An Automatic Emotion Recognition System for Laughter – Team LOL

Sam Holden, Yuxin Zeng, Katie Stuart, Laura Weihl, and Parham Oghabi

1. Introduction

A practical reason for building a system capable of automatically interpreting the emotional content of laughter is to help a robot assistant understand all forms of communication that a human being might use around it. For instance, a laugh might signify that a person is happy, and that the robot is doing a good job attending to the person's needs. However, the laugh could just as well mean that the robot has made a surprising mistake and is being ridiculed for how poorly it is assisting the person. Emotional content of the laughter could be used to differentiate these two situations. The robot assistant imagined for the purposes of this paper consists of software installed on a person's computer and helps to motivate some of the choices made in subsequent sections. With minimal adjustments to data inputs, the confines of the computer can be abandoned. It is with the goal of helping robot assistants derive inference about emotional content in laughter that we have undertaken this project.

2. Methods

2.1 Data Collection

Data was collected for four participants, each of which supplied approximately 10 clean laughs. Eight tasks were performed in the hopes of inducing a variety of laughter. Tongue twisters were used to elicit embarrassment, baby laughter and silly spelling videos for happiness and videos featuring people falling down for schadenfreude (defined as pleasure from another's misfortune [6]). Between each video, the participant was asked to act out various posed laughs.

It is important to make participants feel as comfortable as possible if laugh elicitation is to be successful [7], so they were brought into the experimentation room in pairs and participated in the tasks alongside the experimenters. Only clips of the study participants, i.e. not the experimenters, were used to avoid biasing results. Supplementing the self-collected data, the MAHNOB dataset [1] is used. Full dataset details can be found in table 1.

Table 1. Details of dataset used to model laughter Automatic Emotion Recognition System.

Dataset Details	
Laugh Count	614
% Female	54.7%
Average Clip Length (s)	1.78
Variance of Clip Length (s)	6.72
Happy Laughter	288
Courtesy Laughter	136
Embarrassment Laughter	63
Schadenfreude Laughter	127

2.2 Labelling Process

The labeling process was different for the gathered data than it was for the data from the MAHNOB dataset. The MAHNOB dataset was first clipped into segments using the annotations supplied. The software ffmpeg [8] was employed for this purpose and short clips were created, varying from under a second in length to over seven seconds. Each clip consisted of a single laugh instance. These short clips, containing video and audio from the start of a laugh to the end exclusively, i.e. no contextual information, were each viewed by two volunteers who assigned labels to them. Three labels were affixed: arousal, dominance and emotional content. The arousal and dominance labels were chosen from a six-point scale ranging from zero to five, where higher values mean more arousal and more dominance, respectively. The affective state label was chosen from four categories: happy, embarrassment, schadenfreude and courtesy. These emotions were derived from different basic emotions [4], modified to be appropriate for laughter. Happy and Embarrassment are, therefore, part of Izard's list of basic emotions. Schadenfreude was thought to be the type of laughter likely to be seen with the basic emotion contempt [4]. Taunting laughter, which is more uniformly consistent with contempt [3], was thought to be unlikely in our dataset given that subjects were not being put in a competitive setting. Courtesy laughter was added because it too was likely to be seen both in the wild and during laughter inducement experiments.

For the gathered data, labels used were identical to what is above and two independent volunteers were used. Self-reported labels were also gathered in between tasks, adding the perspective of the subjects themselves. An averaging system for the labels provided by these two labelers is discussed in appendix A.

2.3 Discriminative Features – Audio

The audio signal related to our laughter was processed into Mel Frequency Cepstral Coefficients (MFCC) [9]. MFCCs are presented in the form of frequency quantity over time, where each filter relates to how much of a given frequency is present in a given signal time frame. A gentle introduction to the process is provided in figure 1.

