Supplementary Materials

Motor Imagery Classification for Brain Computer Interface using Deep Convolutional Neural Networks and Mixup Augmentation
Haider Alwasiti, and Mohd Zuki Yusoff, Member, IEEE

I. INTRODUCTION

Motor imagery BCI is commonly based on analyzing the EEG signals in the frequency domain, since the most commonly reported EEG changes associated with the motor imagery are the changes in the amplitude of certain range of frequencies. However, due to the nature of EEG brain waves, there are large trial to trial inter and intra individual differences in the spectral changes. Moreover, each brain region is usually responsible for multiple neurophysiological functions [35]. This leads to the inherent difficulty of detecting the imagery process of the individual.

In literature, BCI classifiers were usually trained on datasets with small to moderate training set size (~10-40 subjects) where a single classifier model is trained on all subjects’ training data. Then, the accuracy of the classification is reported on each subject separately [11], [15], [18], [26], [30], [31]. The main reason for using all subjects’ EEG data to train a single classifier is likely that most classifiers perform better with more training data. However, due to the large inter-individual EEG variability, there is a tradeoff between incorporating large number of subjects to increase the training set size (which improves the classification’s accuracy) and the more variability introduced to the training set due to the inter-individual EEG variability (which has a negative effect on the classification accuracy). Obviously, the best approach would be to collect larger EEG data from each subject and train a single model on each subject. However, collecting EEG signals is time consuming, and EEG recording with long sessions degrade the signal to noise ratio (SNR) of the collected signals because of the scalp-electrode impedance degradation with time. Therefore, most BCI studies adopted the approach of training the classifier on multiple subjects with relatively short EEG sessions. Only a few studies reported training a deep learning classifier on only one subject by incorporating private datasets that are substantially larger than the publicly available MI-BCI EEG datasets [7], [26]. The proposed study employed small training set per each subject from a publicly available dataset. It shows that the designed model can be used effectively with the training of ~120 EEG trials for only one subject per model with high classification accuracy in comparison to the previous Motor Imagery (MI) BCI classifiers. The proposed method eliminates the need for building a large private EEG dataset per subject with long data collection sessions.

In light of this, the motivation in this study is to employ a convolutional neural network (CNN) model to classify MI-EEG signals with a special kind of augmentation that has been applied for the first time on EEG signals to achieve high classification performance with a single model for each subject with a limited training set size. The aim is to effectively mitigating the inter-individual variability issue of motor imagery brain computer interface classification.

In literature, various types of classical machine learning (ML) classifiers have been reported for BCI classification. For example, Huang et al. [12] used a decision tree classifier combined with mahalanobis linear distance classifier achieving 0.57 classification accuracy. This was optimized by a model fine tuning to decode EEG activity. On the other hand, Handiru et al. [11] reported an iterative multiobjective optimization to decrease the number of feature space dimensions. Moreover, multiple channel selection and other dimension reduction algorithms were employed reporting 0.61 average accuracy. They had achieved 0.80 classification accuracy when the 35 top-performing subjects were chosen from a total of 109 subjects of the same public dataset used in this study [8].

II. MATERIALS AND METHODS

A. Dataset
Physiobank Motor/Mental Imagery (MMI) database [8] has been employed for this study, which consists of 64 channels EEG recorded from 109 subjects based on the 10-10 Electrode placement standard [32] performing different motor imagery trials. Each subject was asked to imagine the movement of the right or left hand if a target appeared on either the right or left side of the screen. The subjects imagined opening and closing the fist until the target removed, which was preceded by a short period of rest without activity. There were trials that had no target appearing on the screen and were labeled as rest. On average, 150 trials of 4-seconds segments sampled at 160 samples per second were obtained per each subject with roughly equal numbers of right, left, and rest trials.

B. Approach overview
The aim is to classify the EEG signals that are correlated with the imagery movement of the right, left hand, or the rest state with a limited training dataset. For each 4 seconds segment trial, the 64 EEG channels were transformed to the frequency domain by plotting the EEG power over time. Spectrograms were placed on their associated EEG channels following the 10-10 EEG standard to produce a topographical plot for each trial and fed to convolutional Neural Network (CNN). The traditional method is to combine all the user’s EEG data and try to build one classifier for all the users. However, we found that we could get around 0.10 more average accuracy by training CNN classifier for each user. The most challenging part of this approach is the extremely limited dataset that is used for training deep learning CNN which is
known for the need for a relatively large dataset for training convergence. Mixup augmentation [16] was essential to prevent the CNN from overfitting on such a small dataset and enabling a profile based approach of training a classifier for each user. Mixup augmentation is basically a linear interpolation between 2 items of different classes and it’s label is the weighted average of the original labels in the same mixup ratio of their corresponding images. This would enable expanding the dataset with a different range of linear combinations of any two pairs of items, which can generate orders of magnitude of synthetic data from a limited dataset to avoid overfitting [16].

