Ensemble Co-Training of Image and EEG-based RSVP Classifiers for Improved Image Triage

Steven Gutstein
Army Research Laboratory
Human Research and Engineering Directorate
Aberdeen Proving Ground, MD 21005
steven.m.gutstein.ctr@mail.mil

Vernon Lawhern
Army Research Laboratory
Human Research and Engineering Directorate
Aberdeen Proving Ground, MD 21005
vernon.j.lawhern.civ@mail.mil

Brent Lance
Army Research Laboratory
Human Research and Engineering Directorate
Aberdeen Proving Ground, MD 21005
brent.j.lance.civ@mail.mil

Abstract

Our ability to gather large sets of images far outstrips our ability to review them for targets of interest. Image triage techniques quickly filter image sets in order to more rapidly present the most relevant items to a human. In this paper, we will employ two fundamentally different such techniques - a computer vision classifier, based deep convolutional neural net (DCNN) with an SVM, and a Brain-Computer Interface (BCI) for electroencephalogram (EEG) classifier that relies upon detection of a P300 response with the xDAWN algorithm and a Bayesian linear discriminant. Because the classifiers are so dissimilar, they are excellent candidates for co-training. Due to their moderate precision, it is reasonable to use active learning to ensure no noisy data enters the training set via co-training. In this paper, we demonstrate that, by using co-training in conjunction with active learning, that it should be possible to augment training sets for an offline ensemble made up of image based and EEG based classifiers. These classifiers will form the core of an offline ensemble, which combines human and computer capabilities on two levels. A human generates a neural signal, indicating target recognition; a computer both classifies the EEG signal and directly classifies the source image; lastly, a human manually verifies the unlabelled samples used for co-training.

1 Introduction

Manually examining large volume imagery for infrequent targets is a challenging, yet common task. Examples range from scanning large numbers of MRIs for medical abnormalities to scanning large numbers of satellite images for various objects. Although crowd sourcing is a popular and generally effective technique for filtering through such large image collections, there are times when it is not appropriate, such as when expertise is required, when there are privacy concerns, or when there are security issues. For these situations, image triaging methods are an attractive alternative.

One way this can be done is to take advantage of the success of Deep Convolutional Neural Nets (DCNNs) in labelling images. Another is to use Brain-Computer Interfaces (BCI) to rapidly display images to a person suitably familiar with what would constitute a target, at a rate of about 5 Hz or more. This technique is known as Rapid Serial Visual Presentation (RSVP). As images are being shown to the person, an electroencephalogram (EEG) is used to record the person's neural activity.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

When an image containing a target of interest is shown to the person, a characteristic neural response called the P300 is elicited. The P300 is a waveform with a large positive deflection approximately 300ms post stimulus presentation. Rather than having the person manually identify images with the target object, their EEG is monitored for any sign of a P300 response. It is a promising area of active research for image triage [3, 12, 13]. These techniques not only look at different raw signals (raw image pixels vs. EEG signals) but also at their base use two very different classifiers - a pure computer classifier (DCNN), which directly classifies each image and human-computer classifier, which depends on a computer to classify a person's EEG response to each image.

The strength of any ensemble relies on two main characteristics - the base accuracy and the diversity of its members. The diversity of the ensemble members refers to the tendency of those members not to make identical errors on a given input. The stark difference in our classification algorithms and the raw data they work from ensures a high degree of diversity. To increase the accuracy of these algorithms, we seek to augment their training by allowing them to co-train. We will also augment their training by employing active learning. Before allowing co-training to occur on any sample labelled by either classifier, we will use a human to ensure that the sample has been correctly labelled.

In this paper, we will specifically examine co-training between two classifiers - an EEG-based classifier for detecting the P300 response produced by a human, engaged in Rapid Serial Visual Presentation (RSVP) (xDAWN + Bayesian LDA [11]) and a deep convolutional neural net (DCNN) in combination with a Support Vector Machine (SVM), which will directly classify images. The DCNN is a variation of the AlexNet architecture, which revolutionized the Imagenet competition[6]. It acts as a feature extractor for our SVM. We will refer to this combination classifier as a DCNN+SVM.

