

On the Metrics and Adaptation Methods for Domain Divergences of sEMG-based Gesture Recognition

István Ketykó^a and Ferenc Kovács^b

Nokia Bell Labs, Budapest, Hungary

Keywords: Machine Learning, Time-series Modeling, sEMG/EMG, Divergence Metrics, Domain Adaptation.

Abstract: We propose a new metric to measure domain divergence and a new domain adaptation method for time-series classification. The metric belongs to the class of probability distributions-based metrics, is transductive, and does not assume the presence of source data samples. The 2-stage method utilizes an improved autoregressive, RNN-based architecture with deep/non-linear transformation. We assess our metric and the performance of our model in the context of sEMG/EMG-based gesture recognition under inter-session and inter-subject domain shifts.

1 INTRODUCTION

Machine Learning (ML) is widely used for several tasks with time-series and biosensor data such as for human activity recognition, electronic health records data-based predictions (Ismail Fawaz et al., 2019), and real-time biosensor-based decisions. Various classification goals are addressed related to electrocardiography (ECG) (Jambukia et al., 2015), electroencephalography (EEG) (Craik et al., 2019; Dose et al., 2018), and electromyography (EMG) (Ketykó et al., 2019; Hu et al., 2018; Patricia et al., 2014; Du et al., 2017).

Sensing hand gestures can be done by means of wearables or by means of image or video analysis of hand or finger motion. A wearable-based detection can physically rely on measuring the acceleration and rotations of our body parts (arms, hands or fingers) with Inertial Measurement Unit (IMU) sensors or by measuring the myo-electric signals generated by the various muscles of our arms or fingers with EMG sensors. Surface EMG (sEMG) records muscle activity from the surface of the skin which is above the muscle being evaluated. The signal is collected via surface electrodes.

We are interested in sEMG-sensor placement to the forearm and performing hand gesture recognition with ML. In this context, all ML prediction models suffer from inter-session and inter-subject domain

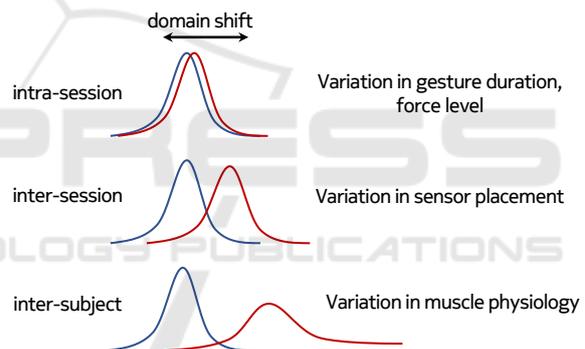


Figure 1: Domain shift in case of different scenarios.

shifts (see Figure 1).

- *Intra Session Scenario*: the device is not removed, and the training and validation data are recorded together in the same session of the same subject. In this situation the gesture recognition accuracy is generally above 90%.
- *Inter-session Scenario*: the device is reattached, and the validation data is recorded separately in a new session of the same subject. Under this domain shift the validation accuracy degrades below 50%.
- *Inter-subject Scenario*: The validation data is on another subject. In this case, the validation accuracy degrades below 50% as well.

Our focus is to investigate: 1) the metrics of these domain discrepancies, and 2) the adaptation solutions with special attention on those, which do not rely on

^a <https://orcid.org/0000-0003-4931-4580>

^b <https://orcid.org/0000-0002-9777-0372>

source data samples.

This paper is organized as follows, Section 2 provides a summary of ML, model risks, domains, domain divergences, and domain adaptation methods. Then our source data-absent metric and adaptation model is introduced in Section 3. Next, we validate our approaches using publicly available sEMG datasets: the experimental setup and results are described in Section 4. Finally, we conclude and summarize our results.

2 RELATED WORK

2.1 Machine Learning

At the most basic level, ML seeks to develop methods for computers to improve their performance at certain tasks on the basis of observed data. Almost all ML tasks can be formulated as making inferences about missing or latent data from the observed data. To make inferences about unobserved data from the observed data, the learning system needs to make some assumptions; taken together these assumptions constitute a model. The probabilistic approach to modeling uses probability theory to express all forms of uncertainty. Since any sensible model will be uncertain when predicting unobserved data, uncertainty plays a fundamental part in modeling. The probabilistic approach to modeling is conceptually very simple: probability distributions are used to represent all the uncertain unobserved quantities in a model (including structural, parametric and noise-related) and how they relate to the data. Then the basic rules of probability theory are used to infer the unobserved quantities given the observed data. Learning from data occurs through the transformation of the prior probability distributions (defined before observing the data), into posterior distributions (after observing data) (Ghahramani, 2015).

We define an *input space* \mathcal{X} which is a subset of d -dimensional real space \mathbb{R}^d . We define also a random variable X with probability distribution $P(X)$ which takes values drawn from \mathcal{X} . We call the realisations of X feature vectors and noted x_i .

A generative model describes the marginal distribution over X : $P_{\Theta}(X)$, where samples x_i of X are observed at learning time in a dataset D and the probability distribution depends on some unknown parameter Θ . A generative model family which is important for time-series analysis is the autoregressive one. Here, we fix an ordering of the variables X_1, X_2, \dots, X_n and the distribution for the i -th random variable depends on the values of all the preceding random variables

in the chosen ordering X_1, X_2, \dots, X_{i-1} (Bengio et al., 2015). By the chain rule of probability, we can factorize the joint distribution over the n -dimensions as:

$$P_{\Theta}(X) = \prod_{i=1}^n P_{\Theta}(X_i | X_1, X_2, \dots, X_{i-1}) \quad (1)$$

We define also an *output space* \mathcal{Y} and a random variable Y taking values y_i drawn from \mathcal{Y} with distribution $P(Y)$. In the supervised learning setting Y is conditioned on X (i.e., $Y \sim P(Y | X)$), so the joint distribution $P(X, Y)$ is actually $P(X)P(Y | X)$.

A discriminative model is related to the posterior conditional probability distribution: $P_{\Theta}(Y | X)$ and it reflects straight the discrimination/classification task with lower asymptotic errors than the generative models. This transductive learning setting has been introduced by Vapnik (Ng and Jordan, 2001).