C. Data preprocessing

The main advantage of Stockwell Transform over STFT is the implicit phase-normalized frequency bands, making the timing of the frequency components of the Stockwell transform distortion-free. Hence it is known for its ability to recover the input signal from the transform in a lossless way [29]. Moreover, the window function of Stockwell Transform is proportional to the frequency, allowing the Stockwell Transform to perform better with low frequency signals like EEG signals.

The spectrograms of the 64 channels have been plotted on a single image for each trial. The placement was following the 10-10 electrode placement standard. However, for reasons that will be elaborated in the results, we have changed the placement of spectrograms in a way that eliminates the background area and maximizes the areas of spectrograms (see Fig. 1).

![Fig. 1. Concatenated spectrogram images of a single trial. The figure shows 8x8 spectrogram images, each from a single EEG channel for a single trial. The total resolution is 696x512 pixels.](image)

D. Convolutional neural network

The Average Pool 2d layer was replaced by Adaptive Average Pool 2d layer and added an Adaptive Max Pool 2d layer to capture more features from the CNN body. Both these layers take input from the 512 features output of the final conv2d layer. The output of both pool layers was flattened and fed into Batch Normalization layer with a dropout of 0.5 probability. Finally, two linear layers instead of the standard one linear layer have been added. The first linear layer takes 1024 features from the two pooling layers and outputs 2048 features to be fed into ReLu, BatchNorm, dropout layer, and finally, the last linear layer. This custom head proved to improve the accuracy since it has a higher capacity for learning a larger number of features in the spectrogram images. Our modified model has increased the model size and parameters by only 16% (total params: 13,288,003).

Training has been performed on Nvidia GTX-1080Ti video card with Intel Core i9-9900k CPU. Pytorch and fastai library were used with Adam optimizer, batch size of 32, and input image size of 696x512. The optimal learning rate was estimated by a learning rate finder, where the loss was plotted against different learning rates. The best maximum learning rate was chosen around one order of magnitude less than the inflection point where the loss was starting to increase. Hence, 1e-03 learning rate has been assigned as the maximum learning rate in the first stage of learning. Furthermore, the learning rate was changed over the training iterations using the one cycle policy [28], where the learning rate starts with a very small learning rate and reaches to its peak at 30% of the whole training iterations and decreases gradually again.

The training set (80% of the dataset) portion was not used in the evaluation, and 20% of the dataset was used to estimate the performance of the model. The training has been performed in 2 stages with objective function minimizing the cross-entropy loss using the Adam optimizer [14]. Initially, all layers except the model’s head have been frozen (i.e., non-trainable during training) in stage 1. This allowed the model to train only the head in stage 1 before performing the full model’s training in stage 2. Empirically, this showed that the model converged faster. This is likely because the head was initialized with random weights, unlike the CNN body with the pretrained ImageNet weights. Faster convergence can be achieved - albeit with comparable results - if we gradually make changes to the weights when we first train the random weights of the linear layer, then train the whole model at the second stage. The first stage was trained for 300 epochs, and 4500 epochs for the second stage.

The second stage training was performed with discriminative layer training using three different learning rates. The CNN body was divided into two equal parts, and the third layer group represented the head of the model. Then we have applied the maximum learning rate max\(_l\) (which has been estimated using the learning rate finder algorithm) on the head layer group (1e-4, 8e-5 for ResNet18 and DenseNet121 respectively), and max\(_l\)/3 for the first and second layer groups, respectively. We found that discriminative layer training learning rates improved the accuracy over one learning rate for all the model layer groups. The reason is likely that the first layer group with ImageNet pretrained weights has almost optimum weights for any type of dataset since the first group CNN layers are responsible for detecting simple features in the images like edges and lines which are the most general knowledge [36]. For the second layer group of the CNN, a higher learning rate is desirable, since it is responsible for higher abstract features detection, like eyes or faces in the ImageNet dataset. For our dataset, those abstract features are more divergent from the ImageNet abstract features, so higher learning rates for that layer group could help. Finally, the head
was initialized from random weights, and no pretraining was used, so the highest learning rate could help the model to converge faster and better.