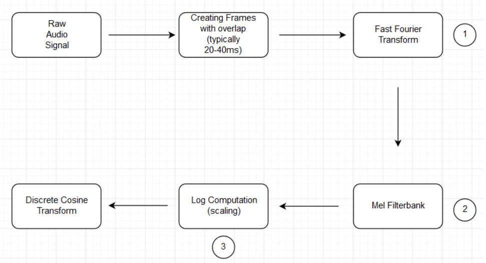
Raw Audio Signal
Creating Frames with overlap (typically 20-40ms)
Fast Fourier Transform
1
Mel Filterbank
2
Log Computation (scaling)
3
Discrete Cosine Transform

Figure 1: Inner workings of the creation of Mel Frequency Cepstral Coefficients from raw input. For clarity, point 1 can intuitively be thought of as separating the different frequencies present in a sound. Point 2 can be understood as grouping like-frequencies.

All this pre-processing results in a dataframe that is 614 observations by 216 time steps by 20 transformed Mel filter banks. Static features like gender and length of clip are then used to add levels to the detection of affective state through auditory signal. The idea is to use gender as an imperfect baselining tool when associating prevalence of certain frequencies to emotional states. One might expect the distributions of frequencies for men and women to differ naturally.

Length of laughter episode is also added, as this information was removed from the sequence data in the creation of the MFCCs. One could easily imagine this feature being relevant for determining arousal or for differentiating emotional content, so it was added back as a static variable.

2.4 Discriminative Features – Visual

From each of the 614 laughter clips, we sample 25 frames in a way which ensures they are evenly spread across the length of the clip. We extract three datasets: two containing all gray-scale and coloured frames and one containing 68 x,y-coordinates of facial landmarks. The library dlib and opencv was used to detect a bounding box around the face and to crop the face from the image with a 20 pixel increase of the border in x and y direction on both sides. A total of 68 facial points (x, y-coordinates within the frame) are extracted with 19 points around the mouth region and 9 and 11 points around the eye and eyebrow regions respectively. An example is shown in figure 2 [20]. We capture these facial features and feed them into our model directly, unlike the functioning of the FACS model, which encodes movements of individual facial muscles. We do not make any distance calculations explicitly, preferring to let the model discover these features from data. FACS encoding is highly subjective and therefore very prone to bias, after all. [19].

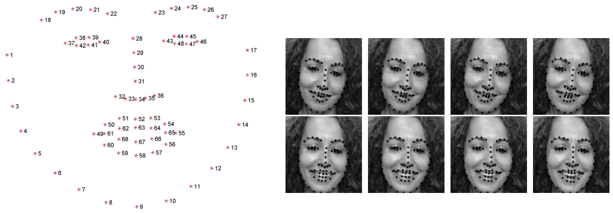

Figure 2: dlib facial feature points (from 1 to 68) and sample frames from the database wih the extracted facial points. [20]

3. Models

3.1 Neural Nets Used

To obtain predictions for emotional content, arousal and dominance from each of the modalities independently, we explored using both an LSTM and Convolutional LSTM (ConvLSTM) model. These are both state-of-the-art methods to deal with sequence data. Since the audio and image features both hold sequential information, it's very important that this context is captured by the model to make better predictions. An LSTM network consists of individual units. Each unit receives an individual audio window/frame from the sequence. This unit stores relevant information and controls how much information passes to the next state [15]. This continues through the time steps before reaching a flattened layer where all this memory from the sequence is used to make a final prediction. The model we used is many-to-one, since we have a sequence input but are making only one prediction. This can be seen in Figure 3. The function of the convolutional architecture is explained in appendix B.