For all modeling approaches, the learning is to fit their distributions over the observed variables x_i in our dataset D . With other words (of a statistician), a good estimate of the unknown parameter Θ would be the value of Θ that maximizes the likelihood of getting the data we observed in our dataset D . Formally, the goal of the Maximum Likelihood Estimation (MLE) is to find the Θ :

$$\Theta = \underset{\Theta}{\operatorname{argmax}} P_{\Theta}(D) = \underset{\Theta}{\operatorname{argmax}} \sum_{x_i \in D} \log P_{\Theta}(x_i) \quad (2)$$

Learning in neural networks is about solving optimization problems. In case of probabilistic (and differentiable) cost functions, the backpropagation method (Rumelhart et al., 1986) is an estimator for MLE.

2.2 Risk of Machine Learning Models

We start with the general and probabilistic description then we move on to the discriminative models.

Let us introduce the concept of *loss function*. A loss function $L(x_i, P_{\Theta})$ measures the loss that a model distribution P_{Θ} makes on a particular instance x_i . Our goal is to find the model P_{Θ} that minimizes the expected loss or *risk*:

$$R(P_{\Theta}) = \mathbb{E}_{x_i \sim P(X)} [L(x_i, P_{\Theta})] \quad (3)$$

Note that the loss function which corresponds to MLE is the log loss $-\log P_{\Theta}(x_i)$. The risk evaluated on D is the in-sample error or *empirical risk*:

$$\hat{R}(P_{\Theta}) = \frac{1}{|D|} \sum_{x_i \in D} L(x_i, P_{\Theta}) \quad (4)$$

The generalization gap is the $R(P_{\Theta}) - \hat{R}(P_{\Theta})$ and our model intends to have small probability that this difference is larger than a very small value ϵ .

In case of supervised learning setting, we have a paired dataset of observations $D = \{(x_1, y_1), \dots, (x_m, y_m)\}$. The loss function becomes the conditional log-likelihood $L(x_i, y_i, P_\Theta) = -\log P_\Theta(y_i | x_i)$ and the risk is described as:

$$R(P_\Theta) = \mathbb{E}_{(x_i, y_i) \sim P(X, Y)} [L(x_i, y_i, P_\Theta)] \quad (5)$$

Let us introduce the *hypothesis space* \mathcal{H} which is the set/class of predictors $\{h : \mathcal{X} \rightarrow \mathcal{Y}\}$. A hypothesis h estimates the target function $f : \mathcal{X} \rightarrow \mathcal{Y}$ from D . The target function f is a proxy for the conditional distribution $P(Y | X)$, such as: $y_i = f(x_i) + \zeta$, where ζ is the noise term. Substituting h into Equation (5) we get the transductive version of the risk:

$$R^{td}(h) = \mathbb{E}_{(x_i, y_i) \sim P(X, Y)} [L(y_i, h(x_i))] \quad (6)$$

Substituting h into Equation (4):

$$\hat{R}^{td}(h) = \frac{1}{|D|} \sum_{(x_i, y_i) \in D} L(y_i, h(x_i)) \quad (7)$$

In the transductive setting the generalisation gap can be quantified:

$$\Pr[\sup_{h \in \mathcal{H}} |R^{td}(h) - \hat{R}^{td}(h)| > \varepsilon] \quad (8)$$

2.3 Domain and Domain-shift Concepts

First, it is necessary to clarify what a domain is and what kind of domain discrepancies there can be. There are several good survey papers that describe this field deeply e.g., (Kouw, 2018), (Kouw and Loog, 2019), and (Csurka, 2017). In this paper, the domain adaptation-related problem statement and notations follow (Kouw, 2018).

The problem statement is introduced from a classification point of view to simplify the definitions, but it can be generalized to other supervised machine learning tasks. A domain contains three elements: *Input space* \mathcal{X} , *Output space* \mathcal{Y} and $P(X, Y)$ joint distribution over \mathcal{X} and \mathcal{Y} .

Two domains are different if at least one of their above mentioned components are not equal. In case of DA the input spaces and output spaces of the domains are the same but the distributions are different. More general cases belong to different fields of transfer learning, a detailed taxonomy of transfer learning tasks can be found in (Pan and Yang, 2010). During DA there is a trained machine learning model on a so-called source domain (S) and there is an intent to apply it on a target domain (T). From this point onwards S and T in subscript refer to source and target domains.

Let us analyze the risk of a source classifier (Equation (6)) on a target domain T in the cross-domain setting:

$$\begin{aligned} R_T(h) &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_T(x, y) dx \\ &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(x, y)}{P_S(x, y)} dx \quad (9) \end{aligned}$$

It can be seen that the ratio of the source and target joint distributions ($P_T(X, Y)/P_S(X, Y)$) defines the risk $R_T(h)$. The investigation of this ratio allows us to define domain shift cases (Moreno-Torres et al., 2012): prior shift, covariate shift and concept shift.

In case of prior shift, the marginal distribution of the labels are different between the source domain and the target domain $P_S(Y) \neq P_T(Y)$, but the conditional distributions are equal $P_S(X | Y) = P_T(X | Y)$. Typical example for prior shift: the symptoms of a disease are usually population independent but the distribution of the diseases is population dependent. These conditions allow us to simplify the risk:

$$\begin{aligned} R_T(h) &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(x, y)}{P_S(x, y)} dx \\ &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(x | y) P_T(y)}{P_S(x | y) P_S(y)} dx \\ &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(y)}{P_S(y)} dx \quad (10) \end{aligned}$$

This means, that the complete labeled dataset from the target domain is not needed but the estimation of the marginal distribution of the labels $P_T(Y)$ is needed on the target domain.

Covariate shift is a well-studied domain shift, for further reference see (Kouw, 2018). It is defined as follows: the posterior distributions are equivalent, this means $P_T(Y | X) = P_S(Y | X)$, but the marginal distributions of the samples are different $P_S(X) \neq P_T(X)$. The typical cause of covariate shift is the sample selection bias. Only the sample distributions determine the risk:

$$\begin{aligned} R_T(h) &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(x, y)}{P_S(x, y)} dx \\ &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(y | x) P_T(x)}{P_S(y | x) P_S(x)} dx \\ &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(x)}{P_S(x)} dx \quad (11) \end{aligned}$$

In case of concept shift, the marginal distributions of input vectors are similar on both source and target domains $P_S(X) = P_T(X)$, on the other hand, the

posterior distributions differ from each other $P_T(Y | X) \neq P_S(Y | X)$. Usually, non-stationary environment causes this data drift (Widmer and Kubat, 1996). It is not possible to simplify significantly the cross-domain risk and the DA cannot be done without labeled target data:

$$\begin{aligned} R_T(h) &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(x, y)}{P_S(x, y)} dx \\ &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(y | x) P_T(x)}{P_S(y | x) P_S(x)} dx \\ &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} L(h(x) | y) P_S(x, y) \frac{P_T(y | x)}{P_S(y | x)} dx \quad (12) \end{aligned}$$

In general, none of the above mentioned assumptions is valid, thus it is not possible to simplify the risk on target domain. The differing posterior distributions cause the major domain shift related issues. The optimal transport approach assumes that there is transport $t()$ that satisfies $P_T(Y | t(X)) = P_S(Y | X)$ (Courty et al., 2016). Finding this transportation map is intractable but it is possible to relax it to a simpler optimization problem, where $t()$ is estimated via a Wasserstein distance minimization between the two domains (Courty et al., 2016; Kouw and Loog, 2019).