Although we will only be experimenting with a pair of classifiers, there is no reason not to use a larger ensemble of various EEG-based and image-based classifiers. The approaches we describe here readily extend to such cases. One interesting aspect of our ensemble of image-based and EEG-based classifiers is that the 'hard' part of training a EEG-based classifier is calibrating it for use by a particular subject. The 'hard' part of training a image-based is training it to recognize a previously unseen class of image. One can imagine that an ensemble of image-based and EEG-based classifiers will be able to use trained image-based to efficiently generate data for an EEG-based classifier for a new subject, and similarly the humans behind the EEG-based classifiers will be able to efficiently label images of new targets to train the image-based classifiers upon. The contributions of this paper include not only an improved method for constructing image triaging ensemble, but also a unique integration of human and machine agents for a classification task.

2 Background

2.1 Training Techniques

The "wisdom of crowds" is a colloquial way to acknowledge that a collection of independent classifiers generally achieves better results than any individual member of the collective. An important characteristic of a good ensemble is that the errors of its members are independent of each other. Otherwise, one would be best off just using the most accurate classifier in the ensemble. For machine learners, this independence is frequently obtained by exposing the individual learners to different training sets, using different algorithms to learn the same datasets and, for stochastic machine learners, using different initial conditions. These techniques are an attempt to create classifiers that are less likely to make the same mistakes.

One way to ensure independence of errors is to have each classifier use different, conditionally-independent data attributes to reach a decision. An example of this would be recognizing people using one classifier based upon facial recognition and another based upon voice recognition. These two classifiers would be very unlikely to make similar errors, since the features they rely upon are unrelated. Once we have guaranteed that our ensemble members will have mistakes that are as uncorrelated as possible, we can use the ensemble members to augment each other's training sets. This is the main idea behind Ensemble Co-Training [16]. Furthermore, we can improve the efficiency of the co-training by using active learning to remove any noisy data that co-training might introduce. Active Learning [14] entails letting a machine learner choose groups of unlabelled data, which would be most useful for it to have labelled for training. By having a human labeller verify any data elements being used for co-training, we not only avoid noise in the new members of the training set, but also correct the mistakes that the responsible classifier is most certain of.

2.1.1 Co-Training

Co-Training was first introduced by Blum & Mitchell [1] as a method to boost the performance of a classifier when there was only a small set of labelled samples available. By creating classifiers based upon different feature sets, or 'views', they envisioned the data space as being made up not of individual samples, but as loosely correlated sample pairs. Ideally, these samples would be conditionally independent of each other, given their classifications. When a pair of such classifiers (e.g. A & B) are given the paired data (x_a, x_b), then even if x_b is an unusual sample and correspondingly hard for B to classify, x_a is likely not as unusual and resultingly easier for A to classify. This may allow A to provide B with a new training sample in a region of data space where it (B) needs more training samples. Furthermore, using standard semi-supervised learning, A can also add that piece of data to its own training set to improve its own accuracy. This makes it very natural to use co-training to improve the accuracy of an ensemble by improving the accuracy of each specific classifier. Several authors have demonstrated this fact and have described this approach with terms such as 'Disagreement Based Semi-Supervised Learning' [19], 'Agreement-Boosting'[7] and 'Ensemble Co-Training' [17]. In this paper, we will use the term 'Ensemble Co-Training', since it succinctly highlights the two main techniques.

2.1.2 Active Learning

Another approach to improve classifier accuracy by increasing the amount of available labelled data, while minimizing human effort, is active learning. This approach differs from semi-supervised learning in that it relies on people to label additional data, but depends upon machines to select the most useful data to label. This usually involves identifying data elements which present the greatest degree of uncertainty for the classifier - i.e. those closest to a decision boundary [14]. Because the machine is uncertain about those points, training on them will have the greatest benefit for the accuracy of the function learned by the machine.

In our experiments, we insert active learning into the co-training process. Once a classifier is sufficiently confident of the label it assigns to a piece of data, but before it's actually added to the training set, a human verifies the prospective label. When the images are correctly labelled, this verification is akin to active learning for the classifier that was not confident of the appropriate label. This approach is very similar to Co-testing [8], which is a variation of Seung et al.'s[15] 'Query-By-Committee' approach to active learning. Since our data is markedly unbalanced, we are primarily interested in finding samples of the target class, which is under represented. So, we are only going to use a person to label data that at least 1 classifier views as belonging to the target class. Although it would reduce the workload on the human to automatically label data on which both ensemble members agree, this does not, at the moment, appear to be a significant savings, since there are extremely few such samples.