With mixup augmentation, we could improve the classification accuracy, and more importantly, converging the model was possible on only a much more limited dataset with only one subject’s EEG data. It is well known that there are substantial inter-individual differences in terms of brain waves correlate to motor imagery [9]. Hence training a model for each person would yield better model performance since it makes the model more adapted for the specific motor imagery EEG changes for that particular subject.

III. DISCUSSION

For comparison with the previous studies that used the same dataset [8] like the one employed in this study, the results of seven studies are reported herein. Tolić et al. [33] preprocessed the EEG with a discrete wavelet transform. Signal features were used as inputs for a shallow classical neural network classifier, achieving a mean accuracy of 0.68 on the same dataset. A study by Park et al. [22] used complex-valued common spatial pattern (CSP) algorithms for EEG preprocessing in order to generate complex signals with noncircular probability distributions. They filtered the EEG signals to restrict the frequency range into 8-30 Hz. With the aid of augmented complex statistics and the strong-uncorrelating transform (SUT) methods, they could show that the SUT algorithm with SVM maximized the inter-class difference between 2 tasks achieving 0.72 mean accuracy.

Kim et al. [13] used multivariate empirical mode decomposition (MEMD) followed by strong uncorrelating transform complex common spatial patterns algorithm as a preprocessing approach for MI-EEG. They used 14 electrodes with a wide frequency range of 1-80 Hz. A random forest classifier yielded 0.80 mean classification accuracy on the same dataset of this study. Similarly, Handiru et al. [11] proposed a Support Vector Machine classifier (SVM) with a channel selection algorithm that utilized an iterative multi-objective optimization reporting 0.61 classification accuracy. Loboda et al. [17] explored using the phase values for BCI applications. They used a total of 9 electrodes with EEG filtered into 8-30 Hz achieving an average accuracy of 0.72. Using both the EEG phase and EEG signal amplitudes is potentially a promising future research direction that can be based on the current research work. Ensembling different signal modalities into one model most likely will improve the accuracy more than if each model alone is considered to be adopted.

Interestingly, X. Ma et al. [18] used a simpler preprocessing pipeline where the EEG signals were only filtered in the time domain and used in the form of raw signal values. They used all the 64 channels with a bidirectional LSTM classifier achieving 0.68 mean classification accuracy. Although they have used the Physionet EEG dataset, which consisted of 109 subjects, however they have chosen twelve subjects only from the dataset. Within our experiments, we have observed several subjects who performed substantially worse in terms of classification accuracy than the others and had negative impact on the average accuracy. Several studies showed that 10-30% of healthy subjects were not able to modulate their EEG for motor imagery BCI control [2], [3], [5], [6], [20], [21], [23]–[25], [27], [34]. Few researchers coined the term BCI illiteracy on this condition [21], [25]. Multiple studies showed that BCI illiteracy is more severe in motor imagery classification in comparison to other BCI modalities like SSVEP or P300 [1], [2], [4], [10], [19]. For a comparison with classical ML methods, X. Ma et al. [18] carried out 2 baseline classical ML models training. They showed that the proposed DL method outperformed the classical ML methods where the CSP+LDA and the FBCSP+LDA methods achieved 0.59 and 0.60 classification accuracy, respectively. The study reported by X. Ma et al. [18] was similar to the approach reported by Zhang et al. [37] where similarly, they used a minimal EEG preprocessing approach by utilizing the time domain EEG as raw signal values only. The classification were performed by an LSTM with an orthogonal Array method for hyper-parameters tuning, achieving 0.95 classification accuracy. However, again they have selected only 10 subjects from the total of 109 subjects of the Physionet public dataset.

In this study, the proposed CNN model outperformed the previous models that used the same dataset, provided that the reported accuracy was on a model that used the entire 109 subjects of the same dataset. The reason can be related to two characteristics of the proposed CNN methods. First, the CNN classifier was able to converge and classify EEG signals with a small number of training samples for only single subject per model, effectively solving the issue of inter-individual variability of BCI EEG. Second, a special kind of augmentation (mixup augmentation) has been applied for the first time on EEG signals, which allowed high performance classification with a very limited dataset. Both factors were crucial to overcome the challenge of training on small dataset per each subject, which allowed a cleaner training set to train the CNN model without the need to combine all subjects to increase the training size. Otherwise, if the model is not able to converge on such a small training set of a single subject, combining multiple subjects’ data to increase the training size will eventually introduce the inter-individual variability leading to less classification accuracy.