2.4 Divergence Metrics and Theoretical Bounds

As the input space and output space are common in case of DA, the distance and divergence metrics of the distributions can measure and quantify the domain discrepancies. We elaborate the most common metrics in the field of DA.

The Kullback-Leibler divergence (Cover and Thomas, 1991) is a well-known information theory-based metric between two distributions. It measures the relative entropy between two distributions. One of its main disadvantage is, that it is difficult to calculate it from samples in some cases (Ben-David et al., 2010).

$$\begin{aligned} D_{KL}(P_S || P_T) &= \int_{\mathcal{X} \times \mathcal{Y}} P_S(x, y) \log \left(\frac{P_S(x, y)}{P_T(x, y)} \right) dx dy \\ &= \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} P_S(x, y) \log \left(\frac{P_S(x, y)}{P_T(x, y)} \right) dx \quad (13) \end{aligned}$$

In general, the KL divergence is an asymmetric metric as $D_{KL}(P_S || P_T) \neq D_{KL}(P_T || P_S)$. A commonly used symmetric version is the Jensen-Shannon divergence (Lin, 1991). It measures the total divergence from the

average divergence.

$$\begin{aligned} M &= \frac{P_S + P_T}{2} \\ D_{JS}(P_S || P_T) &= \frac{1}{2} (D_{KL}(P_S || M) + D_{KL}(P_T || M)) \quad (14) \end{aligned}$$

The origin of the Wasserstein distance is the optimal transport problem: a distribution of mass should be transported to another distribution of mass with minimal cost. Usually Wasserstein-1 distance is used with the Euclidean distance (Arjovsky et al., 2017).

$$D_W(P_S, P_T) = \inf_{\gamma \in \Pi(P_S, P_T)} \mathbb{E}_{(s, t) \sim \gamma} [\|s - t\|] \quad (15)$$

where $\Pi(P_S, P_T)$ is the set of all joint distributions with marginals P_S and P_T . This distance metric allows us to construct a continuous and differentiable loss function (Arjovsky et al., 2017). In case of DA, this distance is calculated between the marginal distributions $P_S(X)$ and $P_T(X)$ to get a tractable problem (Kouw, 2018).

The H divergence allows to find upper bound to cross-domain risk (Kifer et al., 2004; Ben-David et al., 2006; Ben-David et al., 2010). The definitions and formulas are provided for binary classification because of simplification, but they can be generalized to multi-class problems, as well:

$$\begin{aligned} D_H(P_S, P_T) &= \\ &= 2 \sup_{h \in \mathcal{H}} |\Pr(h(x) = 1) - \Pr(h(x) = 1)| \quad (16) \end{aligned}$$

(Ben-David et al., 2006) provide two different techniques to estimate H divergence: from finite sample and from empirical risk of domain classifier. If the hypothesis space is symmetrical, the empirical H divergence can be calculated from finite samples of the source and target domains:

$$\begin{aligned} \hat{D}_H(P_S, P_T) &= \\ &= 2 \left(1 - \min_{h \in \mathcal{H}} \left[\frac{1}{m} \sum_{i=1}^m I[h(x_{S,i}) = 1] + \frac{1}{m'} \sum_{i=1}^{m'} I[h(x_{T,i}) = 0] \right] \right) \quad (17) \end{aligned}$$

where I is an indicator function which gives 1 if predicate is correct, otherwise 0. For the computation of $\hat{D}_H(P_S, P_T)$ during the minimization, the whole hypothesis space \mathcal{H} must be tackled. (Ben-David et al., 2006) introduced an approximation to empirical H divergence, which is called Proxy-A Distance:

$$\hat{D}_A(P_S, P_T) = 2(1 - 2\hat{R}_{S/T}), \quad (18)$$

where $\hat{R}_{S/T}$ is the empirical risk of a linear domain classifier, which is trained (in a supervised fashion) to distinguish the source and target domains.

The cross-domain risk can be estimated by the empirical H divergence:

$$R_T(h) \leq R_S(h) + R_{S+T}(h^*) + \hat{D}_H(P_S, P_T) + C(\mathcal{H}), \quad (19)$$

where $C(\mathcal{H})$ is a complexity measure of hypothesis space, $R_{S+T}(h^*)$ is the risk of the so-called *single good hypothesis*. The h^* is the best classifier that can generalize on both domains:

$$\begin{aligned} R_{S+T} &= R_S(h) + R_T(h) \\ h^* &= \operatorname{argmin}_{h \in \mathcal{H}} (R_{S,T}(h)) \end{aligned} \quad (20)$$

The minimization of H divergence gives a better result, however the the risk of a single good hypothesis can ruin the performance of the DA. In other words, if there is no single good hypothesis, the domains are too far from each other to build an efficient DA.

2.5 Domain Adaptation Techniques

All the adaptive learning strategies focus on identifying how to leverage the information coming from both the source and target domains. Incorporating exclusively the target domain information is disadvantageous because sometimes there is no labelled targeted data at all, or typically the amount of the labelled target data is small. Building on information present in the source domain and adapting that to the target is generally expected to be the superior solution (Patricia et al., 2014).

We make a split in the viewpoint of source-sample availability at DA time. We discuss separately methods that assume source sample availability and methods that do not, first generally, and later in the context of sEMG-based gesture recognition.

2.5.1 Source Data-based

The majority of the approaches incorporate the unlabeled source data samples at DA time. Cycle Generative Adversarial Network (Cycle-GAN) (Zhu et al., 2017) is a state-of-the-art deep generative model which is able to learn and implicitly represent the source and target distributions to pull them close together in an adversarial fashion. It is composed of two Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and learns two mappings ($A : T \rightarrow S$ and $B : S \rightarrow T$) to achieve the cycle-consistency between the source and target distributions ($A(B(S)) \approx S$ and $B(A(T)) \approx T$) via the minimax game.

