Applying Big Transfer-based classifiers to the DEAP dataset

Taylor Sweet and David E. Thompson

Abstract— **Affective brain-computer interfaces are a fast-growing area of research. Accurate estimation of emotional states from physiological signals is of great interest to the fields of psychology and human-computer interaction. The DEAP dataset is one of the most popular datasets for emotional classification. In this study we generated heat maps from spectral data within the neurological signals found in the DEAP dataset. To account for the class imbalance within this dataset, we then discarded images belonging to the larger class. We used these images to fine-tune several Big Transfer neural networks for binary classification of arousal, valence, and dominance affective states. Our best classifier was able to achieve greater than 98% accuracy and 99% balanced accuracy in all three classification tasks. We also investigated the effects of this balancing method on our classifiers.**

I. Introduction

Emotion classification is a growing area of research. Researchers have traditionally approached the classification of emotions by using one of two fundamental models: discrete and dimensional. This work will focus on the dimensional model which provides ways to express a wide range of emotional states. The affective states are expressed in a multi-dimensional space with each feature as a dimension. Common features include arousal, valence, and dominance [1]. In psychological research, the most common method of emotion classification is self-report. Self-reporting can have many issues such as dependence on external factors including the wording of the question [2]. Classification of emotion using physiological signals would lessen these effects and thus increase the consistency of responses. Additionally, the direct estimation of a user's affective state is of much interest to the field of affective computing [3].

There are several publicly available datasets that could be used for emotional classification from biological signals, however the Database for Emotion Analysis (DEAP) is the most heavily investigated [4]. Many groups have managed to achieve high accuracy using this dataset. However, the issue of class imbalance in the DEAP is often overlooked, which can cause misleading accuracy values [5].

Deep neural networks perform well in various computer vision tasks, including image classification. Many studies have applied these classifiers to a variety of medical imaging tasks [6]. Neural networks have shown success in a variety of common Brain-Computer Interface (BCI) tasks including P300 classification, robotic arm control, cursor control, classification of epileptic states, and emotion classification [7-11]. In particular, convolutional neural networks (CNNs) have been shown to produce better predictions in emotional classification compared to classifiers that use a shallower network [12]. However, one problem with using neural networks is that they often require lengthy training times [13]. One way to lessen the training time is to take advantage of transfer learning. In transfer learning a neural network is first trained on a large dataset comprised of many different tasks. Then this neural network is fine-tuned for the specific task of interest.

Big Transfer (BiT) is a family of convolutional neural networks (CNN), based on residual neural network (ResNet) architectures. These models were pre-trained on various sizes of popular computer vision datasets. The BiT-S models are pre-trained on the ILSVRC-2012 dataset, the BiT-M on the ImageNet-21k dataset, and the BiT-L on the JFT-300M dataset. When tested on other computer vision datasets these models achieved high performance. Interestingly, these models performed quite well even when only given a single example in each class for training [14].

The main objective of this study was to detect the affective state of an individual using electroencephalography (EEG) signals. To accomplish this task, we transformed the EEG signals from the DEAP dataset into images. This allows us to take advantage of the publicly available BiT family of classifiers.

II. Methods

A. DEAP dataset

The DEAP is a publicly available dataset consisting of 32-channel EEG along with other physiological signals. To elicit emotion, each participant viewed 40 one-minute music videos. Following the video, participants were asked to rate the video on a 9-point scale in valence, arousal, and dominance. The signals were collected from 16 male and 16 female participants, with an average age of 24.9 years. This study only uses the EEG channels from the preprocessed version of this dataset. This preprocessing consisted of the following: down sampling to 128 Hz, electrooculogram (EOG) artifacts removal, bandpass filtering from 4 to 45 Hz, common average referencing, and segmentation into 60 second trials with a 3 second baseline [4].

B. Class Balancing

The goal of this study was to build a binary classifier on all three emotional axes that the DEAP dataset includes. On all three axes we chose any video rated greater than or equal to 6 to belong to the high class and any rating equal to or below 3 to belong to the low class. Videos rated between 3 and 6 were discarded. One issue here is that using this method of labeling

* Funded by National Science Foundation under Grant No. 1910526

Taylor Sweet is with the Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506 (phone: 417-396-5088 email: tasweet@ksu.edu) .

