Transfer Learning with Large-Scale Data in Brain-Computer Interfaces

Chun-Shu Wei, *Student Member, IEEE*, Yuan-Pin Lin, *Member, IEEE*, Yu-Te Wang, *Member, IEEE*, Chin-Teng Lin, *Fellow, IEEE*, and Tzyy-Ping Jung, *Fellow, IEEE*

***Abstract*— Human variability in electroencephalogram (EEG) poses significant challenges for developing practical real-world applications of brain-computer interfaces (BCIs). The intuitive solution of collecting sufficient user-specific training/calibration data can be very labor-intensive and time-consuming, hindering the practicability of BCIs. To address this problem, transfer learning (TL), which leverages existing data from other sessions or subjects, has recently been adopted by the BCI community to build a BCI for a new user with limited calibration data. However, current TL approaches still require training/calibration data from each of conditions, which might be difficult or expensive to obtain. This study proposed a novel TL framework that could nearly eliminate requirement of subject–specific calibration data by leveraging large-scale data from other subjects. The efficacy of this method was validated in a passive BCI that was designed to detect neurocognitive lapses during driving. With the help of large-scale data, the proposed TL approach outperformed the within-subject approach while considerably reducing the amount of calibration data required for each individual (~1.5 min of data from each individual as opposed to a 90 min pilot session used in a standard within-subject approach). This demonstration might considerably facilitate the real-world applications of BCIs.**

I. Introduction

On the pathway of moving laboratory-developed brain-computer interfaces (BCIs) toward real-world applications, one of the grand challenges is to solve the variability of human electroencephalographic (EEG) dynamics, both across different individuals and within the same individual over time. In the past, a long and tedious calibration session was typically required to train a BCI classifier prior to the usage of a BCI system, which may be time-consuming and labor-intensive to collect sufficient training data depending on the types of BCIs [1]. To reduce the calibration time, transfer learning (TL) approaches were recently adopted in BCI research, such as motor imagery [2]-[4] and workload classification [4], to effectively reuse existing data from other sessions or subjects for obtaining satisfactory performance.

Most current TL approaches for BCIs could work well when sufficient representative training data under each of conditions are available [2]-[4]. However, it might be idealistic because collecting calibration data from each condition could be difficult or expensive in real life. This study proposes a novel TL framework that could leverage large-scale existing data from other subjects to minimize the need for collecting individual calibration data. The advantage of a large-scale TL approach has been previously demonstrated in an offline simulated BCI for neurocognitive lapse detection [5]. Yet the study was established on a large amount of calibration data (a full 90-min pilot session) from each new user. In this study, the proposed TL framework constructed a new model for a target subject using a minimal amount (~1.5 min) of 'calibration' data from him/her. This study further compared both the accuracy and the calibration time of BCIs based on the proposed TL approach with that of the conventional within-subject approach (WSA) that required a full pilot session from each target subject to build a classifier [6]-[8]. The impact caused by the amount of existing data on the performance of the large-scale TL approach was also investigated.

II. Large-Scale Transfer Learning for BCI

The concept of applying large-scale existing data and TL on initializing BCI for a new user without calibration phase was first introduced in [9]. However, to the best of our knowledge, no attempt has been made to experimentally validate the efficacy of such TL framework. In our previous study [5], we demonstrated that leveraging large-scale data from other subjects could improve the BCI performance over the conventional WSA that solely used a full 90-min calibration session from each individual [6]-[8]. This study further extended the TL framework by using much fewer calibration data while retaining a comparable classification performance, compared to that of the conventional WSA.

A. Transfer Learning Framework

Fig. 1 shows the schematic diagram of the proposed TL framework given an existing large-scale dataset from source subjects. For each target (*i.e.*, new) subject, the TL approach fuses a set of source classification models built upon data from other source subjects into a new model. This framework requires *a priori* information of the target subject for measuring its similarity to each source session/subject. In fact, the type of *a priori* information determines the calibration time of BCI usage for the target subject. We suggested using easily accessible *a priori* information, such as a short period of EEG recording from the beginning of the experimental session, namely calibration data. With appropriate *a priori* information, the source models could be fused and optimized according to 1) the generalizability of each source model to other source subjects, and 2) the similarity between the target subject's and the source subjects' calibration data.

