Predicting Human Emotions using EEG-based Brain computer Interface and Interpretable Machine Learning

Tommaso Colafiglio[1,2], Paolo Sorino[1], Angela Lombardi[1], Domenico Lofù[1] and Tommaso Di Noia[1]

[1]Dept. of Electrical and Information Engineering (DEI), Politecnico di Bari, Bari (Italy)
[2]Dept. of Computer, Automatic and Management Engineering (DIAG), Sapienza Università di Roma, Roma (Italy)

Abstract
EEG-based brain-computer interface (BCI) devices have proved to be powerful tools for predicting human emotions. Although Deep learning (DL) techniques have been extensively used to build emotion recognition architectures using EEG-based BCI, they lack interpretability. We propose a prototype of an EEG-based emotion recognition system that can detect the user's emotional state using a deep learning model embedded into an interpretable framework to analyze the decisions of the model and the contributions of the features. The proposed model achieves high performance while showing relevant information on the impact of frequency and spatial features used to predict the emotional states.

Keywords
Brain-computer interface, Artificial Intelligence, Emotion Recognition, Interpretable AI, Explainable AI

1. Introduction

The research field of affective computing has achieved remarkable results enabling the integration of emotion recognition algorithms in different clinical settings. On the one hand, the availability of increasingly low-cost device, and considerable advances in artificial intelligence algorithms have triggered the rapid development of applications for emotional recognition via brain waves. In particular, EEG-based brain-computer interface (BCI) devices have proved to be remarkably powerful tools for brainwave acquisition, both due to their rapid deployment and their wide application in different scenarios and contexts.

EEG signals are primarily used to diagnose and treat various brain disorders, including epilepsy, tremor, concussions, strokes, and sleep disorders. Machine learning (ML) as an analysis method has been used in recent EEG applications. ML Methods for automated EEG analysis have attracted great interest, especially in clinical diagnostics. For example, ML enables the automation of the process of EEG-based sleep stages [1] and neurological diagnosis of specific diseases such as Alzheimer's disease [2], autism spectrum disorders [3], depression [4], or general EEG pathology[5, 6]. Several factors contribute to the interest in automatic clinical EEG diagnosis. According to physiological studies, the cerebral cortex is the primary controller of humans' higher emotional cognitive capabilities. Hence, it would be advantageous to identify brain areas that are strongly associated with emotions using EEG-based emotion detection, mainly in clinical trials for neuromotor rehabilitation or psychological therapies.

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
tommaso.colafiglio@poliba.it (T. Colafiglio); paolo.sorino@poliba.it (P. Sorino); angela.lombardi@poliba.it (A. Lombardi); domenico.lofu@exprivia.com (D. Lofù); tommaso.dinoia@poliba.it (T. Di Noia)
0000-0001-7184-310X (T. Colafiglio); 0000-0002-9081-2648 (P. Sorino); 0000-0003-1815-9522 (A. Lombardi); 0000-0001-6413-9886 (D. Lofù); 0000-0002-0939-5462 (T. Di Noia)

2. Modeling the emotions

Basically, two techniques have been used to describe emotions: the discrete emotion model and the dimensional emotion model. Dimensional models classify emotions on the scale or dimensions, while individual emotional models include multiple major emotions and have two types of emotion: Positive and Negative Emotions. Several theorists have conducted experiments to identify basic emotions and offered a number of models which can be distinguished from one another.

The most common use is for Russell's 2D emotion model [7]. As is clearly shown in Figure 1, the vertical axis represents arousal dimension and expresses intensity of experience ranging from low to excitement, while the horizontal axis shows valence dimension representing the degree of joy or happiness between negative and positive. In the arousal-valence coordinate system, there are four categories of emotions. On the left hand side of the diagram, negative emotions are visible and positive emotions are shown to the right. The valence axis is represented by positive and negative emotions, while the arousal axis varies from inactive to active emotion. The three-dimensional version of the model also includes the dominance dimension which corresponds to the strength

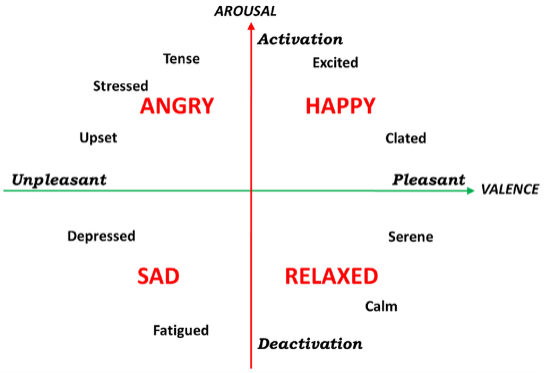


Figure 1: Representation of the Russel's Emotion Model

of the emotion. In contrast to discrete representations, dimensional models attempt to describe emotions using continuous values of their defining properties, which are frequently represented on axes [8].