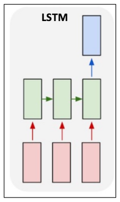

Figure 3: Many-to-one LSTM Architecture that takes a sequence as input and outputs a prediction. [21]

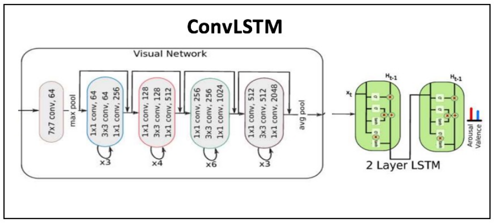

Figure 4: Convolution LSTM Architecture. [22]

3.2 Optimization Process

- **Model Selection**

To choose the best model, we compared performance of ConvLSTM and simple LSTM. Recall results suggest that ConvLSTM improved model performance (recall: 0.59 vs. 0.54). When assessing the confusion matrices, which are shown in tables 2 and 3, the simpler LSTM has better accuracy across all 4 emotions compared to ConvLSTM who's main success is for happy laughter. Overall F1-scores show that simple LSTM is the better model, driven by a better precision (0.53 vs. 0.55). While happy is already the easier emotion to predict, extracting features using ConvLSTM amplified this and so it didn't perform so well across the other emotions.

Table 2. Normalized Confusion Matrix of Emotion using ConvLSTM (with Facial points)

	Courtesy	Embarrassed	Happy	Schadenfreude
Courtesy	0.57	0.40	0.13	0.18
Embarrassed	0.00	0.00	0.00	0.00
Happy	0.43	0.60	0.85	0.65
Schadenfreude	0.00	0.00	0.02	0.18

Table 3. Normalized Confusion Matrix of Emotion using LSTM (with Facial points)

	Courtesy	Embarrassed	Happy	Schadenfreude
Courtesy	0.37	0.00	0.15	0.12
Embarrassed	0.23	0.40	0.02	0.00
Happy	0.33	0.53	0.61	0.18
Schadenfreude	0.07	0.07	0.23	0.71

- **Multistream LSTM**

Studies have proven that temporal dynamics play an important role in interpreting types of smile [5] and these findings can be applied to how a person laughs. One example is in courtesy vs. spontaneous laughter. A spontaneous laugh is one that builds up over a longer period compared to a courtesy laugh, which is quicker and exaggerates an emotion [5]. By taking samples of the audio and video to even their lengths we may have lost some of this important temporal information that would help in classifying these laughter types.

To overcome this we introduced a multi-stream architecture into our model. Alongside Audio and Image data, we used a static variable representing the length of each laugh to make up for some of this lost data. This improved the Audio and Image models for the 4 emotions and for Arousal. In each case the overall prediction accuracy increases by 2% (the confusion matrices for the emotions for the Audio models are shown in tables 4 and 5). With the addition of the static feature, Courtesy and Happy laughter misclassification rates decrease.

Table 4. Normalized Confusion Matrix of Emotion using LSTM (without static feature – only MFCC)

	Courtesy	Embarrassed	Happy	Schadenfreude
Courtesy	0.27	0.04	0.10	0.00
Embarrassed	0.00	0.00	0.02	0.00
Happy	0.71	0.87	0.86	0.98
Schadenfreude	0.02	0.09	0.01	0.02

Table 5. Normalized Confusion Matrix of Emotion using LSTM (with static feature and MFCC)

	Courtesy	Embarrassed	Happy	Schadenfreude
Courtesy	0.43	0.35	0.12	0.12
Embarrassed	0.00	0.00	0.00	0.00
Happy	0.53	0.57	0.80	0.78
Schadenfreude	0.04	0.09	0.07	0.10

- **Regularization Techniques**

Several regularization techniques were used for this experiment including Adam optimizer, Dropout, Batch Normalization and Early Stopping. A description of these can be found in Appendix C.

- **Fusion of Multiple Modalities**

Multiple fusion methods were attempted such as early-fusion and late fusion. The audio and video features were not concatenated in the beginning since the sampling rate of the video (25 frames per second) and audio (44KHz) varied by a large amount and concatenating them would not be reasonable. Early-fusion was attempted, where the outputs of the second LSTM layer in each of the video and audio model were concatenated before being fed into a final dense layer. The model was then trained using the graph described. However, the results were not as good as the late fusion model. The complete architecture of the late fusion weighted ensemble model is shown in Figure 5.

