal to 3 in our experiments.

Finally, using the same DNN architecture and training recipe, one model was trained with soft labels generated from all three annotators as explained in Section III-E. Table IV shows the average recall and F-score of the model on the prototypical and full test sets.

As expected, the model trained with soft labels outperformed the baseline model with a relative improvement of 3.28% and 2.89% in average recall and F-score respectively on the full test set. More interestingly, the model trained with soft labels performed very similarly to the ensembles in Table III, almost retaining the performance gain by the ensemble over the baseline model without the need for training and evaluating N models. This confirms that by using soft labels instead of ground truth hard labels, we were able to incorporate more

TABLE IV
TEST PERFORMANCE USING SOFT LABELS

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

knowledge in the labels and obtain a richer representation of the underlying emotions, leading to an improvement in the classification performance.

The methodology of modeling subjectiveness presented in this paper is not limited to emotion recognition or the neural network architecture used in this paper, but may be extended in a straightforward manner to other subjective classification tasks and using other neural network architectures.

V. CONCLUSION AND FUTURE WORK

In this paper, a deep neural network that recognizes emotions in real-time from one-second windows of speech spectrograms was presented, demonstrating promising results on the IEMOCAP database. It was shown that data prototypicality has a significant effect on the classifier's performance.

In order to incorporate knowledge from multiple annotators to reflect the subjectiveness of emotions, two approaches were investigated: using soft labels, and using an ensemble of models such that each model in the ensemble represented an annotator. Both approaches outperformed the baseline model trained 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.

Future work comprises extension of the proposed methodology to a larger range of subjective classification problems that would harness the use of soft labels.

ACKNOWLEDGMENT

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

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[2] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, School of Architecture and Planning, Massachusetts Institute of Technology, 2004.

[3] P. Kleinginna Jr and A. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and Emotion*, vol. 5, no. 4, pp. 345–379, 1981.

[4] R. Picard, "Emotion research by the people, for the people," *Emotion Review*, 2010.

[5] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion knowledge: Further exploration of a prototype approach," *Journal of Personality and Social Psychology*, vol. 52, no. 6, pp. 1061–1086, 1987.

[6] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[7] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, prototypes, performance: classifying emotional user states." in *INTERSPEECH*, 2008, pp. 601–604.

[8] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[9] F. Eyben, M. Wllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 7–19, 2010.

[10] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on*, December 2015.

[11] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[13] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 535–541.

[14] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, "of all things the measure is man : Automatic classification of emotions and interlabeler consistency," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, March 2005, pp. 317–320.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 1, pp. 6:1–6:29, Mar. 2012.

[16] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[17] M. Shah, C. Chakrabarti, and A. Spanias, "A multi-modal approach to emotion recognition using undirected topic models," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, June 2014, pp. 754–757.

[18] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH 2014*, 2014.

[19] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April 2013.

[20] Y. Kim, H. Lee, and E. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 3687–3691.

[21] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[22] A.-R. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[23] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean *et al.*, "On rectified linear units for speech processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3517–3521.

[24] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, 2012.

[25] Y. Dauphin, H. de Vries, J. Chung, and Y. Bengio, "RMSProp and equilibrated adaptive learning rates for non-convex optimization," *arXiv preprint arXiv:1502.04390*, 2015.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv preprint arXiv:1502.01852*, 2015.

[28] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.