3. The Role of Machine Learning

The design of a machine learning emotion recognition model requires the careful planning of several steps, i.e., collecting EEG data, preprocessing phase, retrieving features, choosing or reducing them, and classifying emotions. The selection of the best classification tool which is capable of accurately classifying individual emotions has been one of the most important elements in developing an effective emotion classification system. Most existing techniques consider emotion recognition as a problem of classification and attempt to identify between category emotions or across different areas in Russell's 2D emotional model. Several ML models have been developed in recent years to handle the categorization of EEG data for human emotion identification. Among these approaches are the commonly used classification methods Support Vector Machines (SVM), Nave Bayes (NB), k-nearest neighbour (K-NN), Decision Trees (DT), Random Forest (RF), and Artificial Neural Networks (ANN) [9].

Although most works in the literature contribute more to the classification of emotions than to the regression of emotional dimensions, the regression approach aiming to predict continuous values in the emotional plane could be suggested in clinical contexts where it is important to track the evolution of a patient's emotional state during treatment. Deep learning (DL) techniques, such as autoencoder, deep belief network (DBN), convolutional neural network, and recurrent neural networks, have been extensively used to build emotion recognition architectures that outperform other standard machine learning approaches [10]. Moreover, deep neural networks offer the benefit of dealing directly with raw data and automating feature extraction and selection via high-level data representation.

One disadvantage of DL approaches, however, is their lack of interpretability. Understanding how the models affect the decisions and how each predictive variable is involved in the decision-making process for each instance is essential both to increase the degree of confidence in the models and to correct and act in case of bias and erroneous decisions [11]. Explainable Artificial Intelligence (XAI) methods have been recently introduced to overcome these limitations. In particular, local post-hoc techniques such as LIME [12] and SHAP [13] have gained popularity due to their ability to provide agnostic explanations for the decisions of most ML and DL algorithms at the local level.

In this work, we extend a prototype of an EEG-based emotion recognition system that can detect the user's emotional state using a deep learning model embedded into an interpretable framework to analyze the decisions of the model and the contributions of the features selected to describe the emotional state.

4. Materials

4.1. Dataset Description

The Dreamer dataset consists of 23 users' EEG signals during emotional elicitation. The emotional elicitation protocol was performed using audio/video clips. Eighteen video clips were used and classified into nine basic emotions such as amusement, excitement, happiness, calmness, anger, disgust, fear, sadness, and surprise. Each user was required to watch all video clips ranging from 65 to 393 seconds. After each video clip ended, users provided a self-assessment based on a 5-point Likert scale for Valence, Arousal, and Dominance. To carry out this task, participants were asked to complete the Self-Assessment Manikin questionnaire at the end of each experiment [14]. Recordings without emotional elicitation (baseline) and recordings during emotion induction were collected in the data set. At the end of the study, the authors published the dataset in Matlab format [1].

5. Methods

5.1. Preprocessing approach

One of the significant problems with EEG signals is the strong presence of artifacts (noise) or faulty EEG channels that can impair data analysis. An essential aspect of our study was automated preprocessing to create a user-friendly routine for acquiring real-time EEG signals.

[1] https://it.mathworks.com/products/matlab.html

The preprocessing flow is critical because our prototype aims to provide a real-time detection system of the user's emotional state. Our Preprocessing approach is also applied to the Dreamer dataset for training deep learning models. Afterward, we checked the efficiency of the automatic preprocessing technique by visually inspecting all EEG trials. More specifically 414 samples were selected containing information related to Valence, Arousal, and Dominance values. The first pre-processing step was the removal of the DC offset (DC offset). Then the data format conformed to the MNE framework respecting the 10-20 standard. A notch filter calibrated to the cutoff frequency of 50hz was used to remove noise due to the commercial electric current. The trial was normalized in the frequency range 1-40Hz. Epochs of length equal to 1 second were created from the continuous EEG signal. independent component analysis (ICA), was used to remove all noisy epochs and for the identification of EEG signal components. All artifacts in the signal were correctly removed. identification and interpolation of defective channels and epochs were performed using the pyprep framework [2]. However, epochs exceeding a certain noise threshold are removed and not interpolated. Finally, the continuous EEG signal is reconstructed by merging all the various preprocessed epochs.