Besides the GANs, the autoassociative Auto Encoder (AE) models are capable of building domain-invariant representations in their latent space. The non-linear Denoising AE (DAE) (Glorot et al., 2011)

builds strong representation of the input distribution with the help of mastering to denoise the input (augmented with noise or corruptions). As a side effect, the multi-domain input ends up with a domain-invariant latent representation in the model. Inspired by the DAE, a linear counterpart: the Marginalized Denoising AE (mDA) (Chen et al., 2012) has been proposed to keep the optimization convex with closed-form solution and achieve orders-of-magnitude faster computation (at the expense of the representation power limited to be linear).

Data augmentation with marginalized corruptions has been studied for the transductive learning setting (Ng and Jordan, 2001) also. The Marginalized Corrupted Features (MFC) classifier (van der Maaten et al., 2014) has strong performance in case of validation under domain shift. In particular, as the corrupting distribution may be used to shift the data distribution in the source domain towards the data distribution in the target domain - potentially, by learning the parameters of the corrupting distribution using maximum likelihood.

In the transductive learning setting a classifier can be explicitly guided to learn a domain-invariant representation of the conditional distribution of class labels $P(Y | X)$ among two domains. Domain-Adversarial Neural Network DANN (Ganin et al., 2016) adversarially connects a binary domain classifier into the neural network directly exploiting the idea exhibited by Equation (19).

The binary domain classifier of the DANN (Ganin et al., 2016) and the mDA (Chen et al., 2012) have been paired in (Clinchant et al., 2016) to get DA regularization for the linear mDA model. Hence, the mDA has been explicitly guided to develop a latent representation space which is domain-invariant. Linear classifiers built in that latent space have had comparable performance results in several image classification tasks.

The 2-Stage Weighting framework for Multi-Source Domain Adaptation (2SW-MDA) and the Geodesic Flow Kernel (GFK) methods in (Patricia et al., 2014) tackle inter-subject DA for sEMG-based gesture recognizers. In 2SW-MDA all the data of each source subject are weighted and combined with the target subject samples with a linear supervised method; for GFK the source and target data are embedded in a low-dimensional manifold (with PCA) and the geodesic flow is used to reduce the domain shift when evaluating the cross domain sample similarity.

2.5.2 Source Data-absent

The overwhelming majority of existing DA methods makes an assumption of freely available source domain data. An equal access to both source and target data makes it possible to measure the discrepancy between their distributions and to build representations common to both target and source domains. In reality, such a simplifying assumption rarely holds, since source data are routinely a subject of legal and contractual constraints between data owners and data customers (Chidlovskii et al., 2016).

Despite the absence of available source samples it is still possible to rely on: 1) statistical information of the source retrieved in advance, 2) model(s) trained on the source data.

CORrelation ALignment (CORAL) (Sun et al., 2016) minimizes domain shift by aligning the second-order statistics of source and target distributions, without requiring any target labels. In contrast to subspace manifold methods (e.g., (Fernando et al., 2013)), it aligns the original feature distributions of the source and target domains, rather than the bases of lower-dimensional subspaces. CORAL performs a linear whitening transformation on the source data then a linear coloring transformation (based on the second-order statistics of the target data). If the statistical parameters of the source data are retrieved in advance of the DA then it can be considered as a source data-absent method.

Adaptive Batch Normalization (AdaBN) (which is an approximation of the whitening transformation, usually applied in deep neural networks) is utilised for DA in (Du et al., 2017) for sEMG-based gesture classification. Furthermore, it builds upon the deep Convolutional Neural Network (CNN) architecture to extract spatial information from the high-density sEMG sensor input. However, it is not modeling the possible temporal information in the time-series data. Apart from that, it has state-of-the-art unsupervised DA performance which has been validated under inter-session and inter-subject domain shifts on several datasets.

(Fernando et al., 2013) introduces (linear) subspace alignment between the source and target domains with PCA. In the common linear subspace classifiers can be trained with comparable performance. The alignments (i.e., the PCA transformations) are learned on the source and target data, respectively. If the source alignment is learned (as a model of the source) in advance of the DA then it can be considered as a source data-absent method.

(Farshchian et al., 2019) introduces the Adversarial Domain Adaptation Network (ADAN) with an AE (trained on the source data) for Brain-Machine Inter-

faces (BMIs). With the representation power of the AE it is possible to capture the source distribution then continuously align the shifting target distributions back to it. ADAN is trained in an adversarial fashion with an Energy-based GAN architecture (Zhao et al., 2017) where the "energy" is the reconstruction loss of the AE, and the domain shifts are represented as the residual-loss distributions of the AE. ADAN learns via the minimax game to pull the target residual distributions to those of the source.

In the transductive learning setting (Ng and Jordan, 2001) there are several source data-absent DA approaches building on the pre-trained source classifier(s).

The Transductive Domain Adaptation (TDA) in (Chidlovskii et al., 2016) utilizes the representation capabilities of the mDA (Chen et al., 2012) to (linearly) adapt the output of a trained source classifier to the target domain. TDA performs unsupervised DA in closed form without the presence of any extra source information.

The transductive Multi-Adapt and the Multi-Kernel Adaptive Learning (MKAL) in (Patricia et al., 2014) both tackle the inter-subject DA for sEMG-based gesture recognizers by the adaptation of trained source classifiers. In Multi-Adapt, an SVM is learned from each source and used as reference (resulted by a convex optimization) when performing supervised learning on the target dataset. In MKAL each SVM source classifier predicts on the target samples and the scores are used as extra input features for the learning of the gesture classifier on the target dataset. Multi-Adapt and MKAL have had comparable performance at that time even though these models do not capture the available temporal information in the time-series data.

(Dose et al., 2018) builds a BMI and investigates DA for multi-variate EEG time-series data classification. The time-series classification of the multi-variate EEG signals is a very similar challenge to the multi-variate sEMG signals. (Dose et al., 2018) captures both the spatial and temporal correlations in the data with a CNN architecture. However, the DA is about supervised fine-tuning of all the model parameters on the target subject which is suboptimal as highlighted by (Du et al., 2017; Ketykó et al., 2019).