2.2 Dataset

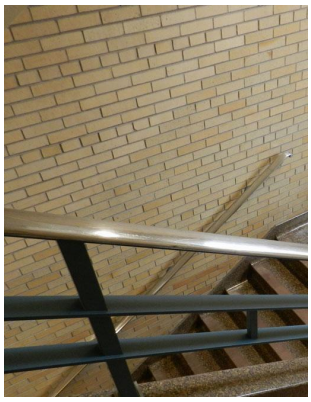
We used existing data from an RSVP experiment with 17 subjects and an image database of scenes in typical indoor/office environments. There were about 300 samples of each target image and about 3000 total images. The target objects belonged to one of five classes - chair, container, door, poster and stair. The target objects appear in a wide variety of distances from the observer, orientation with respect to the observer and partial occlusions. There is also a large variety of distractor (i.e. non-target) objects in each image. Two samples from this set are shown below in Fig 1. We will refer to the image database as the 'Office Object' dataset. Due to this dataset's size and complexity of the images, our DCNN+SVM was only able to achieve AUCs around .70.

Subjects performed five RSVP sessions, where in each session subjects were instructed to identify one of the five target classes against the non-target images. 256-channel EEG was recorded using a Biosemi ActiveTwo system, the RSVP presentation rate was 5 Hz and the percentage of targets varied between 5% - 15%. There would be 1562 non-target images and approximately 100 - 300 target images in each training session. For more information, see Touryan et al. [18].

2.3 Classifiers

The machine classifiers comprising our ensemble are an RSVP classifier, which uses the xDAWN algorithm [11] and a deep convolutional neural net (DCNN) using the bvlc_reference_caffenet

(a) Sample image of chair



(b) Sample image of stairs

Figure 1: Sample images with targets (i.e. chair & stair) from the Office Object dataset - note the oblique views and partially occluded targets.

variation of the AlexNet architecture, which was pretrained on the Imagenet dataset. This net acts as a feature extractor for an SVM, that directly identifies images containing target objects. We take this approach for several reasons. Firstly, our dataset is not sufficiently large to effectively train a DCNN, so if we assume that the various visual features that the DCNN learned to use in order to classify images from the Imagenet data are general visual features, then the lower layers of such a net should be an effective feature extractor. Secondly, because we repeatedly retrain our classifier as we increase the number of labelled samples of targets, it's more computationally tractable than repeatedly retraining a DCNN from scratch.

2.3.1 EEG Classification

A major problem in reading EEG signals is dealing with a low Signal-To-Noise Ratio (SNR). The P300 response that we need to detect is obscured by electrical signals ranging from those generated by the subject's body to those in the external environment. We will accentuate the P300 response using the xDAWN algorithm, first introduced by Rivet et al. [11]. xDAWN is a supervised learning algorithm that estimates a set of spatial filters that improves the SNR of the P300 response. Once the data has been filtered, we train a Bayesian Linear Discriminant Analysis classifier (BLDA) following the approach of Hoffmann et al. [4] This will form the basis of our neural classifier, which we will refer to as an xBLDA classifier.

2.3.2 Deep Learners

Deep learners, in particular, deep convolutional neural nets, have achieved notable success in labelling images. However, training these nets is still an extremely time consuming process, requiring huge training sets. Due to the small size of the Office Object dataset, we used a net that had already been trained on the Imagenet dataset - bvlc_reference_caffe_net, as shown in Fig. 2.

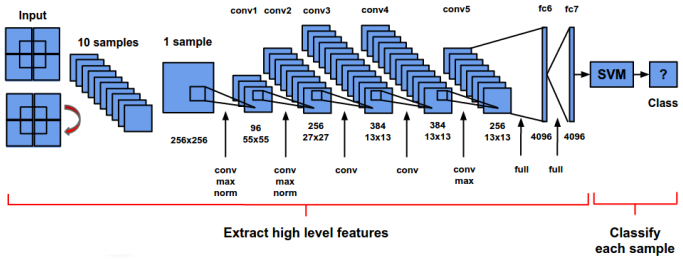


Figure 2: Diagram of the DCNN+SVM classifier. The lower DCNN is a variant of AlexNet and acts as a feature extractor. The upper layer is an SVM that is trained to classify the extracted features. [5]

In order to transfer learning acquired with the Imagenet dataset, to the Office Object dataset, we discarded the output layer of the neural net and trained an SVM to identify our dataset's classes based upon the output of the nets' penultimate layer. When we refer to training the DCNN+SVM, we are only training the SVM, which classifies the DCNN's outputs.