5.2. Training dataset

After obtaining the preprocessed EEG signal, the data set is structured as follows: (i) 4 seconds epochs are extracted from continuous EEG trial; (ii) Overlapp epochs are obtained every 1280 samples. The new generated epochs have the same label as the originals; (iii) Theta, Alpha, Beta1, Beta2, Beta3 bands are extracted with neurokit2 framework[3]. The band range considered are 4-8Hz in Theta, 8-13Hz in Alpha, 13-16 Hz in Beta1, 16-20Hz in Beta2, 20-30Hz in Beta3.

5.3. DL Model Description

All the obtained features were initially split into train, validation, and test with the sklearn train_test_split library in proportion 80% for training and the remaining 20% for testing. The training dataset was then split into 75% for train and 25% for validation. After this operation, the normalization was performed with the MinMax scaler of sklearn [15]. The model used to make regression predictions is a 1D convolutional neural network (CNN) because it is useful in order to predict vectors of features at one size. The reference frameworks for the model are Keras [16] and Tensorflow [17]. The model consists of three convolutional layers, of which two to 128 neurons and a last to 64 neurons. A BatchNormalization was performed at the end of the first two layers of filters. Each layer was then condensed with the MaxPooling-1D in order to extract the most relevant correlation of engineered features. The kernel size is kept at 3, and the activation functions are Relu for convolutional layers. At the end of the convolutional layers, a Flatten operation is performed to create the input arrays for the next neural network. The neural network useful in order to predict regression values is a Fully Connected Layer composed of four layers, one of which is 128 neurons input, a second hidden at 128 neurons, and a third hidden layer at 32 neurons. The activation functions are relatively Tanh for the first two layers and Relu for the 32-neuron layer. Then a Dropout operation of 0.2 was performed in order to regularize learning to avoid overfitting. Finally, there is the last three neurons' output layer with linear output function useful for the purpose of the regression task. The three classes we want to predict are Valence, Arousal, and Dominance. During the learning, the mean_absolute_error was monitored as a loss function, and a callback was set to stop learning if the loss did not improve after ten iterations. (Patience = 10). The optimizer chosen is the Adam algorithm [18]. The structure of the network is presented in Fig. 2.

The main aim of our study is to provide an emotion recognition system that can provide real-time feedback on the user's emotional condition. In order to achieve this goal, the minimum length in terms of seconds was sought in relation to the greater level of accuracy of the R2 metric. In practice, the minimum time that maintains the levels of accuracy above 0.9% of R2 was sought, achieving periods not less than 4 seconds. In the same way, the optimal overlap coefficient was chosen to maintain the value of the metric R2 not less than 0.9%. This optimization of the hyperparameters was carried out experimentally to directly find the best solution that could avoid considering the eras of the EEG signal not too long but neither too short. Assuming to use epochs of 1-second length, or 128 samples, is not representative of an emotional state. Increasing the length of the epoch, the value R2 increases, but consequently, it creates a problem relative to the time of scan of the signal EEG during the acquisition in real-time.

5.4. Interpretable model

The SHAP algorithm has been selected to explain the predictions of the DL model for the independent test set. SHAP represents the marginal contribution of each input variable in the model's decision-making process. This algorithm is based on game theory and in particular Shapley's approach for evaluating the contribution of each player in a cooperative game. SHAP introduces a variant of Shapley's approach through the use of a local

[2] https://pypi.org/project/pyprep/0.2.1/
[3] https://neurokit2.readthedocs.io/en/latest/