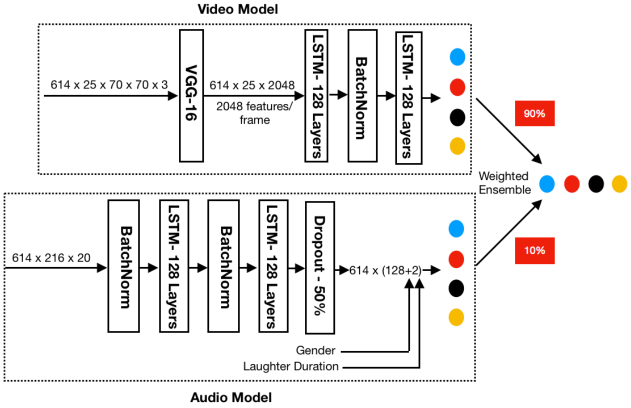

Figure 5: The ensemble model architecture of our best performing model.

The audio and video model were trained separately and the probability of their predictions were averaged. For the video model, a VGG-16 network was used to extract the 2048 features from each video frame. The VGG-16 network was pre-trained on colored images from the ImageNet dataset. [18] The video frames were cropped as a pre-processing step prior to feature extraction since some experiments contained faces of other people in the background and this would skew the results. Cropping the video frames increased the accuracy of the video model from 58% to 60% in predicting emotional content. The features for each video frame were then input into a 128-unit LSTM network, which contained batch normalization as well. Since the VGG-16 network was pre-trained on a large dataset, it was able to effectively extract abstract features from the frames in a way that the non-pre-trained models could not.

For the audio model, the MFCC features were input into a 128-unit LSTM network which also contained batch normalization and dropout layers for regularization. Multi-streaming was also implemented in this model by concatenating the subject's gender and the duration of their laughter as a final step with the extracted LSTM features before inputting them into the softmax classifier. Grid search was performed in order to find the optimal weights for the video model and audio model prediction probabilities. As expected, the video model was given a higher weight (90%) since it was more effective in discriminating between the laughters.

4. Results

A benchmark was created against which the success of our model could be compared for arousal and dominance. This was necessary because mean squared error (MSE) does not have intuitive meaning like accuracy or information in a confusion matrix does. The mean of the test set was chosen, meaning that MSE for this benchmark is simply the variance of the test set. Similarly, for the emotion recognition (laughter classification) task, a random model was chosen as the baseline with an accuracy of 25%, since there are four classes.

4.1 Early Models

The earliest models created in this project and their results can be seen in tables 6, 7 and 8. These early models consisted of using separate audio and video models to predict the arousal, dominance, and laughter types to evaluate the importance of each feature.

Table 6. Results of arousal/dominance prediction using pixelized video frames.

	Prediction Mean	MSE
Without Static Variables		
Arousal	1.80	1.04
Dominance	2.30	0.65
With Static Variables (Gender and Laughter Duration)		
Arousal	1.77	0.91
Dominance	2.40	0.68
Benchmark		
Arousal	1.86	1.29
Dominance	2.31	0.90

Table 7. Confusion matrix of emotion prediction by considering pixelized video frames (with static features).

	Courtesy	Embarrassed	Happy	Schadenfreude
Courtesy	5	1	32	11
Embarrassed	6	1	14	2
Happy	5	1	81	9
Schadenfreude	1	0	33	6

For the video models, the full video frames were used as input first without cropping or facial landmarks. For the audio model, MFCC features were used both with and without the static features (gender and laughter length) to compare the performance.