IV. CONCLUSION

The proposed CNN models with mixup augmentation achieved 0.920 (95% CI 0.908, 0.933) and 0.933 (95% CI 0.922, 0.945) classification accuracy, respectively, which resulted in higher classification performance in comparison to other DL classifiers of previous studies on the same dataset. We show that with the presented approach, we could counteract the limited dataset issue, without sacrificing the performance of the classifier.

Others have attempted to classify MI-EEG signals using deep learning methods but needed a substantially larger training dataset in comparison to what could be achieved in this work, in part because the augmentation techniques that are commonly used with CNNs are not applicable to EEG spectrograms. This study shows that custom CNN with mixup augmentation can
counteract the limited dataset issue, without sacrificing the performance of the classifier.

REFERENCES

[1] Brendan Allison, Thorsten Luth, Diana Valbuena, Amir Teymourian, Ivan Volosyak, and Axel Grazer. Bci demographics: How many (and what kinds of) people can use an ssvep bci? IEEE transactions on neural systems and rehabilitation engineering, 18(2):107–116, 2010.
[2] Brendan Z Allison, Clemens Brunner, Vera Kaiser, Gernot R Mülller-Putz, Christa Neuper, and Gert Pfurtscheller. Toward a hybrid brain–computer interface based on imagined movement and visual attention. Journal of neural engineering, 7(2):026007, 2010.
[3] Brendan Z Allison, Dennis J McFarland, Gerwin Schalk, Shi Dong Zang, Melody Moore Jackson, and Jonathan R Wolpaw. Towards an independent brain–computer interface using steady state visual evoked potentials. Clinical neurophysiology, 119(2):399–408, 2008.
[4] Brendan Z Allison and Christa Neuper. Could anyone use a bci? In Brain-computer interfaces, pages 35–54. Springer, 2010.
[5] Brendan Z Allison, Elizabeth Winter Wolpaw, and Jonathan R Wolpaw. Brain–computer interface systems: progress and prospects. Expert review of medical devices, 4(4):463–474, 2007.
[6] Niels Birbaumer and Leonardo G Cohen. Brain–computer interfaces: communication and restoration of movement in paralysis. The Journal of physiology, 579(3):621–636, 2007.
[7] Antonio Maria Chiarelli, Pierpaolo Croce, Arcangelo Merla, and Filippo Zappasodi. Deep learning for hybrid eeg-fmri brain-computer interface: application to motor imagery classification. Journal of neural engineering, 15(3):036028, 2018.
[8] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mager, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation, 101(23):e215–e220, 2000.
[9] Bernhard Graimann, Brendan Z Allison, and Gert Pfurtscheller. Brain-computer interfaces: Revolutionizing human-computer interaction. Springer Science & Business Media, 2010.
[10] Christoph Guger, Shahab Daban, Eric Sellers, Clemens Holzner, Gunther Kraus, Roberta Carabalona, Furio Gramatica, and Guenter Edlinger. How many people are able to control a p300-based brain–computer interface (bci)? Neuroscience letters, 462(1):94–98, 2009.
[11] Vikram Shenoy Handiru and Vinod A Prasad. Optimized bi-objective eeg channel selection and cross-subject generalization with brain–computer interfaces. IEEE Transactions on Human-Machine Systems, 46(6):777–786, 2016.
[12] Dandan Huang, Kai Qian, Simon Oxenham, Ding-Yu Fei, and Ou Bai. Event-related desynchronization/synchronization-based brain-computer interface towards volitional cursor control in a 2d center-out paradigm. In 2011 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMI), pages 1–8, IEEE, 2011.
[13] Youngjoo Kim, Jiwoo Ryu, Ko Keun Kim, Clive C Took, Danilo P Mantini, and Cheolsoo Park. Motor imagery classification using mu and beta rhythms of eeg with strong uncorrelating transform based complex common spatial patterns. Computational intelligence and neuroscience, 2016:1, 2016.
[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[15] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Egnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering, 15(5):056013, 2018.
[16] Daojun Liang, Feng Yang, Tian Zhang, and Peter Yang. Understanding mixup training methods. IEEE Access, 6:58774–58783, 2018.
[17] Ana Loboda, Alexandra Margineanu, Gabriela Rotariu, and Ana-Mihaela Logar. Discrimination of eeg-based motor imagery tasks by means of a simple phase information method. International Journal of Advanced Research in Artificial Intelligence, 3(10), 2014.
