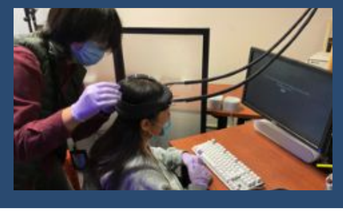
Examine Adversarial Domain Adaptation Model on Tufts fNIRS Dataset with Learning Invariant Representations and Risks

By Irene Chang

CS137: Deep Neural Networks – Tufts University



ABSTRACT

Brain-Computer Interfaces (BCIs) have been widely researched and developed towards assisting physically challenged users. Apart from this application, a growing research area BCIs revolves around examining brain activities of a wider population in daily tasks such as those involving the use of laptops or desktops using brain signals such as EEG (electroencephalogram) and fMRI (Functional magnetic resonance imaging) [2]. The main obstacle in the development of this technology is the wide variation of these signals across human users caused by the differences in cortical and other physiological features, as well as the lack of standardized data acquisition and evaluation protocol. As a part of this ongoing efforts to tackle this issue, a study at Tufts has been done to develop and evaluate generalizable machine learning methods built on a type of brain signals, fNIRS (functional near-infrared spectroscopy), to classify mental workload intensity levels. The outcome of this study suggests a lot of potentials for a generic model for this classification task, as well as indicating a future direction of transfer learning for cross-subject learning. Our study expands on this result to examine the performance of an adversarial domain adaptation technique, Learning Invariant Representations and Risks (LIRR) on fNIRS dataset. LIRR has proved to be effective for image classification tasks, and we want to see if this algorithm can also enhance deep neural networks built on time-series data, from which we can determine how adversarial learning can contribute to the current progress of building efficient BCIs.

INTRODUCTION

There are many studies on using machine learning to extract information about users' mental workload intensity levels from their fNIRS signals. Traditionally, experiments are needed to build individualized classifiers for each user, which proves to be exhaustive and causing lots of discomfort for participants. Moreover, the models trained on one subject often struggle to make accurate predictions on new subjects due to the variations in fNIRS signals across individuals. Generic classifiers and transfer learning models have hence been considered the potential solutions to reduce these efforts.

The Tufts study [1] performed a thorough evaluation of four different classification models, Random Forest, Logistic Regression, EEGNet and DeepConvNet, under both subject-specific and generic paradigms. The outcome suggests that a classification model trained on some subjects can be employed to predict for others. However, these generic classifiers still suffer from the cross-subject inherent variations as mentioned above. Thus, the study recommends domain adaptation in future study as a way to alleviate the gap in performance caused by these noises.

Li et al. [3] introduced LIRR, a domain adaptation technique for semi-supervised settings. The algorithm was tested on image classification tasks and evaluated against other deep learning frameworks such as DANN, CDAN, and ADDA. LIRR was showed to outperform other methods on the classification task when there is a limited number of labeled target data. Motivated by this finding, in this paper we will examine the effectiveness of LIRR mechanism on the fNIRS dataset in a cross-subject multi-classification task. We are going incorporate the LIRR framework, into EEGNet. EEGNet model without LIRR is our baseline for comparison and accuracy score is used as the evaluation metric.

If the LIRR model is successful, it can be employed for the development of efficient and precise BCIs for producing predictions of mental workload in real time. Additionally, we can save costs and effort of setting up the experiment for new subjects in order to get enough data for training new subject-specific models.

DATASET

Here we describe the data collection procedure done by. For each subject, the data collection produces an fNIRS recording lasting over 20 minutes (task duration). A sliding-window approach is applied to extract fixed-duration (2-40 sec.) windows. Experiments on our dataset suggest that 30 second windows give the highest accuracy. We extract overlapping windows with a stride of 0.6 seconds.

Thus, we have each window as one input unit to the deep learning model. The final dataset consists of 8 signal measurements that are used as the features for the model, and 1488 entries for each subject representing 1488 windows. There are 4 mental workload intensity levels that we want to predict {0, 1, 2, 3}, with a larger number representing more intensity.