2.4 Related Work

Detection of the P300 response has already been used as the basis of Brain-Computer Interface (BCI) Spellers, such as the one introduced by Farwell and Donchin [2]. It has also been used studied for more general purposes to allow disabled individuals to interact more easily with computers and their environment [4]. More specifically, it has also been used for image triage by several authors[3, 13, 12], among others.

In many ways EEG-based BCI problems both need and are particularly suited for co-training. The difficulty of assembling large sets of EEG data means that any sort of learning with EEG data will require the greatest possible efficiency in using that data. However, the variety of ways in which EEG data can be processed means that creating different feature sets on which to train is especially straight-forward. Two examples of co-training for EEG-based classifiers are Panicker et al.[9], who used two different classification techniques (Fisher Linear Discriminant analysis & Bayesian Linear Discriminant Analysis) for co-training in order to train a P300 speller and Ren et al.[10], who also use two different classifiers (Biomimetic Pattern Recognition - a shallow neural net with a novel activation function, and Sparse Representation, which is a dictionary learning approach) for motor imagery recognition. Our work differs most significantly in that it employs two truly different feature sets, raw image pixels and EEG signals, to classify images. The use of two such dissimilar feature sets is combined with two very dissimilar classification algorithms - DCNN+SVM and xDAWN + Bayesian Linear Discriminant Analysis (xBLDA).

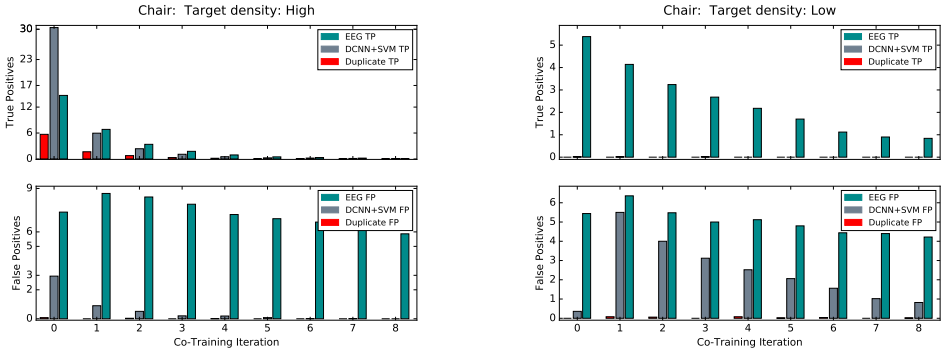
3 Experiments

In order to train the DCNN+SVM and xBLDA classifiers, we divided these images and their corresponding EEG signals into a training, validation and testing sets using a 50%/25%/25% split. On average we would have a training set with about 887 images, of which about 106 were targets, and validation and testing sets with an average of 444 images, of which about 53 were targets. This is too small a dataset for training a DCNN from scratch. However, our use of a pretrained DCNN allows us to achieve reasonable results with far fewer images than training a DCNN from scratch would require.

Co-Training involved letting our DCNN+SVM and xBLDA classifiers learn the training set. Then, using the validation set, each classifier picked out any images in which it had at least 90% confidence contained the target being sought. The resulting images were then filtered so that only images actually containing targets remained (thus, simulating the use of a active learning oracle). These images were then added to the training set of each classifier and the process was repeated 10 times, or until no new targets were correctly identified. Classifier performance was measured with the AUC achieved by each classifier on a separate testing set.

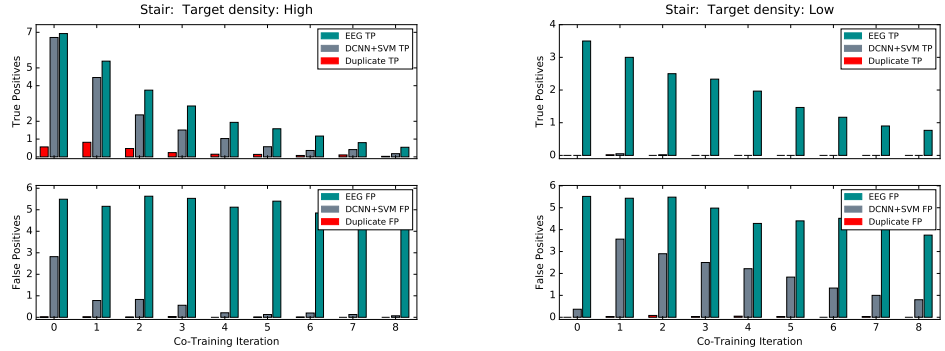
For each subject, we ran 10 trials for each target object using different random splits of their data into training, validation and testing sets. At the end of each trial, images from the validation set, which either classifier was confident contained a target were verified by a human. If the images did contain a target, then they were removed from the validation set and added to the training set for the next trial. We tracked these images in order to verify that the two classifiers really were selecting different images in general.