[18] Xuelin Ma, Shuang Qiu, Changde Du, Jiezhen Xing, and Huiguang He. Improving eeg-based motor imagery classification via spatial and temporal recurrent neural networks. In 2018 48th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1903–1906. IEEE, 2018.
[19] Femke Nijboer, Niels Birbaumer, and Andrea Kubler. The influence of psychological state and motivation on brain–computer interface performance in patients with amyotrophic lateral sclerosis—a longitudinal study. Frontiers in neuroscience, 4:45, 2010.
[20] Femke Nijboer, Adrian Furdea, Ingo Gunst, Jürgen Mellinger, Dennis J McFarland, Niels Birbaumer, and Andrea Kubler. An auditory brain–computer interface (bci). Journal of neuroscience methods, 167(1):43–50, 2008.
[21] Anton Nijholt, Desney Tan, Gert Pfurtscheller, Clemens Brunner, José del R Millán, Brendan Allison, Bernhard Graimann, Florin Popescu, Benjamin Blankertz, and Klaus-R Müller. Brain-computer interfacing for intelligent systems. IEEE intelligent systems, 23(3):72–79, 2008.
[22] Cheolsoo Park, Clive Cheong Took, and Danilo P Mandic. Augmented complex common spatial patterns for classification of noncircular eeg from motor imagery tasks. IEEE Transactions on neural systems and rehabilitation engineering, 22(1):1–10, 2013.
[23] Gert Pfurtscheller, Robert Leeb, Doron Friedman, and Mel Slater. Centrally controlled heart rate changes during mental practice in immersive virtual environment: a case study with a tetraplegic. International journal of psychophysiology, 68(1):1–5, 2008.
[24] Gert Pfurtscheller, GR Mülller-Putz, A Schlogl, Bernhard Graimann, Reinhold Scherer, Robert Leeb, Clemens Brunner, Claudia Keirnarth, F Lee, George Townsend, et al. 15 years of bci research at graz university of technology: current projects. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14(2):205–210, 2006.
[25] Claudia Samelli, Miklo Braun, Michael Tangermann, and Klaus-Robert Müller. Estimating noise and dimensionality in bci data sets: towards illiteracy comprehension. 2008.
[26] Robin Tibor Schirrmieier, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glassettler, Katharina Egneggerspen, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. Human brain mapping, 38(11):5391–5420, 2017.
[27] Eric W Sellers, Andrea Kubler, and Emanuel Donchin. Brain-computer interface research at the university of south florida cognitive psychophysiology laboratory: the p300 speller. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14(2):221–224, 2006.
[28] Leslie N Smith. A disciplined approach to neural network hyperparameters: Part 1—learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820, 2018.
[29] Robert Glenn Stockwell, Lulu Mansinha, and RP Lowe. Localization of the complex spectrum: the s transform. IEEE transactions on signal processing, 44(4):998–1001, 1996.
[30] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial eeg classification. Journal of neuroscience methods, 274:141–145, 2016.
[31] Yousef Rezaei Tabar and Ugur Halici. A novel deep learning approach for classification of eeg motor imagery signals. Journal of neural engineering, 14(1):016003, 2016.
[32] Michal Teplan et al. Fundamentals of eeg measurement. Measurement science review, 2(2):1–11, 2002.
[33] Martina Tolic and Franjo Vukusic. Classification of wavelet transformed eeg signals with neural network for imagined mental and motor tasks. Kinesiology: International journal of fundamental and applied kinesiology, 45(1):130–138, 2013.
[34] Theresa M Vaughan, Dennis J McFarland, Gerwin Schalk, William A Samacki, Dean J Krusienski, Eric W Sellers, and Jonathan R Wolpaw. The wadsworth bci research and development program: at home with bci. IEEE transactions on neural systems and rehabilitation engineering, 14(2):229–233, 2006.
[35] Jonathan Wolpaw and Elizabeth Winter Wolpaw. Brain-computer interfaces: principles and practice. OUP USA, 2012.
[36] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Advances in neural information processing systems, pages 3320–3328, 2014.
[37] Xiang Zhang, Lina Yao, Chaoran Huang, Quan Z Sheng, and Xianzhi Wang. Intent recognition in smart living through deep recurrent neural networks. In International Conference on Neural Information Processing, pages 748–758. Springer, 2017.

Supplementary Materials