The 2-Stage Recurrent Neural Network (2SRNN) model for sEMG gesture recognition and DA in (Ketykó et al., 2019) can be viewed as the deep neural, autoregressive modeling analogy of the MKAL (Patricia et al., 2014). It utilizes a trained source classifier and performs supervised DA to the target (session or subject) via learning a linear transformation between the domains. The transformation is then ap-

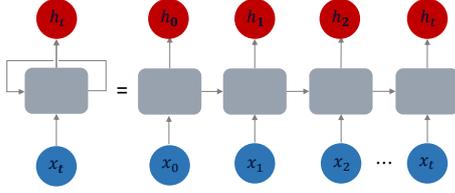


Figure 2: RNN in compact and unrolled representations.

plied to the input (samples coming from the target). Learning is on the divergent (inter-session or inter-subject) domains via the backpropagation (Rumelhart et al., 1986) of the classifier’s cross-entropy loss to its DA layer (which is a linear readout layer of the input). The size of its DA layer is less than 1% of the overall 2SRNN (in terms of the trainable parameters) so it achieves fast computation of the DA, and the 2SRNN has the state-of-the-art performance in inter-session and inter-subject domain shift validations.

3 OUR DIVERGENCE METRIC AND ADAPTATION METHOD

We provide a sequential, source data-absent, transductive, probability-based divergence metric and DA method as well. First, we introduce the RNN architecture for temporal modeling, then the source data-absent and transductive 2SRNN model in details.

3.1 Recurrent Neural Network

Recurrent Neural Network (RNN) (Jordan, 1986) is an autoregressive neural network architecture in which there are feedback loops in the system. Feedback loops allow processing the previous output with the current input, thus making the network stateful, being influenced by the earlier inputs in each step (see Figure 2). A hidden layer that has feedback loops is also called a recurrent layer. The mathematical representation of a simple recurrent layer can be seen in Equation (21):

$$\begin{aligned} \mathbf{h}_t &= \sigma_h(\mathbf{W}_x \mathbf{x}_t + \mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{b}_t) \\ \mathbf{y}_t &= \sigma_y(\mathbf{W}_y \mathbf{h}_t + \mathbf{b}_y) \end{aligned} \quad (21)$$

The hidden state \mathbf{h}_t depends on the input \mathbf{x}_t and the previous hidden state \mathbf{h}_{t-1} . There is a non-linear dependency (via the $\sigma(\cdot)$ wrapper) between them.

However, regular RNNs suffer from the vanishing or exploding gradient problems which means that the gradient of the loss function decays/rises exponentially with time, making it difficult to learn long-term temporal dependencies in the input data (Pascanu et al., 2013). Long Short Term Memory (LSTM)

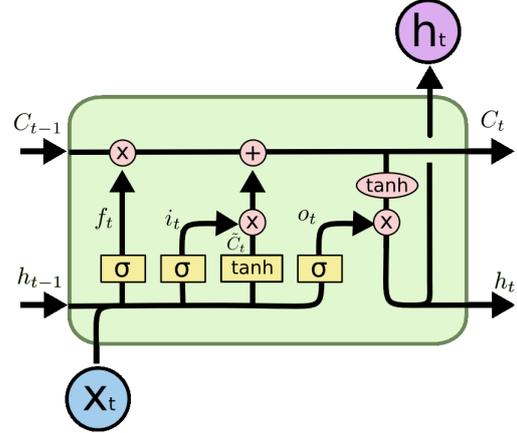


Figure 3: LSTM cell architecture (Olah, 2015).

recurrent cells have been proposed to solve these (Hochreiter and Schmidhuber, 1997).

$$\begin{aligned} \mathbf{f}_t &= \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \\ \mathbf{i}_t &= \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \\ \tilde{\mathbf{C}}_t &= \tanh(\mathbf{W}_C \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C) \\ \mathbf{C}_t &= \mathbf{f}_t * \mathbf{C}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{C}}_t \\ \mathbf{o}_t &= \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \\ \mathbf{h}_t &= \mathbf{o}_t * \tanh(\mathbf{C}_t) \end{aligned} \quad (22)$$

LSTM units contain a set of (learnable) gates that are used to control the stages when information enters the cell (input gate: \mathbf{i}_t), when it is output (output gate: \mathbf{o}_t) and when it is forgotten (forget gate: \mathbf{f}_t) as seen in Equation (22). This architecture allows the neural network to learn longer-term dependencies because it learn also how to incorporate an additional information channel \mathbf{C}_t . In Figure 3 yellow rectangles represent a neural network layer, circles are point-wise operations and arrows denote the flow of data. Lines merging denote concatenation (notation of $[\]$ in Equation (22)), while a line forking denote its content being copied and the copies going to different locations.

For autoregressive modeling of time-series data, RNN with LSTM cells is widely adopted (Hu et al., 2018; Ketykó et al., 2019).

3.2 2-Stage Recurrent Neural Network-based Domain Divergence Metric

Similarly to ADAN (Farshchian et al., 2019), we build a source data-absent, probability-based divergence metric on the validation loss of the source model to measure domain shifts. In ADAN, the distribution of the residual loss (of the AE) is incorporated to express

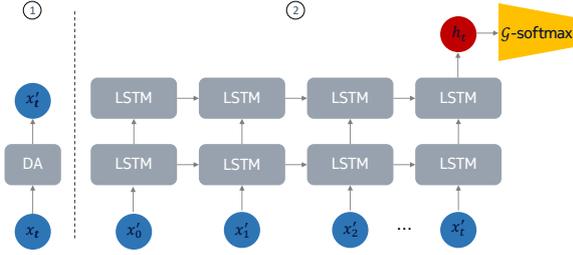


Figure 4: The 2SRNN architecture: 1) DA component, 2) Sequence classifier.

the divergence of target distributions from the one of source. However, we follow the transductive learning setting and directly take the (cross-entropy) loss of the source classifier (exhibiting Equation (5)). Our source classifier is a sequential model (i.e., built on the autoregressive RNN architecture to have temporal modeling capabilities). For this task, we utilise the sequence classifier of the 2SRNN architecture (Ketykó et al., 2019).

The sequence classifier of 2SRNN (visualised as block 2 in Figure 4) is a deep stacked RNN with the many-to-one setup followed by a G -way fully-connected layer (G is the number of gestures to be recognized) and a softmax transformation at the output. The sequence classifier is directly modeling the conditional distribution of $P(Y | X)$, where Y belongs to a categorical distribution with G (gesture) classes. Learning is via the categorical cross-entropy loss (of the ground truth and the predicted Y):

$$\mathcal{L}_{\text{cross-entropy}} = - \sum_{g \in G} I_g \log P_{\Theta, g}, \quad (23)$$

where I_g is the indicator function whether class label g is the correct classification for the given observation and $P_{\Theta, g}$ is the predicted probability that the observation is of class g .