BACKGROUND

Domain-Adversarial Neural Network (DANN):

A representation learning technique for when training and test time come from similar but different distributions. The method aims to learn a set of features that can discriminate between the objective labels for the main classification task but cannot discriminate between the training (source) and test (target) domains i.e, given an input observation, the model cannot identify its domain of origin. At the end, we obtain features that have the same or very similar distributions in the source and the target domains, called invariant representation, and they allow us to make good prediction across all domains.

To accomplish this task, beside the label predictor of the classification task, we also have a domain classifier that attempts to discriminate between the source and the target domains during training. We optimize the parameters to minimize the loss of the label classifier and to maximize the loss of the domain classifier. The latter update thus works adversarially to the domain classifier, and it encourages domain-invariant features to emerge in the course of the optimization [4].

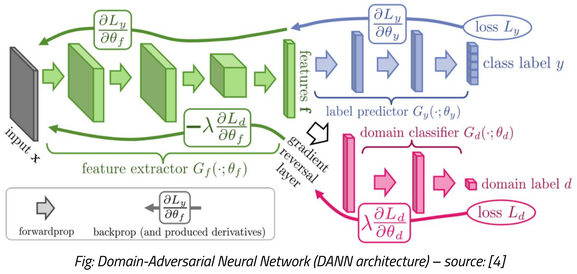


Fig: Domain-Adversarial Neural Network (DANN architecture) – source: [4]

Invariant Risk Minimization (IRM):

While machine learning are extracting complex prediction rules from data, they are inherently subject to biases and confounding factors. In trying to minimize training error, the algorithm can learn spurious correlations in the predictors. IRM thus aims to identify which properties of the training data describe spurious correlations and which properties represent the true causal relations of interest. However, IRM is not sufficient to ensure reduced accuracy discrepancy across domains, and thus we want to conduct representation learning simultaneously [3].

LIRR: combines representation learning and invariant risks learning. As a result, we are able to learn the common features that are causally related to our variable of interest across all domains and make better predictions.

METHODOLOGY

Training models: In order to prepare the dataset for the model, we follow and combine the procedures of both [1] and [2]. We randomly break the dataset from the source subject and the dataset from the target subject into 5 different splits: source training set, source validation set, target labeled set, target unlabeled set, target validation set. The split assignment is done using a random selection on the window indexes.

Then, the classification step of EEGNet from [1] is modified to match with LIRR framework by changing the design of the final classifier layer. The architecture of EEGNet, which consists of first 3 convolutional layers, is retained as the feature extractor of the network. The invariant risks learner (also called the environment predictor) and the domain classifier from LIRR will be piped into these layers. The total loss of the network will hence be calculated using the objective laid out in LIRR.

We use a set of fixed parameter to train the model, with learning rate 0.01, batch size of 32, 50 epochs. We keep track of the model that produces the lowest loss on the target test set, and obtain the accuracy score on the target test set using this model.

Evaluating models: Due to the randomness in splitting the data, in addition to the fact that there are overlapping windows, it may be the case that the training and the testing set of the target contain multiple overlaps and thus the true accuracy is overestimated. Therefore, we apply bootstrapping (30 times) to reduce this confounding factor.

EXPERIMENTS

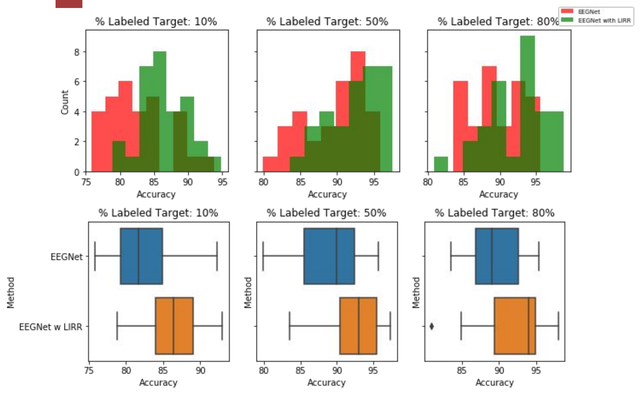


Fig: *Accuracy (%) distribution of EEGNet model and EEGNet with LIRR model on 10%, 50%, 80% labeled target data after bootstrapping. (Subject 7 - Subject 75)*