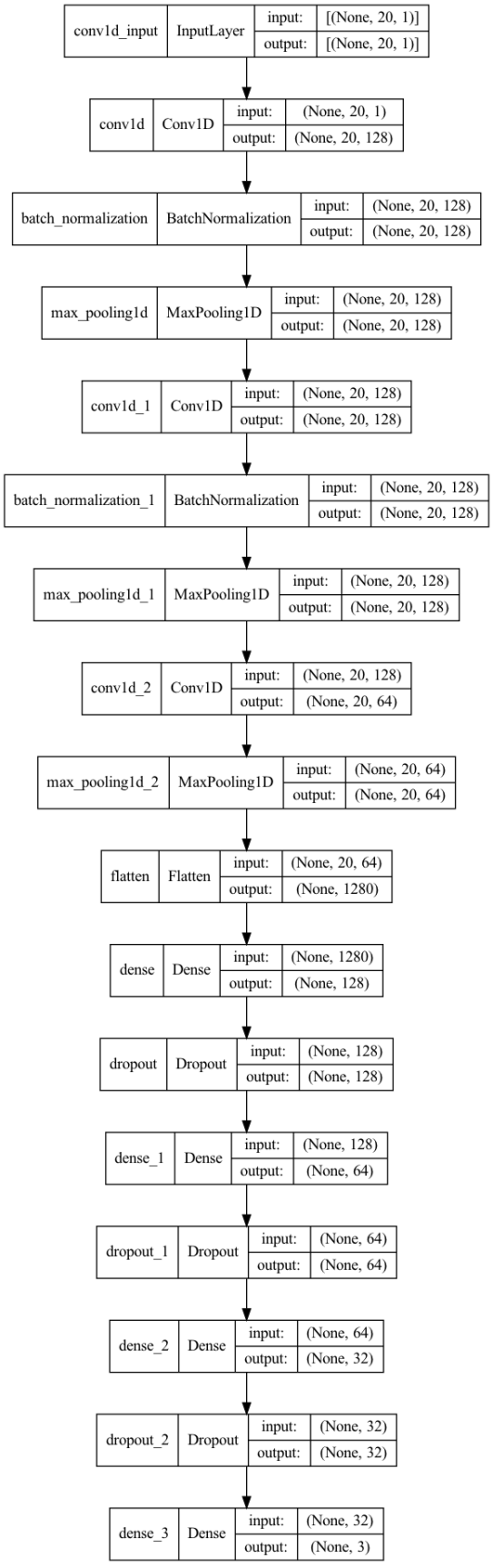


Figure 2: Architecture of CNN-1D.

contribution function, which calculates the contribution of each variable for each input instance of the test set.

For each output variable, we provide a waterfall plot in SHAP representing the contribution of each feature towards the final output of the DL model for a particular instance of data. The plot shows the base value, which is the expected output of the model when no features are observed, and the sum of the contributions of each feature to the final output for the given instance of data. Each feature is represented as a horizontal bar, and the length of the bar represents the magnitude of its contribution. Features that increase the output are shown in blue, while those that decrease the output are shown in red. The plot shows how each feature contributes to the final prediction, highlighting which features are driving the model for the given instance of data.

6. Results and Discussion

The DL model achieves the following levels of predictive accuracy: $R^2 = 0.93$, Mean Absolute Error $= 0.08$, Mean Absolute Percent Error $= 0.07$. All metrics are calculated with sklearn.metrics. Fig. 3 shows the course

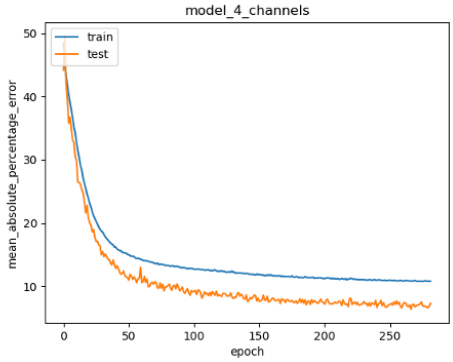


Figure 3: Loss – Mean Absolute Percent Error.

of the function of loss during the learning phase. It can be observed that no overfitting occurs during the training of the model. The proposed architecture has been trained to predict valence, arousal, and dominance levels simultaneously, thus classifying human emotions dynamically over time. From this point of view the model could be used to track the emotional history of a subject in real-time.

The waterfall plots for each outcome shown in Figure 4 suggest that the selected features for the four channels have comparable impacts for all test samples, except for the feature AF3 Beta 3, whose impact is higher than the others for the predictions of the three outputs. This finding encourages further exploration of the impact of

frequency features in the 20-30 Hz band and the use of frontal electrodes for human emotion recognition.

7. Conclusion

The proposed DL model for the three-dimensional regression of valence, arousal and dominance, constitutes a prototype for the continuous tracking of human emotional states and for explaining the impact of spatial and frequency features. It could provide effective information in clinical settings and be used as a tool to support diagnosis. Future developments include training with larger populations and the use of non-linear and complex features to complement the frequency features used in this work.

Acknowledgments

This work was partial support of the projects: Italian P.O. Puglia FESR 2014 – 2020 (project code 6ESURE5) 'SECURE SAFE APULIA', Fincons CdP3, PASSPARTOUT, Servizi Locali 2.0, ERP4.0. Also this work has been carried out while Tommaso Colafiglio was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Politecnico di Bari.

References

[1] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. B. Westover, M. T. Bianchi, J. Sun, Sleepnet: automated sleep staging system via deep learning, arXiv preprint arXiv:1707.08262 (2017).

[2] L. R. Gianotti, G. Künig, D. Lehmann, P. L. Faber, R. D. Pascual-Marqui, K. Kochi, U. Schreiter-Gasser, Correlation between disease severity and brain electric loreta tomography in alzheimer's disease, Clinical Neurophysiology 118 (2007) 186–196.