For the divergence measure of distributions (between the source and target domains), we take the categorical cross-entropy losses of the sequence classifier in the following way: the classifier is trained on the source distribution then evaluated on a target one. Hence, the resulting $\mathcal{L}_{\text{cross-entropy}}$ expresses the domain shift in the loss space of the two domains. The cross-entropy between P_S and P_T :

$$H(P_S, P_T) = H(P_S) + D_{KL}(S||T). \quad (24)$$

The $\mathcal{L}_{\text{cross-entropy}}$ expresses the empirical $H(P_S, P_T)$ by the model. A valid source classifier is expected to model the source with the entropy of $H(P_S) \approx 1/G$, so in fact the cross-entropy $H(P_S, P_T)$ captures the actual Kullback-Leibler divergence among P_S and P_T (Equation (13)).

Furthermore, let μ_S and μ_T be the corresponding means of P_S and P_T . We measure the dissimilarity between these two distributions by a lower bound to the Wasserstein distance (Equation (15)), provided by the absolute value of the difference between the means (Berthelot et al., 2017):

$$D_W(P_S, P_T) \geq |\mu_S - \mu_T|. \quad (25)$$

The difference of the empirical means with the $\mathcal{L}_{\text{cross-entropy}}$ approximates Equation (25).

3.3 2-Stage Recurrent Neural Network-based Domain Adaptation

We build a source data-absent, probability distribution of $\mathcal{L}_{\text{cross-entropy}}$ -driven DA. (Ketykó et al., 2019) implements a linear version (L-2SRNN), we extend it to a deep, non-linear one, and name it the Deep 2SRNN (D-2SRNN). Generally, the DA is applied to the input of the sequence classifier at each timestamp t (visualised as block 1 in Figure 4). L-2SRNN learns the weights of a linear transformation:

$$\mathbf{x}'_t = \mathbf{M}\mathbf{x}_t + \mathbf{b}. \quad (26)$$

D-2SRNN learns the weights of chained non-linear transformations:

$$\mathbf{x}'_t = \sigma(\mathbf{M}_2 \cdot \sigma(\mathbf{M}_1 \mathbf{x}_t + \mathbf{b}_1) + \mathbf{b}_2). \quad (27)$$

Figure 5 presents the two consecutive stages of the DA process:

- I) The DA component initially is the identity transformation, and the weights of it are frozen. The sequence classifier is trained from scratch on the labelled source dataset.
- II) The weights of the sequence classifier are frozen and the DA component's weights are trained on a minor subset of the labelled target dataset: $\mathcal{L}_{\text{cross-entropy}}$ is backpropagated (Rumelhart et al., 1986) to the DA component during the process. Hence, the $D_{KL}(P_S||P_T)$ in Equation (24) or expressed via the $D_W(P_S, P_T)$ in Equation (25) gets minimized.

4 RESULTS

We perform experiments to validate our divergence metric and DA for sEMG-based gesture recognition in case of inter-session and inter-subject scenarios. We follow the exact same hyperparameterization and network implementations as in (Ketykó et al., 2019). The parameters $\mathbf{M}, \mathbf{M}_1, \mathbf{M}_2 \in \mathcal{R}^{f \times f}$ in Equations (26) and

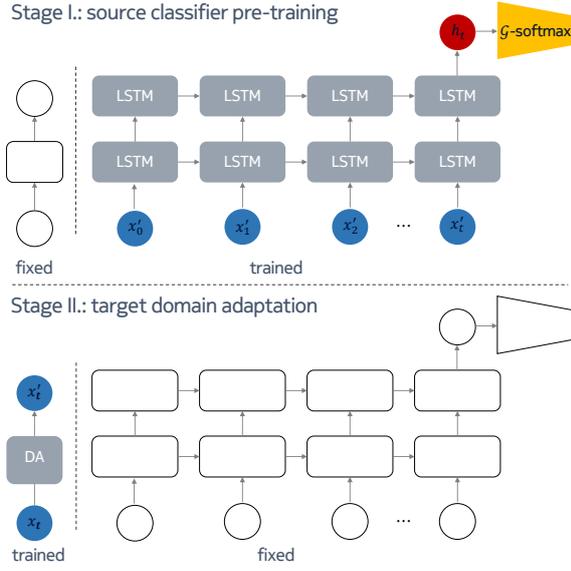


Figure 5: The 2SRNN method: I) Source classifier pre-training, II) Domain adaptation.

(27), where f is equal to the size of the input features (number of sEMG channels). The $\sigma(\cdot)$ non-linearity in Equation (27) is the REctified Linear Unit (Nair and Hinton, 2010).

For the sequence classifier we use a 2-stack RNN with LSTM cells. Each LSTM cell has a dropout with the probability of 0.5 and 512 hidden units. The RNN is followed by a G -way fully-connected layer with 512 units (dropout with a probability of 0.5) and a softmax classifier. Adam (Kingma and Ba, 2014) with the learning rate of 0.001 is used for the stochastic gradient descent optimization. The size of the DA component in both the linear and deep cases is less than 1% of the total trainable network parameters. The gesture recognition accuracy is calculated as given below:

$$\text{Classification Accuracy} = \frac{\text{Correct}}{\text{Total}} * 100\% \quad (28)$$

We investigate the inter-session and inter-subject divergences and validate our DA method on public sparse and dense sEMG datasets. We follow the experiment setups of previous works for comparability. Since we do sequential modeling in all experiments, we decompose the sEMG signals into small sequences using the sliding window strategy with overlapped windowing scheme. The sequence length must be shorter than 300 ms to satisfy real-time usage constraints. To compare our current experiments with previous works, we follow the segmentation strategy in previous studies.

The dense-electrode sEMG CapgMyo dataset has been thoroughly analysed by (Du et al., 2017; Hu

et al., 2018; Ketykó et al., 2019) such as the sparse-electrode sEMG NinaPro dataset by (Patricia et al., 2014; Du et al., 2017; Ketykó et al., 2019).

The CapgMyo dataset (Du et al., 2017): includes HD-sEMG data for 128 electrode channels. The sampling rate is 1 KHz:

1. DB-b: 8 isometric, isotonic hand gestures from 10 subjects in two recording sessions on different days.
2. DB-c: 12 basic movements of the fingers were obtained from 10 subjects.