Table 8. Results of arousal/dominance prediction using MFCC features with and without static features

	Prediction Mean	MSE
Without Static Variables		
Arousal	1.82	1.22
Dominance	2.31	0.78
With Static Variables (Gender and Laughter Duration)		
Arousal	1.86	0.94
Dominance	2.24	0.89
Benchmark		
Arousal	1.86	1.29
Dominance	2.31	0.90

We can see that both audio and video features seem to hold some information, with video clearly doing better than audio. This was a reassuring first step, given the difficulty humans have with this task [3] and the limited data being trained on. These early results also showed exactly where to look next. Static features show an improvement when predicting arousal in both models, and although results got worse for predicting dominance, this was in line with an overfitting problem that was a constant throughout the experiments. It was taken for granted moving forward that the static variables held discriminative information for our prediction tasks. All models were tested on a hold-out set which was not seen during training.

4.2 Using Face Points

The next step we took was to combat overfitting: facial features were used for modeling instead of the whole video frames. This was theorized to help narrow the focus of the flexible models being used. Results were a big improvement as seen in tables 9 and 10. Both arousal and dominance were predicted with significance well below alpha of 0.05 significance level and the accuracy on affective state rises to 52 %.

Table 9. Results of arousal/dominance prediction using face points of the video frames and including static features

	Prediction Mean	MSE
With Static Variables (Gender and Laughter Duration)		
Arousal	1.88	0.70
Dominance	2.44	0.62
Benchmark		
Arousal	1.86	1.29
Dominance	2.31	0.90

Table 10. Confusion matrix of emotion prediction using face points.

	Courtesy	Embarrassed	Happy	Schadenfreude
Courtesy	26	10	14	4
Embarrassed	0	0	0	0
Happy	22	13	80	35
Schadenfreude	1	0	2	2

The confusion matrices shown thus far have both shown a class imbalance problem that was persistent throughout the experiments. It was decided that the distribution of laughter was likely similarly (although not exactly) imbalanced in the real world so it should not be rectified.

The probabilities of each affective state could be used by a robot assistant rather than a hard assignment, so this class imbalance problem exists more in this report than it does in real application. Interestingly, the perceived class imbalance may be a reflection of the prediction problem as well. Schadenfreude is overwhelmingly being predicted as happy (in a way that is not happening to courtesy, which has similar class weighting). Schadenfreude is a mix of one positive basic emotion, like happy, and one negative, like contempt [3]. So it is in line with human recognition that our model confuses these two affective states in laughter.

4.3 Late Fusion (Combining Early Models)

The fusion results are best for the late-fusion model, mainly because of an overfitting problem in the early-fusion one. The dataset used is simply not big enough to avoid these issues. Results are presented in tables 11 and 12 and are slightly better in all respects than the single-modality models.

Table 11. Confusion matrix of affection prediction by late fusion model

	Courtesy	Embarrassed	Happy	Schadenfreude
Courtesy	23	8	12	2
Embarrassed	0	0	0	0
Happy	26	13	83	38
Schadenfreude	0	2	1	1

Table 12. Results of arousal/dominance prediction using late fusion of modalities and static features

	MSE
With Static Variables (Gender and Laughter Duration)	
Arousal	0.68
Dominance	0.62
Benchmark	
Arousal	1.29
Dominance	0.90

4.4 Late Fusion Leveraging Transfer Learning (Best Model)

Another late fusion model was also implemented which performed the best for affective state prediction. The architecture was shown in Figure 5. VGG-16, a model trained on the Imagenet dataset, was very effective in extracting features from the frames and the LSTMs exploited these features. The ensemble model had an accuracy of 62% and the results are in tables 13 to 15. The previous late fusion model had a recall of 0 for the “embarrassment” laughter but this model achieves a recall of 33%. The size of the validation set was 20% of the training set as opposed to the 33% chosen for the previous models. However, the model was also tested with a validation size of 33% and outperformed the previous models, achieving an accuracy of 58%.

Table 13. Confusion matrix of emotion prediction for our best model as shown in figure 5; validation set size was 20%.