David E. Thompson is with the Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506 (email: davet@ksu.edu).

causes more than 70% of all the videos used to belong to the high class on all three axes. To prevent label biasing issues, we first determined how many trials were in the smaller of the two classes (N). Then we selected the N highest or lowest rated trials for the remaining class and discarded the rest. This method forces both the classes to have the same number trials within them. Participant and emotional axis combinations that had less than 5 trials in the smaller class were not included in the final results. Two neural networks were also trained without balancing for comparison. In this case, results from any classifier that did not have at least one trial from each class in the training data were discarded.

C. Image Generation

We started with the preprocessed DEAP data described in Section II.A. First, the 3 second baseline and 60 second trials were separated. For each trial, the Fourier transform was applied to the signal and the baseline. To reduce the dimensionality and convert the baseline and signal to the same size, we summed the frequency coefficients over 0.5 Hz intervals ranging from 4 to 45 Hz. The final feature set was the decibel ratio between signal and baseline power. This process was repeated for each channel and all decibel outputs were concatenated. This array was then plotted as a heat map which had dimensions of channel by frequency as shown in Fig. 1. Finally, each image was saved as a portable network graphics file, with a size of 875x656 pixels

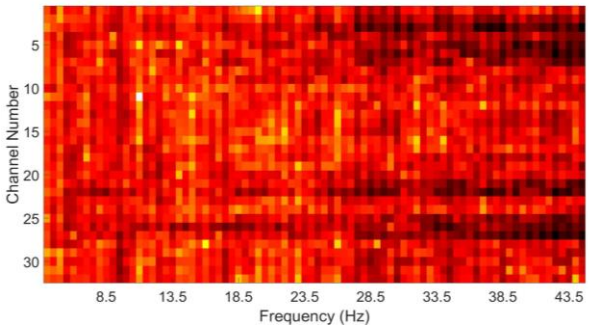


Figure 1. Example heat map produced by the image generation step. Axis labels and scales here are for display purposes only and were not included on the images used in classification.

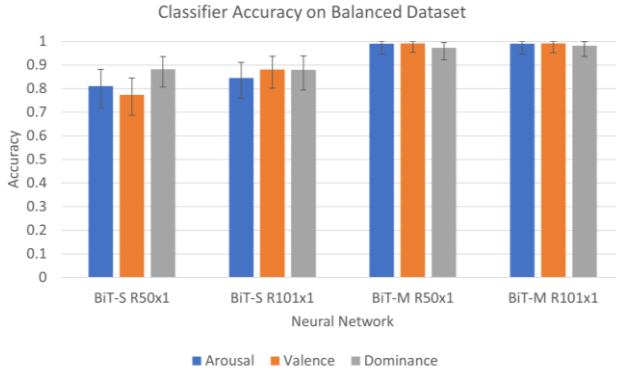
D. Emotion Classification

A separate classifier was built for each participant and emotional dimension combination. To match the input requirements of the BiT models, the heat map images that were generated for each trial were resized to 512 by 512 pixels and the color values were divided by 255 so that they would range from 0 to 1. The images were randomly sorted into train and tests with 70% of the images going into the train set. The images and labels were then used to fine-tune one of the following publicly available neural networks: BiT-S R50x1, BiT-S R101x1, BiT-M R50x1, or BiT-M R101x1. The fine-tuning step took place over 30 epochs, with 10 steps per epoch. The neural network was set to use a stochastic gradient descent optimizer and a sparse categorical cross-entropy loss function. To illustrate the effect of pre-balancing the labels we included results from the BiT-S R50x1 and BiT-M R101x1 models when the pre-balancing is forgone. Kansas State University's high-performance computing cluster, Beocat, was used to decrease the overall training time and allow the training of many classifiers simultaneously.