We downloaded the pre-processed version from <http://zju-capg.org/myo/data> to work with the exact same data as (Du et al., 2017; Ketykó et al., 2019) for fair comparison. In that version, the power-line interference was removed from the sEMG signals by using a band-stop filter (45–55 Hz, second-order Butterworth). Only the static part of the movements was kept in it (for each trial, the middle one-second window, 1000 frames of data). They used the middle, one-second data to ensure that no transition movements are included in it. We rescaled the data to have zero mean and unit variance, then we rectified it and applied smoothing (as low-pass filtering).

The NinaPro DB-1 dataset (Patricia et al., 2014) contains sparse 10-channel sEMG recordings:

1. Gesture numbers 1–12: 12 basic movements of the fingers (flexions and extensions). These are equivalent to gestures in CapgMyo DB-c.

The data is recorded at a sampling rate of 100 Hz, using 10 sparsely located electrodes placed on subjects' upper forearms. The sEMG signals were rectified and smoothed by the acquisition device. We downloaded the version from <http://zju-capg.org/myo/data/ninapro-db1.zip> to use the exact same data as (Du et al., 2017; Ketykó et al., 2019) for fair comparison. For each trial, we used the middle 1.5-second window, 180 frames of data to get the static part of the gestures.

4.1 Divergence Metric Validation

We validate the proposed domain divergence metric in Section 3.2 on the CapgMyo DB-b dataset which covers both the inter-session and inter-subject scenarios.

The divergence results are shown in Figure 6 and Figure 7. In both figures the empirical distributions of $\mathcal{L}_{\text{cross-entropy}}$ are illustrated by their histograms and mean values.

Figure 6 presents the inter-session and inter-subject divergences before DA. The values with red (μ_0, μ_1, μ_2) represent the means of $\mathcal{L}_{\text{cross-entropy}}$; μ_0 is

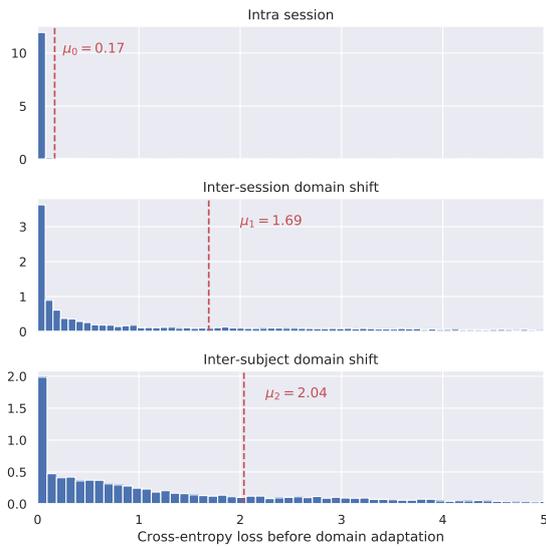


Figure 6: Histograms and mean values in case of the different domain shifts: histograms represent the distributions cross-entropy loss of the classifier on validation data before domain adaptation. The values with red (μ_0, μ_1, μ_2) represent the means of validation loss; μ_0 is the low mean loss in case of intra session; μ_1 and μ_2 (along with their histograms) show high inter-session and inter-subject distribution divergences.

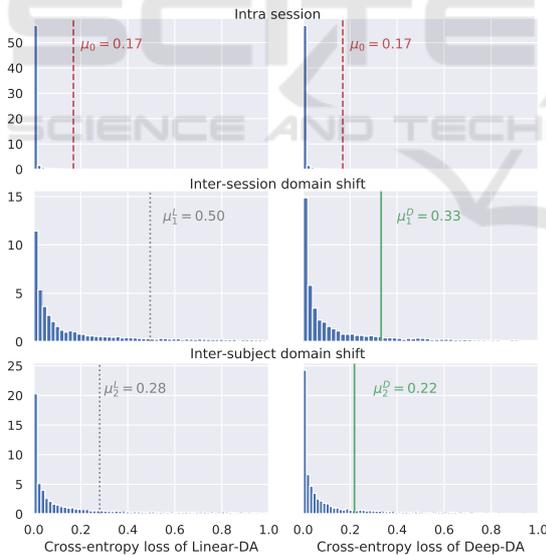


Figure 7: Histograms and mean values in case of different domain shifts and adaptation solutions: histograms represent the distributions of cross-entropy loss of the classifier on validation data after domain adaptation. Intra session statistics (μ_0) represent the source distribution (towards the divergent distributions are aimed to be adapted). The histograms and corresponding mean values with gray (μ_1^L, μ_2^L) represent the validation loss after linear domain adaptation; the histograms and corresponding mean values with green (μ_1^D, μ_2^D) represent the validation loss after deep domain adaptation.

Table 1: Inter-session recognition accuracy results on CapgMyo DB-b.

| | pre-DA | post-DA |
|-------------------------------|--------|--------------|
| AdaBN (Du et al., 2017) | 47.9% | 63.3% |
| L-2SRNN (Ketykó et al., 2019) | 54.6% | 83.8% |
| D-2SRNN | 54.6% | 85.9% |

the low mean loss in case of intra session; μ_1 and μ_2 (along with their histograms) show high inter-session and inter-subject domain shifts. μ_0 shows the power of the sequence classifier: $\mu_0 = 0.17$ is close to the theoretical lower bound of cross-entropy which is $H(S) = 0.125$ in the current case.

Figure 7 presents the inter-session and inter-subject divergences after DA. Intra session statistics (μ_0) represent the source distribution (towards the divergent distributions are aimed to be adapted). The histograms and corresponding mean values with gray (μ_1^L, μ_2^L) represent the validation loss after L-2SRNN DA; the histograms and corresponding mean values with green (μ_1^D, μ_2^D) represent the validation loss after D-2SRNN DA. In all cases, the post-DA $\mathcal{L}_{\text{cross-entropy}}$ distributions appear to be close to one of the source.

4.2 Domain Adaptation Validation

For comparison purposes, we take the exact same pre-trained source classifiers from (Ketykó et al., 2019) and perform D-2SRNN DA (described in Section 3.3). The evaluation of the D-2SRNN DA is exactly the same as of the L-2SRNN and the AdaBN (Du et al., 2017) approaches. Furthermore, the comparison to the MKAL (Patricia et al., 2014) also is exactly the same as in (Du et al., 2017; Ketykó et al., 2019).

Table 1 presents the inter-session recognition accuracy results on the dense CapgMyo DB-b dataset. The L-2SRNN and D-2SRNN share the exact same pre-trained source classifier models. The D-2SRNN DA brings 57.3% improvement which is better by 2.1 percentage points than the L-2SRNN.