B. Model Generalizability

The generalizability of a BCI model refers to the predictive capability of a BCI model from one session or subject to others. We defined the generalizability of a source model based on its overall performance on classifying/predicting the data of other subjects. Given a BCI model for session n, its generalizability G is expressed below

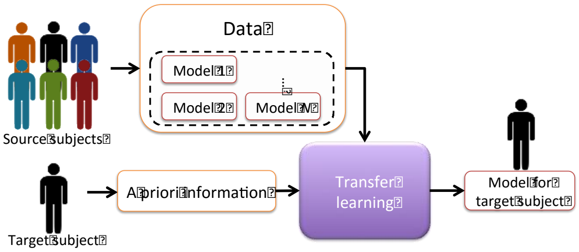

Fig. 1. A schematic diagram of the proposed TL framework. Source subjects provide sufficient existing data that could be used to build BCI models for the target subject. The target subject's *a priori* information is utilized for positively transferring source models into a new model for the target subject.

$$G = \overline{perf(n, \Phi(n))} \qquad (1)$$

where $\Phi(n)$ is the set of indices of sessions from all other source subjects given a session n in the data pool. $\overline{perf(n, \Phi(n))}$ is the median performance of session n's model on the classification/prediction of all other source subjects' sessions, *i.e.*, $\Phi(n)$.

C. *Data Similarity*

Data similarity is a factor commonly considered for applying transfer learning, since feeding dissimilar data could lead to negative transfers, *i.e.*, deteriorated performance [4], [5]. We proposed to estimate the similarity between subjects by comparing the individual spatio-spectral EEG patterns under the same condition. The spectral topographic pattern is represented as a 2-dimensional matrix $A_{F \times C}$, where F is the number of frequency bins, and C is the number of electrodes. This study calculated the Pearson's correlation coefficient, ρ, and used it to quantify the similarity between subjects. Note that the calculation of similarity was purely based on the short calibration data from the source and target subjects.

D. *Model Ranking and Fusion*

The performance of transfer learning mostly depends on how the information/data of source subjects are reused for the target subject. In this study, given a pool of K source models, each model was ranked according to their generalizability, G, and similarity to target subject, ρ, and the model n's rank was denoted as $k_n \in \{1, 2, \ldots, K\}$. This study then adopted a sigmoid weighting function to assign a higher score to a source model if it returned a higher generalizability and similarity, which is defined by the following formula:

$$w(k, s, b) = 1 - 1/(1 + e^{-s(\frac{k}{K} - b)}) \qquad (2)$$

where k is the ranking of a source model, $s \in \{10, 20, \ldots, 100\}$ and $b \in \{0.1, 0.2, \ldots, 1\}$ are parameters that adjust the slope and the position of half maximum, respectively, of the sigmoid function. It is worth noting that s and b were optimized using leave-one-subject-out cross validation within the source-subject pool. Lastly, the predictive output of the source model ensemble for a target subject was determined using the following equation:

$$\hat{y} = \sum_n \frac{w(k_n, s, b)}{\sum_n w(k_n, s, b)} \widehat{y_n} \qquad (3)$$

where $\widehat{y_n}$ is the output of the source model n for a given trial.

III. Data

A. Experiment

The data for testing the TL framework were collected from a sustained-attention driving task, during which both participants' EEG and task performance were continuously and simultaneously measured [10]. Participants were driving on a nighttime straight highway in a virtual-reality driving simulator with a coupe car body mounted on a 6-degree-of-freedom interactive motion platform. The experiment involved a lane-keeping task, in which the vehicle automatically cruised at 100 km/h speed, and deviated toward left or right randomly every 6-10 seconds. To maximize the chance of neurocognitive lapses from subjects during driving, we conducted the experiment at early afternoon when afternoon slump often occurs. The subjects were asked to keep the vehicle on course by steering the wheel once they sense the deviation. For each lane-deviant event, the duration from the onset of deviation to the onset of movement was measured as the response time (RT) for that trial, which reflects the level of neurocognitive state at a given moment. Trials with RT shorter than 1.5× (alert RT) were categorized as 'alert' trials, whereas those with RT longer than 2.5× (alert RT) were 'lapse' trials. The alert RT was individually estimated for each session as suggested in [5]. Forty-six sessions from 28 subjects were collected, in which 11 subjects participating in multiple sessions were selected as target subjects for comparing the TL performance to that obtained by conventional WSA approach [7].