III. Results

Table I. shows the average accuracy and balanced accuracy values for each neural network used. Balanced accuracy is defined as the mean of true positive rate and the true negative rate. This is metric is more suitable for imbalanced datasets such as the DEAP [5]. The BiT-M R101x1 model had the best performance except in balanced accuracy on the arousal dimension and accuracy in the valence dimension, in both cases the BiT-M R50x1 had the best performance. Overall, our best classifier was the balanced BiT-M R101x1 model which had over 98% accuracy and balanced accuracy in all classification tasks. Fig. 2 displays the accuracy ratings for each neural network on the balanced dataset. The BiT-M models achieved significantly higher average accuracies than the corresponding BiT-S models in arousal and valence, but no significant difference was observed in dominance. No significant difference in accuracy was observed between the R101x1 models and their corresponding R50x1 models.

Figure 2. Classifier accuracies on the balanced DEAP dataset with 95% confidence intervals.



In Table I, we can see that classifiers that were trained on unbalanced datasets achieved similar accuracies in all categories compared to their balanced counterparts. The unbalanced BiT-S R50x1 model achieved higher accuracy in arousal and valence than the corresponding balanced classifier. The unbalanced BiT-M R101x1 model however achieved lower accuracy in all three tasks compared to its counterpart. Additionally, the balanced classifiers did achieve higher average balanced accuracy in comparison to the corresponding unbalanced classifiers in all tasks. In Fig. 3, we compare the balanced accuracy results of our classifiers on the balanced and unbalanced datasets. Here we can see the only significant difference is the difference in balanced accuracy between the balanced and unbalanced classification on the dominance task using the BiT-S R50x1 model.

IV. Discussion

Table I. demonstrates that the classification strategy provided in this work was able to achieve high accuracy and balanced accuracy. Generally, models based on the ResNet 101x1 architecture achieved higher average accuracy and balanced accuracy than their counterparts using the smaller ResNet 50x1 architecture. Also, the BiT-M models outperformed the corresponding BiT-S models. These observations mostly agree with the results reported by [14].

TABLE I. Classifier Results

Balanced Classifiers	**Accuracy (%)**	**Balanced Accuracy (%)**
BiT-S R50x1	81.00 (Arousal) 77.31 (Valence) 88.18 (Dominance)	83.16 (Arousal) 78.84 (Valence) 89.11 (Dominance)
BiT-S R101x1	84.54 (Arousal) 88.12 (Valence) 87.91 (Dominance)	86.14 (Arousal) 88.86 (Valence) 88.39 (Dominance)
BiT-M R50x1	99.00 (Arousal) 99.16 (Valence) 97.27 (Dominance)	99.29 (Arousal) 99.23 (Valence) 97.46 (Dominance)
BiT-M R101x1	99.00 (Arousal) 99.11 (Valence) 98.18 (Dominance)	99.14 (Arousal) 99.23 (Valence) 98.27 (Dominance)
Unbalanced Classifiers		
BiT-S R50x1	83.47 (Arousal) 80.66 (Valence) 83.47 (Dominance)	76.14 (Arousal) 69.04 (Valence) 70.57 (Dominance)
BiT-M R101x1	97.11 (Arousal) 98.35 (Valence) 98.73 (Dominance)	95.70 (Arousal) 97.50 (Valence) 97.54 (Dominance)

In Table II five recent publications using the DEAP dataset for binary classification on arousal, valence, and dominance were selected for comparison. We can see that the average of our proposed models achieves comparable performance to these recent works. Our best model, the balanced BiT-M R101x1, achieved higher accuracy in all three classifications compared to these works. All the publications in Table II defined the high and low class differently than our proposed method. This limits the ability to make a direct comparison between the classifiers, as it is possible that changing the class definitions made our classification task less difficult. Furthermore, how the high and low classes are defined determines how imbalanced the dataset will be, which can also affect classifier performance. None of the works included in Table II. provided a metric that is less sensitive to class imbalance, such as balanced accuracy or an F1 score, which would provide more insight into how the class imbalance affected their performance.

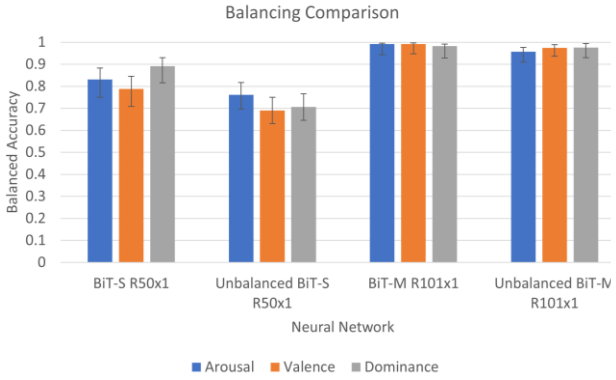


Figure 3. Effect of pre-balancing the dataset on balanced accuracy of multiple models, error bars correspond to 95% credible intervals.