Because the responsiveness of a subjects neural signals could vary due to target density, we divided subjects into 2 groups - those who saw a high percentage of targets (10% - 15%) and those who saw a low percentage of targets (5% - 10%). The effectiveness of co-training on each pair of classifiers for a given subject was evaluated by observing the median change in AUC over the 10 trials. The degree to which co-training occurred was measured by observing the number of target images that were correctly identified in the validation set by each classifier and the degree to which there was overlap in the images each classifier confidently labelled.



(a) Target: Chair with high target density

(b) Target: Chair with low target density



(c) Target: Stair with high target density

(d) Target: Stair with low target density

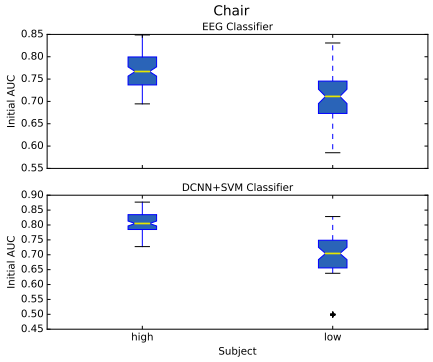
Figure 3: True and False Target Detection by EEG and DCNN+SVM classifiers. The red bars indicate the number of images that both classifiers identified; the gray bars indicates the number identified by the DCNN+SVM and the green bars indicate the numbers selected by the EEG classifier. The ordering of the bars for each iteration is red/gray/green.

4 Results

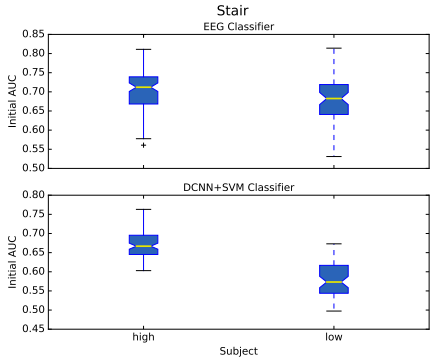
The main observations we have when observing the detection of true positives and false positives by the two classifiers are:

1. There is very little overlap between the images confidently identified as containing targets by the two classifiers, so we are likely operating in a realm where co-training will be a productive strategy
2. Our classifiers initially identify targets with a precision approaching 50%, which decreases over the various training iterations. This may be seen as reflecting the extreme imbalance of our datasets. Each iteration removes a significant fraction of the target images, but not of the non-target. Since there are only slight changes in the probabilities of true and false detections, our precision decreases. This seems to indicate that the further a particular target image is from being confidently identified, that more training images are needed to confidently identify it.
3. At low target frequencies, the EEG classifier is able to begin confidently identifying target images with fewer training samples
4. At high target frequencies, both classifiers do markedly better, but the DCNN+SVM classifier outperforms the EEG classifier

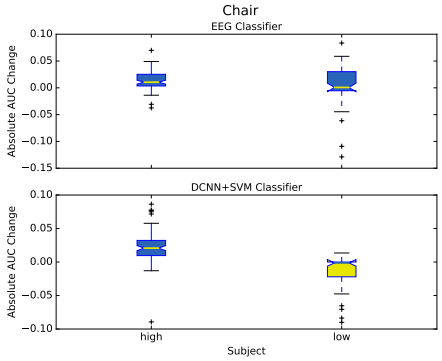
These points are illustrated in Fig 3, which shows our results for chairs and stairs. The other target objects gave similar results.



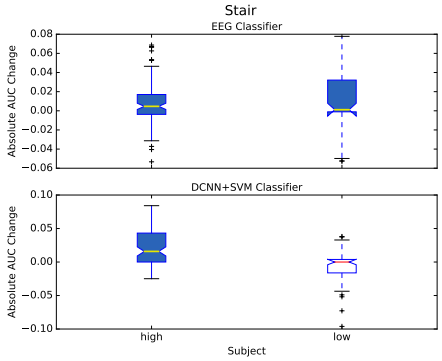
(a) Target: Chair - Initial AUCs for high & low target density



(b) Target: Stair - Initial AUCs for high & low target density



(c) Target: Chair - AUC Changes



(d) Target: Stair - AUC Changes

Figure 4: Median Initial AUCs and Changes in AUC after co-training EEG and DCNN+SVM classifiers. A blue box indicates the median(gold) is positive; a gold box indicates the median (blue) is negative; a clear box indicates the median (red) is zero. The notches in each box indicate a 95% confidence interval for the median. We only see statistically significant improvement in the median of the AUCs for the high density experiments.