Accuracy=62%	Courtesy	Embarrassed	Happy	Schadenfreude
Courtesy	13	1	5	1
Embarrassed	0	5	0	1
Happy	16	7	50	7
Schadenfreude	1	2	6	8

Table 14. Precision and Recall for each of the classes

	Precision	Recall
Courtesy	65%	43%
Embarrassed	83%	33%
Happy	62%	82%
Schadenfreude	47%	47%

Table 15. Results of arousal/dominance prediction

	Prediction Mean	MSE
With Static Variables (Gender and Laughter Duration)		
Arousal	1.99	0.76
Dominance	2.33	0.53
Benchmark		
Arousal	1.87	1.32
Dominance	2.39	0.74

In conclusion, some preliminary success in the task of automatic prediction of affective state from laughter can be seen. A model leveraging learned abstract features from the Imagenet dataset was particularly successful. This is a promising result given some strong dataset constraints and should serve as a starting point for further research.

5. Appendix

5.1 Appendix A

As discussed, we used two labelers for our data. We knew from [3] that humans were imperfect emotion recognizers when it comes to laughter: see table 16. We developed an averaging system for labels coming from our 2 labelers (and self-report for self-gathered data) to get the best ground truth possible. This was straightforward when it came to the numerical labels for arousal and dominance. It was less straightforward for emotion. Emotions on Izard's list of basic emotions [4] were averaged using equal weighted sampling. This was the case for 'happy' and 'embarrassment'. 'Schadenfreude', having a mix of contempt and joy [3], was always chosen when averaged with another emotion because we wanted to acknowledge that one of the labelers had noticed contempt in the laughter. A tie involving courtesy was always labeled the other emotion. This is because courteous laughter does not hold emotional content per se. It is a more complex social interaction and is the result of a perceived anticipation of a response. We felt that weight should be given to the emotion perceived.

5.2 Appendix B

Convolutional Neural Networks (CNNs) are often used to extract features from image data (Figure 4). The CNN architecture takes advantage of the 2D structure of an input image by learning "kernels", which are iteratively moved across the image's axes to create abstract representations of the image that might encompass local features. This movement is what is known as convolution.

The number of these features is a hyper-parameter of the model, and so an arbitrary number of these abstract image representations can be created. Given the image data is a sequence, these features were then fed into a sequence model after being processed by the CNN. This is what is known colloquially as a ConvLSTM [10]. ConvLSTMs have led to state-of-the-art results in human activity recognition and part of the motivation for trying it here was the resemblance between human activity recognition's local temporal features and those present in facial features over time [10]. Because the audio features did not have obvious need for abstraction of features, the ConvLSTM was used for video exclusively.

5.3 Appendix C

Adam optimizer is an adaptive gradient algorithm with a per-parameter learning rate, rather than a fixed single learning rate as with SGD, therefore using an Adam optimizer rather than SGD is crucial in ensuring a good accuracy was met [16]. Such design choice ensures quicker convergence and avoids parameters bouncing chaotically in the optimization process by taking huge steps, and therefore get stuck in a swinging stage from one side of the gradient landscape to another [17].

Dropout technique was used to reduce the model over-fitting to the training data by ignoring randomly selected neurons in the model architecture. The purpose of adding dropout is to inhibit one neuron from being too influential compared to the others and force other neurons in the network to 'learn' on their own. A dropout value of 50% of neurons was optimal for this model. Anything lower had a minimal effect while a higher dropout drops too much information and the model under fits [13].

Batch normalization was also used alongside with dropout to prevent over-fitting. Since the LSTM (and in particular ConvLSTM) are deep complex models, they suffer from inviable activation functions throughout training (for ex. Relu activations can stop learning completely). Batch normalization regulates the values at each layer entering these functions and makes sure they stay viable and the model learns optimally [12].