B. EEG Feature Extraction

The relationship between human EEG and fatigue, drowsiness, and lapse has been studied for the past two decades [5]-[8], [10]-[12]. This study inherited the lapse-related EEG correlates found in our previous study [8], *i.e.*, pre-event EEG band powers including delta (2-5 Hz), theta (5-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and all-band (2-30 Hz). For each band-pass filtered EEG of each channel, the pre-event EEG segment from -3 to 0 second of each deviation onset was excerpted and converted into logarithmic band power using the following equation:

$$Power = \log_{10}(var(X)) \qquad □□□$$

where X is the band-passed EEG signal. The above procedure collected a 150-dimensional (30 channels × 5 bands) feature vector of pre-event EEG powers. Note that the 3s-long pre-event EEG data only included the brain activity during steady driving while avoiding the motor-related potentials. Because subjects usually started with nearly perfect task performance, the pre-event band powers (unsmoothed) of each channel of the first 10 alert trials from each subject were averaged and regarded as the

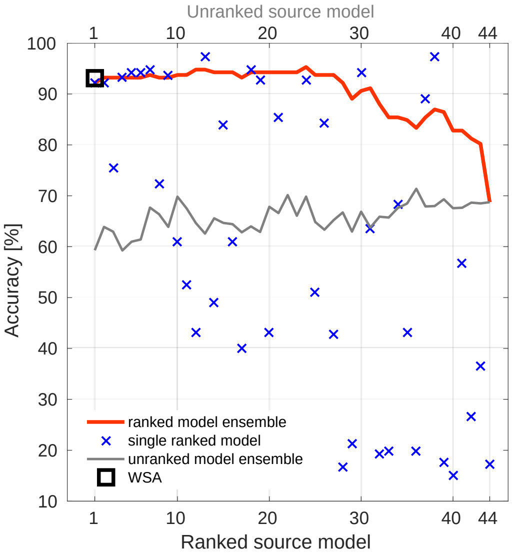

Fig. 2. A comparison of lapse classification accuracy across different BCI models (please see text for details) for a sample target session. Blue crosses indicate the performance of transferring each ranked source model to the sample target subject. The performance of uniform-weighted model ensemble based on the 1st to *k*th (red curve) source models, shows higher accuracy when *k*<28. The gray curve represents the performance of a uniform-weighted ensemble based on randomized models. Black square marks the accuracy of using calibration data from another 90-min session from the same target subject (the WSA model).

calibration data for calculating the similarity between subjects. In addition, an early study [6] suggested applying a smoothing window to eliminate unrelated power perturbation that accounted for slow (tonic) changes in EEG spectrum for EEG-based alertness estimation. Therefore, we proposed using both smoothed and unsmoothed pre-event power rather than using only the unsmoothed power. For each feature in the 150-dimensional feature vector, the smoothed power was calculated by averaging the pre-event powers of the adjacent trials occurred within a 90-second causal window from -90 to 0 second for each trial. At the end, the congregated 300-dimensional feature vector was used for lapse classification. The classifier employed in this study was support vector machine (SVM) implemented in the LIBSVM [13] with linear kernel and default settings.

IV. Results and Discussion

This study experimentally investigated the feasibility of the proposed large-scale TL framework for lapse detection. Fig. 2 illustrates the classification performance using different BCI models for a sample target subject (S8). X-axis represents the individual source models ranked by the generalizability and similarity (w.r.t. the target subject). We compared the performances obtained using (1) a single model approach that used the data from an individual source model to classify/predict the lapses in the target session (blue cross), (2) an ensemble model that used all models from the first *k* source models, *e.g.*, the results for *k*=10 used all models from the first 10 source models (equally weighed) (red curve), (3) an ensemble model that used models from unranked (random) *k* source models (equally weighed) (the gray curve using the top ticks), and (4) the conventional WSA that used a full 90-min calibration session from the test subject as the training data to build a within-subject model (black square).