Given our relatively small datasets it is to be expected that we will have moderate AUCs and similarly small improvements in the AUCs. This is shown in Fig 4, which displays notched box-whiskers plots of the AUC's achieved by our two classifiers and the co-training improvements we experienced. It is important to observe, that although the improvements were small, for the high target frequency sets, the 95% confidence interval for those improvements is entirely positive.

5 Conclusions & Future Work

We observe that based upon the images selected by the two classifiers as being very likely to contain a target, that their differing raw inputs lead to different approaches to the problem of target location. The degree to which these differences arise from the different machine learning algorithms (i.e. DCNN+SVM & xBLDA) and the degree to which they arise from the fact that machine learners still approach image recognition problems differently than people do is open to speculation. However, it strongly indicates that our co-training approach should help to make an ensemble consisting of these classifiers more effective. The improvements we've seen, though generally positive have been small. This is likely due to the size of our dataset. We are currently working with a similar dataset that is considerably larger and anticipate more significant improvements.

References

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[2] Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.

[3] Adam D Gerson, Lucas C Parra, and Paul Sajda. Cortically coupled computer vision for rapid image search. *IEEE Transactions on neural systems and rehabilitation engineering*, 14(2):174–179, 2006.

[4] Ulrich Hoffmann, Jean-Marc Vesin, Touradj Ebrahimi, and Karin Diserens. An efficient p300-based brain–computer interface for disabled subjects. *Journal of Neuroscience methods*, 167(1):115–125, 2008.

[5] Jeremy Karnowski. Alexnet + svm, 2015. Online; Oct. 10, 2016.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] Boaz Leskes and Leen Torenvliet. The value of agreement a new boosting algorithm. *J. Comput. Syst. Sci.*, 74(4):557–586, June 2008.

[8] Ion Muslea, Steven Minton, and Craig A Knoblock. Selective sampling with redundant views. In *AAAI/IAAI*, pages 621–626, 2000.

[9] Rajesh C Panicker, Sadasivan Puthusserypady, and Ying Sun. Adaptation in p300 brain–computer interfaces: A two-classifier cotraining approach. *IEEE Transactions on Biomedical Engineering*, 57(12):2927–2935, 2010.

[10] Yuanfang Ren, Yan Wu, and Yanbin Ge. A co-training algorithm for eeg classification with biomimetic pattern recognition and sparse representation. *Neurocomputing*, 137:212–222, 2014.

[11] Bertrand Rivet, Antoine Souloumiac, Virginie Attina, and Guillaume Gibert. xdawn algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Transactions on Biomedical Engineering*, 56(8):2035–2043, 2009.

[12] Paul Sajda, Eric Pohlmeyer, Jun Wang, Barbara Hanna, Lucas C Parra, and Shih-Fu Chang. Cortically-coupled computer vision. In *Brain-Computer Interfaces*, pages 133–148. Springer, 2010.

[13] Paul Sajda, Eric Pohlmeyer, Jun Wang, Lucas C. Parra, Christoforos Christoforou, Jacek Dmochowski, Barbara Hanna, Claus Bahlmann, Maneesh Kumar Singh, and Shih-Fu Chang. In a blink of an eye and a switch of a transistor: Cortically coupled computer vision. *Proceedings of the IEEE*, 98(3):462–478, 2010.

[14] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[15] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 287–294, New York, NY, USA, 1992. ACM.

[16] Jafar Tanha, M Someren, Hamideh Afsarmanesh, et al. Ensemble based co-training. In *BNAIC*, number 23, pages 223–231, 2011.

[17] Jafar Tanha, Maarten van Someren, and Hamideh Afsarmanesh. Disagreement-based co-training. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pages 803–810. IEEE, 2011.

[18] Jon Touryan, Gregory Apker, Brent J Lance, Scott E Kerick, Anthony J Ries, and Kaleb McDowell. Estimating endogenous changes in task performance from eeg. *Frontiers in Neuroscience*, 8:155–155, 2014.

[19] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.