Table 2 shows the inter-subject recognition accuracy results on the dense CapgMyo DB-b & DB-c and the sparse NinaPro DB-1 datasets. The L-2SRNN and D-2SRNN share the exact same pre-trained source classifier models. The D-2SRNN DA achieves: 74.9% improvement on the DB-b, 156.3% improvement on the DB-c, 107.4% improvement on the DB-1. The performance ratio (between the deep and the linear solutions) is 3.4% in case of the dense datasets, and 11.7% in case of the sparse one which suggests that there is higher gain by non-linear adaptation in case of a sparse-electrode situation.

Table 2: Inter-subject recognition accuracy results on dense CapgMyo and sparse NinaPro datasets.

| | pre-DA | | | post-DA | | |
|-------------------------------|--------|-------|-------|--------------|--------------|--------------|
| | DB-b | DB-c | DB-1 | DB-b | DB-c | DB-1 |
| AdaBN (Du et al., 2017) | 39.0% | 26.3% | - | 55.3% | 35.1% | - |
| MKAL (Patricia et al., 2014) | - | - | 30% | - | - | 55% |
| L-2SRNN (Ketykó et al., 2019) | 52.6% | 34.8% | 35.1% | 89.9% | 85.4% | 65.2% |
| D-2SRNN | 52.6% | 34.8% | 35.1% | 92.0% | 89.2% | 72.8% |

5 CONCLUSIONS

We showed that the divergences between the empirical distributions of the cross-entropy losses by a source classifier trained on the source distribution and evaluated on the target one is a valid measure for the domain shifts between source and target. It works in the absence of source data and a domain adaptation method built on minimizing that divergence is an effective solution in the transductive learning setting. Furthermore, we pointed out that this metric and the corresponding adaptation method is applicable to investigate and improve sEMG-based gesture recognition performance in inter-session and inter-subject scenarios under severe domain shifts. The proposed deep/non-linear transformation component enhances the performance of the 2SRNN architecture especially in a sparse sEMG setting.

The code is available at <https://github.com/ketyi/Deep-2SRNN>.

ACKNOWLEDGEMENTS

Project no. FIEK_16-1-2016-0007 has been implemented with the support provided of the National Research, Development and Innovation Fund of Hungary, financed under the FIEK_16 funding scheme.

REFERENCES

- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223.
- Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1):151–175.
- Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2006). Analysis of representations for domain adaptation. In *Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS’06*, pages 137–144, Cambridge, MA, USA. MIT Press.
- Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1171–1179. Curran Associates, Inc.
- Berthelot, D., Schumm, T., and Metz, L. (2017). BEGAN: Boundary Equilibrium Generative Adversarial Networks. *ArXiv*, abs/1703.10717.
- Chen, M., Xu, Z., Weinberger, K. Q., and Sha, F. (2012). Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning, ICML’12*, pages 1627–1634, USA. Omnipress.
- Chidlovskii, B., Clinchant, S., and Csurka, G. (2016). Domain adaptation in the absence of source domain data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*, pages 451–460, New York, NY, USA. ACM.
- Clinchant, S., Csurka, G., and Chidlovskii, B. (2016). A domain adaptation regularization for denoising autoencoders. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 26–31, Berlin, Germany. Association for Computational Linguistics.
- Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.
- Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA.
- Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001.
- Csurka, G. (2017). Domain adaptation for visual

- applications: A comprehensive survey. *CoRR*, abs/1702.05374.
- Dose, H., Möller, J. S., Iversen, H. K., and Puthusserypadu, S. (2018). An end-to-end deep learning approach to MI-EEG signal classification for BCIs. *Expert Systems with Applications*, 114:532 – 542.
- Du, Y., Jin, W., Wei, W., Hu, Y., and Geng, W. (2017). Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation. *Sensors (Basel, Switzerland)*, 17(3):458. 28245586[pmid].
- Farshchian, A., Gallego, J. A., Cohen, J. P., Bengio, Y., Miller, L. E., and Solla, S. A. (2019). Adversarial domain adaptation for stable brain-machine interfaces. In *International Conference on Learning Representations*.
- Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. (2013). Unsupervised visual domain adaptation using subspace alignment. In *ICCV*.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.
- Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In Getoor, L. and Scheffer, T., editors, *ICML*, pages 513–520. Omnipress.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*.
- Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.
- Hu, Y., Wong, Y., Wei, W., Du, Y., Kankanhalli, M., and Geng, W. (2018). A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition. *PLOS ONE*, 13(10):1–18.
- Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.
- Jambukia, S. H., Dabhi, V. K., and Prajapati, H. B. (2015). Classification of ECG signals using machine learning techniques: A survey. In *2015 International Conference on Advances in Computer Engineering and Applications*, pages 714–721.
- Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 531–546. Hillsdale, NJ: Erlbaum.
- Ketykó, I., Kovács, F., and Varga, K. Z. (2019). Domain adaptation for sEMG-based gesture recognition with Recurrent Neural Networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.
- Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30, VLDB '04*, pages 180–191. VLDB Endowment.
- Kingma, D. P. and Ba, J. (2014). (adam): A method for stochastic optimization. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
- Kouw, W. M. (2018). An introduction to domain adaptation and transfer learning. *CoRR*, abs/1812.11806.
- Kouw, W. M. and Loog, M. (2019). A review of single-source unsupervised domain adaptation. *CoRR*, abs/1901.05335.
- Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.
- Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recogn.*, 45(1):521–530.
- Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of ICML'10*, pages 807–814, USA.
- Ng, A. Y. and Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01*, pages 841–848, Cambridge, MA, USA. MIT Press.
- Olah, C. (2015). Understanding LSTM Networks.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.
- Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of ICML'13*.
- Patricia, N., Tommasi, T., and Caputo, B. (2014). Multi-source Adaptive Learning for Fast Control of Prosthetics Hand. In *Proceedings of the 22nd International Conference on Pattern Recognition*, pages 2769–2774. IEEE.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning Internal Representations by Error Propagation, pages 318–362. MIT Press, Cambridge, MA, USA.
- Sun, B., Feng, J., and Saenko, K. (2016). Return of frustratingly easy domain adaptation. In *AAAI*.
- van der Maaten, L., Chen, M., Tyree, S., and Weinberger, K. Q. (2014). Marginalizing corrupted features. *CoRR*, abs/1402.7001.
- Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Mach. Learn.*, 23(1):69–101.
- Zhao, J., Mathieu, M., and LeCun, Y. (2017). Energy-based generative adversarial networks. *ICLR 2017*.
- Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.