Our proposed balancing strategy largely achieved higher average accuracies and balanced accuracies than the unbalanced approach. However, only one of these differences was significant. There are a few reasons that could limit the effectiveness of this balancing step. The first being that the balancing strategy reduces the number of trials and participants that are available. On average the balanced classifiers trained on 12.3 images and tested on 5.3 images. These numbers represent less than half of the total images provided by the DEAP dataset. Additionally, since the results of any participant and emotional dimension with less than 5 trials in its smallest class were discarded, the number of useable participants in each emotional dimension was reduced by an average of 38%. Another possible concern is that the separation into the train and test sets for the classifier was performed randomly. This implies that the balanced datasets can become unbalanced due to this separation step. However, in the balanced data classifiers, we did not observe any large difference between accuracy and balanced accuracy. Similar accuracy and balanced accuracy imply that the classifiers are not favoring either class, which suggests that the separation step did not cause any issues with regards to class balancing. The unbalanced classifier using the BiT-M R101x1 also performed quite well, so it is possible this balancing step is altogether unnecessary. However, the unbalanced dataset often yielded a test set with only one class. Such a test set could allow a classifier that cannot discriminate between the two classes at all to nevertheless show a very high accuracy.

Future work will first focus on making this study more directly comparable to other works, by testing this dataset with a different class definition. Additionally, we would use a more rigorous classifier validation scheme such as k-fold cross validation. This would limit the effect of the one class test set issue in the unbalanced dataset. We are also interested in exploring other methods to adjust for the class imbalance in the DEAP. The trial reduction issue in the balanced datasets could be corrected by individually adjusting the thresholds for the high and low classes for each participant and emotional dimension pair. This would allow the classes to be balanced

while keeping more data. Finally, we would like to investigate the performance of the BiT family of classifiers on other emotional BCI datasets. No attempt was made to optimize the various hyperparameters of these classifiers; doing so may result in even better performance, thus hyperparameter optimization of these classifiers is also of interest.

TABLE II. COMPARISON TO RECENT STUDIES

Source	Classifier	Class Definition	Accuracy (%)
Zhao et al [15]	SCC-MPGCN	High >5 Low <5	97.02 (Arousal) 96.37 (Valence) 96.72 (Dominance)
Ahmed et al [16]	AsMap+CNN	High >5.5 Low <5.5	95.21 (Arousal) 95.45 (Valence)
Yang et al [17]	GDCSBR	High ≥5 Low <5	85.84 (Arousal) 84.91 (Valence)
Li et al [18]	BRS with similarity measure	High ≥5 Low <5	75.66 (Arousal) 72.86 (Valence)
Peng et al [19]	DW-FBCSP	High >5 Low <5	84.45 (Arousal) 81.14 (Valence)
This work (average)	Various BiT Models	High ≥6 Low ≤3	90.69 (Arousal) 90.45 (Valence) 92.29 (Dominance)
This work (best)	Balanced BiT-M R101x1	High ≥6 Low ≤3	99.00 (Arousal) 99.14 (Valence) 98.18 (Dominance)

V. CONCLUSION

In this work we presented a novel classifier for the DEAP data set. To prevent the class imbalance issue that is present in this dataset, we first discarded trials belonging to the larger class. We then generated an image from the neurological signals. This allows us to take advantage of recent advances in image processing and transfer learning. Then we used these images to fine-tune several BiT models. Our best model achieved above 98% accuracy and 99% balanced accuracy in the binary classification of arousal, valance, and dominance dimensions.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1910526. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Some of the computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, EPS-0919443, ACI-1440548, CHE-1726332, and NIH P20GM113109. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

REFERENCES

[1] J. A. Russell, "A circumplex model of affect," J. Pers. Soc. Psychol., vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.