Most of top-ranked source models ($k<10$) led to high classification performance, suggesting the effectiveness of using generalizability and similarity to rank source models. The ranked model ensemble that uniformly fused source models ranked from 1 through k could achieve high performance when k is small as top-ranked source models dominated the ensemble. As the blue crosses showed, the low-ranked source models poorly classified lapses in the target session (at ~25% accuracy). Therefore, over-recruiting source models could be destructive to performance as low-ranked source models might cause negative transfer. More importantly, the ranked model ensemble resulted in comparable performance with the WSA that used calibration data from another 90-min session of the same target subject. Furthermore, without effective ranking, the randomized model ensemble performed poorly, and the performance did not increase as more models were involved. The experimental results evidenced the effectiveness of the proposed model-ranking method that selected source models based on their generalizability and calibration-data similarity for this sample subject. Other subjects had comparable trends but different maximal performances for top-k models.

Next, we systematically investigated the effect of the size of source-model pool on the classification performance across all target subjects. To this end, we randomly reconstructed the source-model pool with a smaller size than the original pool size, and fused them according to the proposed weighting scheme (Eq. (3)) on the outputs from the source models. This procedure was repeated 20 times for each reduced size of the randomly reconstructed source-model pool. Fig. 3 exhibits the relationship between the TL performance and the size of source-model pool. The overall TL performance across 11 target subjects (red curve) monotonically increased with the size of the source-model pool, suggesting that large-scale existing data from other subjects could improve the TL performance if the optimization procedure was involved. The classification performance of the proposed large-scale TL approach reached and exceeded that of the conventional WSA approach when the pool size increased above 10 and 15, respectively, but there was no statistical significance; $p>0.05$). In addition, Fig. 3 also shows the performance of combing the selected source models with random weights (TL-random - gray curve), as opposed to using the proposed ranking with generalizability and similarity. The proposed TL method significantly outperformed the random TL method ($p<0.05$, paired t-test) as the size of source-model pool was greater than 30. Without considering the model generalizability and calibration-data similarity, negative transfer was inevitable due to the substantial EEG variability across subjects. Using model generalizability and calibration-data similarity, though might not guarantee the best selection, is capable of separating auxiliary source models that more likely lead to positive transfer from those dissimilar source models that might deteriorate TL performance.

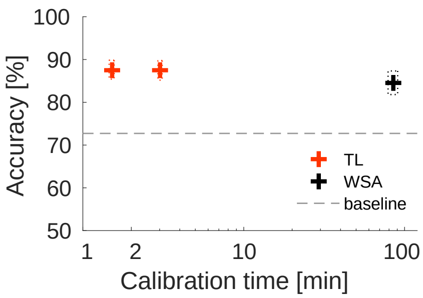

Fig. 4. A classification-performance comparison between the proposed TL and the conventional WSA in terms of both accuracy and calibration time across 11 target subjects who performed multiple sessions. Red and black crosses show the average accuracy and calibration time for TL and WSA respectively. Gray dashed line denotes the blind classification baseline (73.75%) since the amounts of data points between two classes were uneven. The calibration time was drastically reduced using the TL approach, whereas no significant difference in accuracy was observed between TL and WSA approaches.