Method	10% labeled target	50% labeled target	80% labeled target
7 to 75			
EEGNet	82.57 ± 4.44	89.17 ± 4.06	89.50 ± 3.68
EEGNet with LIRR	86.27± 3.37	92.60 ± 3.40	92.27 ± 4.13
75 to 36			
EEGNet	74.35 ± 3.33	78.12 ± 4.20	78.75 ± 4.01
EEGNet with LIRR	77.73 ± 4.38	82.67 ± 3.47	82.57 ± 4.71

Table: *Accuracy (%) comparison (higher means better) of EEGNet model and EEGNet with LIRR model on 10%, 50%, 80% labeled target data (mean ± std).*

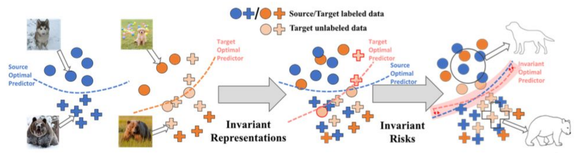


Fig: *Overview of the LIRR model. Learning invariant representations induces indistinguishable representations across domains, but there can still be misclassified samples (as stated in red circle) due to misaligned optimal predictors. Besides learning invariant representations, LIRR model jointly learns invariant risks to better align the optimal predictors across domains. [3]*

RESULTS & RECOMMENDATIONS

Results: The comparisons show that the mean accuracy scores are higher with the model that uses LIRR learning technique, for both cases. This suggests that LIRR can be applied on sequential data to improve cross-domain classification. We also see that the more labeled target is present, the better the learning model is, which is aligned with the intuition for semi-supervised domain adaptation. However, from the bootstrapped distributions and the variances of the accuracy scores, the model with LIRR are not distinctly better than the baseline models. LIRR produces the most improvement from the baseline model when there are less labeled target data. This gap decreases as we have more labels in the target domain, since the baseline method approaches a supervised training in such scenarios. More bootstrapping in addition to tuning the model can give us a more rigorous bounds of these algorithms' performance

Future directions:

- Include more sources: [1] and many other studies utilize data from multiple different individuals to increase the training set size, which is especially important in this study setting given that data collection is a significant barrier that has motivated many research in the field. The current LIRR algorithm is currently only tested on transferring from one domain to another. We can expand upon this method to include more subjects in a multi-source transfer learning model.
- Cosine: [2] suggests the use of cosine module to improve the performance of semi-supervised domain adaptation. This is showed in the same study to even further improve the results of both the image classification task and the regression task. We can apply this method on our fNIRS dataset to examine if this holds in the case of sequential data.
- Having recognized the potential of LIRR in improving the mental workload classification task, a next step will be to carry out an extensive tuning of the models to develop a more rigorous model that can be deployed in practice. This includes experimenting with other networks for domain classification and invariant risk learning tasks.

REFERENCES

[1] Zhe Huang and Liang Wang. The Tufts fNIRS to Mental Workload Dataset: Toward Brain-Computer Interfaces that Generalize. 2021. url: https://openreview.net/forum?id=QzNHE7QHhut.
[2] Xu Y Wu D and Lu B.-L. "Transfer learning for EEG-based brain-computer interfaces: A review of progress made since 2016". In: IEEE Transactions on Cognitive and Developmental Systems 14 (2022), pp. 4–19.
[3] Bo Li et al. "Learning Invariant Representations and Risks for Semi-supervised Domain Adaptation". In: CoRR abs/2010.04647 (2020). arXiv: 2010.04647. url: https://arxiv.org/abs/2010.04647.
[4] Yaroslav Ganin et al. "Domain-Adversarial Training of Neural Networks". In: Journal of Machine Learning Research 17.59 (2016), pp. 1–35. url: http://jmlr.org/papers/v17/15-239.html.