Early Stopping was attempted for the purpose of this experiment, which provides a threshold on how many iterations the model should run for before the generalization error increases [14]. In most networks, early stopping is a useful regularization technique. However, due to the sequential nature of the data of LSTMs the model is more complex and can take a lot longer to learn, the model can often get stuck at a learning rate for many iterations before the accuracy improves. The LSTM did not benefit from Early Stopping, the full potential of the model just became restricted.

Table 16. Confusion Matrix from [3] showing the human ability to distinguish affective state from laughter.

	Joy	Tickle	Schadenfreude	Taunt
Joy	44%	20%	28%	8%
Tickle	26%	45%	25%	5%
Schadenfreude	30%	17%	37%	16%
Taunt	14%	7%	30%	50%

6. REFERENCES

[1] Petridis, Stavros, Brais Martinez, and Maja Pantic. "The MAHNOB laughter database." Image and Vision Computing 31.2 (2013): 186-202.

[2] Ekman, Paul Ed, and Richard J. Davidson. The nature of emotion: Fundamental questions. Oxford University Press, 1994.

[3] Szameitat, Diana P., et al. "Differentiation of emotions in laughter at the behavioral level." Emotion 9.3 (2009): 397.

[4] Izard, Carroll E. "Basic emotions, relations among emotions, and emotion-cognition relations." (1992): 561.

[5] M. F. Valstar, H. Gunes, M. Pantic. "How to Distinguish Posed from Spontaneous Smiles using Geometric Features." Proc. of ACM ICMI, 2007.

[6] Dictionary, Oxford English. "OED online." (2018).

[7] McKeown, Gary, et al. "Laughter induction techniques suitable for generating motion capture data of laughter associated body movements." Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 2013.

[8] Tomar, Suramya. "Converting video formats with FFmpeg." Linux Journal 2006.146 (2006): 10.

[9] Zheng, Fang, Guoliang Zhang, and Zhanjiang Song. "Comparison of different implementations of MFCC." Journal of Computer Science and Technology 16.6 (2001): 582-589.

[10] Guan, Yu, and Thomas Plötz. "Ensembles of deep lstm learners for activity recognition using wearables." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1.2 (2017): 11.

[11] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[12] Medium.com (2017) Glossary of Deep Learning: Batch Normalisation. Available at: https://medium.com/deeper-learning/glossary-of-deep-learning-batch-normalisation-8266dcd2fa82. [Accessed 5 Apr. 2018].

[13] machinelearningmastery.com. (2016). Dropout Regularization in Deep Learning Models With Keras. Available at:https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/. [Accessed 5 Apr. 2018].

[14] Medium.com (2016) Regularization in Deep Learning. [online] Available at: https://chatbotslife.com/regularization-in-deep-learning-f649a45d6e0. [Accessed 5 Apr. 2018].

[15] Colah.github.io (2015) Understanding LSTM Networks. [online] Available at: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ [Accessed 5 Apr. 2018].

[16] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[17] Cs231n.github.io. (2018). CS231n Convolutional Neural Networks for Visual Recognition. [online] Available at: http://cs231n.github.io/neural-networks-3/ [Accessed 5 Apr. 2018].

[18] Karen Simonyan, Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. 2014.

[19] Hamm, Jihun et al. “Automated Facial Action Coding System for Dynamic Analysis of Facial Expressions in Neuropsychiatric Disorders.” Journal of Neuroscience Methods 200.2 (2011): 237–256. PMC. Web. 13 Apr. 2018.

[20] Davis E. King. Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research 10, pp. 1755-1758, 2009.

[21] Karpathy.github.io (2015) The Unreasonable Effectiveness of Recurrent Neural Networks. [online] Available at: http://karpathy.github.io/2015/05/21/rnn-effectiveness/ [Accessed 17 Apr. 2018]

[22] Tzirakis, Panagiotis et al. “End-to-End Multimodal Emotion Recognition using Deep Neural Networks.” arXiv preprint arXiv: 1704.08619v1 (2017)