Fig. 4 shows the classification performance of the TL approach in terms of accuracy and required calibration time, compared to that using within-subject cross-session classification (WSA). TL marginally outperformed the within-subject approach (87.62±7.32% vs. 84.55±9.12%, $p=0.24$ assessed by paired t-test) across 11 target subjects who had multiple sessions. Most importantly, TL approach required much fewer calibration data than that of the WSA (1.51±0.23 min vs. 85.97±22.57 min, $p<10^{-17}$). In addition, as the number of calibration trials for measuring subjects'/sessions' similarity was increased from the first 10 to the first 20 trials, no appreciable difference was observed in the TL performance. It is worth noting that a major drawback of other current TL approaches [2]-[4] and the WSA is that they require representative calibration data from all of the conditions (alert vs. lapse in our study) to build a reliable BCI model, which can be difficult or expensive to obtain as the subject might neither experience lapses until the later part of the experiment nor have sufficient numbers of lapses in the pilot session. In contrast, the proposed TL approach only needs minimum calibration data (~1.5 min) from the target subject under alert condition, which is undoubtedly more accessible in real-world applications.

V. Conclusion

This study is the first attempt to test the feasibility of leveraging large-scale existing data from other subjects while reducing the need for time-consuming calibration-data collection in BCIs. We found that the TL approach marginally outperformed the within-subject approach while considerably reducing the required calibration data for the target subject (first ~1.5 min vs. a 90 min pilot session used in the WSA). The proposed TL and optimization frameworks can facilitate numerous real-world applications, not limited to lapse detection, of BCIs.

References

[1] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, “Towards Zero Training for Brain-Computer Interfacing,” *PLoS ONE*, vol. 3, no. 8, p. e2967, Aug. 2008.

[2] S. Dalhoumi, G. Dray, and J. Montmain, “Knowledge Transfer for Reducing Calibration Time in Brain-Computer Interfacing,” in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2014, pp. 634–639.

[3] M. Arvaneh, I. Robertson, and T. E. Ward, “Subject-to-subject adaptation to reduce calibration time in motor imagery-based brain-computer interface,” in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2014, pp. 6501–6504.

[4] F. Lotte, “Signal Processing Approaches to Minimize or Suppress Calibration Time in Oscillatory Activity-Based Brain-Computer Interfaces,” *Proceedings of the IEEE*, vol. 103, no. 6, pp. 871–890, Jun. 2015.

[5] C.-S. Wei, Y.-P. Lin, Y.-T. Wang, T.-P. Jung, N. Bigdely-Shamlo, and C.-T. Lin, “Selective Transfer Learning for EEG-Based Drowsiness Detection,” in *2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2015, pp. 3229–3232.

[6] T.-P. Jung, S. Makeig, M. Stensmo, and T. J. Sejnowski, “Estimating alertness from the EEG power spectrum,” *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 1, pp. 60–69, Jan. 1997.

[7] C.-T. Lin, R.-C. Wu, S.-F. Liang, W.-H. Chao, Y.-J. Chen, and T.-P. Jung, “EEG-based drowsiness estimation for safety driving using independent component analysis,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 12, pp. 2726–2738, Dec. 2005.

[8] C.-S. Wei, Y.-T. Wang, C.-T. Lin, and T.-P. Jung, “Toward non-hair-bearing brain-computer interfaces for neurocognitive lapse detection,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 6638–6641.

[9] M. Congedo, A. Barachant, and A. Andreev, “A New Generation of Brain-Computer Interface Based on Riemannian Geometry,” *arXiv:1310.8115 [cs, math]*, Oct. 2013.

[10] R.-S. Huang, T.-P. Jung, A. Delorme, and S. Makeig, “Tonic and phasic electroencephalographic dynamics during continuous compensatory tracking,” *NeuroImage*, vol. 39, no. 4, pp. 1896–1909, Feb. 2008.

[11] S. K. L. Lal, A. Craig, P. Boord, L. Kirkup, and H. Nguyen, “Development of an algorithm for an EEG-based driver fatigue countermeasure,” *Journal of Safety Research*, vol. 34, no. 3, pp. 321–328, Aug. 2003.

[12] Y.-T. Wang, K.-C. Huang, C.-S. Wei, T.-Y. Huang, L.-W. Ko, C.-T. Lin, C.-K. Cheng, and T.-P. Jung, “Developing an EEG-based on-line closed-loop lapse detection and mitigation system,” *Front Neurosci*, vol. 8, p. 321, 2014.

[13] C.-C. Chang and C.-J. Lin, “LIBSVM: A Library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.