[3] L. Billeci, F. Sicca, K. Maharatna, F. Apicella, A. Narzisi, G. Campatelli, S. Calderoni, G. Pioggia, F. Muratori, On the application of quantitative eeg for characterizing autistic brain: a systematic review, Frontiers in human neuroscience 7 (2013) 442.

[4] X. Li, B. Hu, S. Sun, H. Cai, Eeg-based mild depressive detection using feature selection methods and classifiers, Computer methods and programs in biomedicine 136 (2016) 151–161.

[5] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, Journal of neural engineering 16 (2019) 051001.

[6] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, M. S. Hossain, Multilevel weighted feature fusion using convolutional neural networks for eeg motor imagery classification, Ieee Access 7 (2019) 18940–18950.

[7] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (1980) 1161.

[8] H. Gunes, B. Schuller, Categorical and dimensional affect analysis in continuous input: Current trends and future directions, Image and Vision Computing 31 (2013) 120–136.

[9] E. H. Houssein, A. Hammad, A. A. Ali, Human emotion recognition from eeg-based brain–computer interface using machine learning: a comprehensive review, Neural Computing and Applications 34 (2022) 12527–12557.

[10] C. Ardito, I. Bortone, T. Colafiglio, T. Di Noia, E. Di Sciascio, D. Lofù, F. Narducci, R. Sardone, P. Sorino, Brain computer interface: Deep learning approach to predict human emotion recognition, in: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2022, pp. 2689–2694.

[11] A. Lombardi, D. Diacono, N. Amoroso, A. Monaco, J. M. R. Tavares, R. Bellotti, S. Tangaro, Explainable deep learning for personalized age prediction with brain morphology, Frontiers in neuroscience (2021) 578.

[12] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

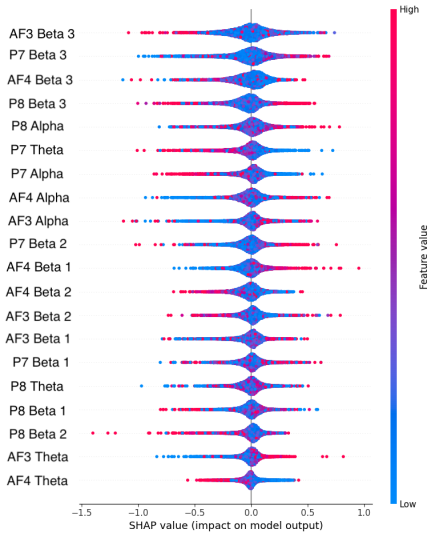
[13] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[14] M. M. Bradley, P. J. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, Journal of behavior therapy and experimental psychiatry 25 (1994) 49–59.

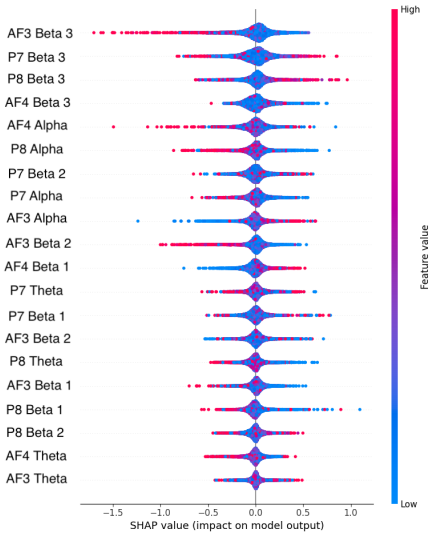
[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[16] F. Chollet, et al., Keras, 2015. URL: https://github.com/fchollet/keras.

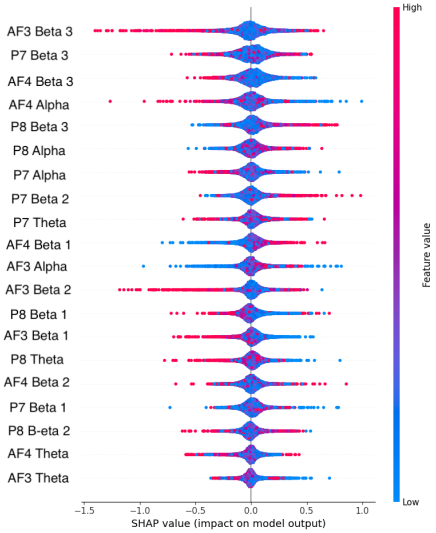
[17] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens,



(a) Waterfall Plot Valence.



(b) Waterfall Plot Arousal.



(c) Waterfall Plot Dominance.

Figure 4: Representation of shapley values for each emotion under study

B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[18] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).