[2] G. Kalton et al., "Experiments in Wording Opinion Questions Published by : Wiley for the Royal Statistical Society Experiments in Wording Opinion Questions," vol. 27, no. 2, pp. 149–161, 1978.

[3] S. Brave and C. Nass, "Emotion in Human-Computer Interaction," Expand. Front. Vis. Anal. Vis., no. Cmc, pp. 239–262, 2012, doi: 10.1007/978-1-4471-2804-5_14.

[4] Koelstra, S., Muhl, C., Soleymani, M., Jong-Seok Lee, Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. and Patras, I., 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. IEEE Transactions on Affective Computing, 3(1), pp.18-31.

[5] M. R. Mowla, R. I. Cano, K. J. Dhuyvetter, and D. E. Thompson, "Affective brain-computer interfaces: Choosing a meaningful performance measuring metric," Comput. Biol. Med., vol. 126, no. August, p. 104001, 2020, doi: 10.1016/j.compbiomed.2020.104001.

[6] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep Learning Applications in Medical Image Analysis," IEEE Access, vol. 6, pp. 9375–9379, 2017, doi: 10.1109/ACCESS.2017.2788044.

[7] H. Shan, Y. Liu, and T. Stefanov, "A simple convolutional neural network for accurate P300 detection and character spelling in brain computer interface," IJCAI Int. Jt. Conf. Artif. Intell., vol. 2018-July, pp. 1604–1610, 2018, doi: 10.24963/ijcai.2018/222.

[8] Z. Tayeb et al., "Validating deep neural networks for online decoding of motor imagery movements from eeg signals," Sensors (Switzerland), vol. 19, no. 1, 2019, doi: 10.3390/s19010210.

[9] J. R. Stieger, S. A. Engel, D. Suma, and B. He, "Benefits of deep learning classification of continuous noninvasive brain-computer interface control," J. Neural Eng., vol. 18, no. 4, 2021, doi: 10.1088/1741-2552/ac0584.

[10] Y. Gao, B. Gao, Q. Chen, J. Liu, and Y. Zhang, "Deep convolutional neural network-based epileptic electroencephalogram (EEG) signal classification," Front. Neurol., vol. 11, no. May, pp. 1–11, 2020, doi: 10.3389/fneur.2020.00375.

[11] J. Liu et al., "EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder," Front. Syst. Neurosci., vol. 14, no. September, pp. 1–14, 2020, doi: 10.3389/fnsys.2020.00043.

[12] H. R. Kim, Y. S. Kim, S. J. Kim, and I. K. Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," IEEE Trans. Multimed., vol. 20, no. 11, pp. 2980–2992, 2018, doi: 10.1109/TMM.2018.2827782.

[13] H. Bhavsar and A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning," Int. J. Soft Comput. Eng., vol. 2, no. 4, pp. 74–81, 2012.

[14] A. Kolesnikov et al., "Big Transfer (BiT): General Visual Representation Learning," 2020, pp. 491–507.

[15] H. Zhao, J. Liu, Z. Shen, and J. Yan, "SCC-MPGCN: Self-attention coherence clustering based on multi-pooling graph convolutional network for EEG Emotion Recognition," Journal of Neural Engineering, 2022.

[16] M. Z. I. Ahmed, N. Sinha, S. Phadikar, and E. Ghaderpour, "Automated Feature Extraction on AsMap for Emotion Classification Using EEG," *Sensors*, vol. 22, no. 6, pp. 1–17, 2022, doi: 10.3390/s22062346.

[17] L. Yang, S. Chao, Q. Zhang, P. Ni, and D. Liu, "A Grouped Dynamic EEG Channel Selection Method for Emotion Recognition," *Proc. - 2021 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2021*, pp. 3689–3696, 2021, doi: 10.1109/BIBM52615.2021.9669889.

[18] J. W. Li *et al.*, "EEG-based Emotion Recognition Using Similarity Measure of Brain Rhythm Sequencing," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 31–34, 2021, doi: 10.1109/EMBC46164.2021.9629520.

[19] H. Peng, W. Lin, G. Cai, S. Huang, Y. Pei, and T. Ma, "DW-FBCSP: EEG emotion recognition algorithm based on scale distance weighted optimization," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 430–433, 2021, doi: 10.1109/EMBC46164.2021.9629850.