Variational Adversarial Deep Domain Adaptation for Health Care Time Series Analysis

Sanjay Purushotham*, Wilka Carvalho*, Tanachat Nilanon, Yan Liu
Department of Computer Science
University of Southern California
Los Angeles, CA 90089, USA
{spurusho,wcarvalh,nilanon,yanliu.cs}@usc.edu

Abstract

Data-driven machine learning, in particular deep learning, is improving state-of-the-art in many healthcare prediction tasks. A current standard protocol is to collect patient data to build, evaluate, and deploy machine learning algorithms for specific age groups (say source domain), which, if not properly trained, can perform poorly on data from other age groups (target domains). In this paper, we address the question of whether it is possible to adapt machine learning models built for one age group to also perform well on other age groups. Additionally, healthcare time series data is also challenging in that it is usually longitudinal and episodic with the potential of having complex temporal relationships. We address these problems with our proposed adversarially trained Variational Adversarial Deep Domain Adaptation (VADDA) model built atop a variational recurrent neural network, which has been shown to be capable of capturing complex temporal latent relationships. We assume and empirically justify that patient data from different age groups can be treated as being similar but different enough to be classified as coming from different domains, requiring the use of domain-adaptive approaches. Through experiments on the MIMIC-III dataset we demonstrate that our model outperforms current state-of-the-art domain adaptation approaches, being (as far as we know) the first to accomplish this for healthcare time-series data.

1 Introduction

Healthcare data has become a frontier for machine learning. While, data-driven machine learning models, in particular deep learning models, are improving the state-of-the-art in many healthcare prediction tasks [4]; they still face a big challenge for real-world deployment due to the nature of healthcare data. A current standard protocol in healthcare is to collect patient data for population of a specific age group (say source domain), and build & evaluate machine learning models for this age group. These models tend to perform poorly on patient data from other age groups since the patient data of different age groups (target domains) usually come from different data distributions [2]. Moreover, many machine learning models do not capture the complex temporal relations which are inherent in the longitudinal and episodic healthcare data. For example, children with Acute Hypoxemic Respiratory Failure (AHRF) (specifically Acute Respiratory Distress Syndrome (ARDS)) have a lower pulse oximetry saturation ratio (SF = SpO2/FiO2) compared to the adults, which affects risk stratification [14]. Also, the changes in ventilator settings temporally affects the PF ratio (which is a surrogate for SF ratio) which further impacts the health state of the patients with AHRF. Thus,

*: Co-first authors

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

we believe a fundamental challenge is to develop new machine learning approaches which can model the complex temporal relations and also handle the varying data distributions of healthcare data.

An underlying assumption for many machine learning algorithms is that the training data and testing data are in the same feature space and have the same distribution [18]. While this makes the model easier to train, it also limits its utility. When the feature space or data distribution change, many models must be rebuilt or re-trained. Transfer learning is a learning framework that attempts to circumvent this obstacle and transfer knowledge from a source domain $\mathbf{x_{src}}$ to a target domain $\mathbf{x_{tgt}}$. When the feature spaces of the source and target domains are the same but the marginal probability distributions are different, i.e. $P(\mathbf{x_{src}}) \neq P(\mathbf{x_{tgt}})$, this is known as domain adaptation. Domain adaptation has been well studied, mainly in computer vision [15, 9] and natural language processing fields [3, 6, 11]. Most of the existing methods learn a shallow representation model to achieve domain adaptation by reducing the domain discrepancy between the source and target domains. Recently, the deep learning paradigm has been explored for domain adaptation [8, 7, 20, 17] to learn invariant nonlinear representations which suppress domain-specific factors. While these deep learning based domain adaptation approaches have achieved excellent performance for non-sequential data (such as images); they are not suited for sequential data (such as healthcare patient data) since they do not adapt/transfer complex temporal latent relationships from one domain to another. In this paper, we propose a novel domain adaptation model, termed as **Variational Adversarial Deep Domain Adaptation (VADDA)**, for performing unsupervised domain adaptation of complex multimodal sequential data such as healthcare time series data. The VADDA uses variational methods to capture the underlying temporal relationships and adversarial training to achieve domain adaptation. As far as we know, this is the first model capable of accomplishing unsupervised domain adaptation of complex healthcare time series data.

2 Related Work

A good survey on Transfer learning and Domain adaptation approaches has been done in several previous works [18, 10, 19]. Here, we briefly discuss two deep domain adaptation approaches which are closely related to our proposed model. Domain Adversarial Neural Networks (DANN) [7] is a deep domain adaptation model which uses two core components to create domain-invariant representations, a feature extractor that produces the data's latent representation, and an adversarial domain labeler that attempts to classify that data's domain. In [16], a Deep Adaptation Network (DAN) was proposed to learn hidden representations of all task-specific layers by embedding them in a reproducing kernel Hilbert space where the mean embeddings of different domain distributions are explicitly matched. While, these deep learning approaches learn domain-invariant representations, they fail to capture and transfer the underlying complex temporal latent relationships from one domain to another domain since they use convolutional or feed forward neural networks which are not suitable for healthcare time series data. In the following section, we discuss our approach, Variational Adversarial Deep Domain Adaptation (VADDA), to model and transfer complex temporal latent relationships while learning the domain invariant representations.

3 Variational Adversarial Deep Domain Adaptation

VADDA utilizes deep neural networks with adversarial training (similar to DANN [7]), to create domain-adaptive invariant representations by assuming that training and testing data share similar feature spaces but come from different distributions. Like in DANN, our VADDA is trained on labeled data from the source domain (Eg. older patients) and unlabeled data from the target domain (Eg. younger patients). Unlike DANN, VADDA uses Variational Recurrent Neural Networks (VRNN) [5] to capture the latent temporal dependencies. During training, VADDA finds feature representations that are (i) discriminative for the main learning task on the source domain and are (ii) indiscriminate with respect to the shift between the domains. We show the block diagram of our VADDA model in Figure 1 and briefly explain the model below.

We denote the healthcare time series of N patients as $\{\mathbf{x^i} = (x_t^i)\}_{i=1}^N$ and treat them as N data points. We denote $\{\mathbf{x^i_{src}}\}_{i=1}^n$ with indices $i = 1, \ldots, n$ as source domain data samples and $\{\mathbf{x^i_{tgt}}\}_{i=n+1}^N$ as target domain data samples. We assume that source domain data comes with labels $y_i \in \{0, 1\}$ (these labels correspond to a clinical outcome, for example: mortality), while target domain has no labelled data points. We assign a domain label $d_i \in \{0, 1\}$ to each data sample to indicate if it

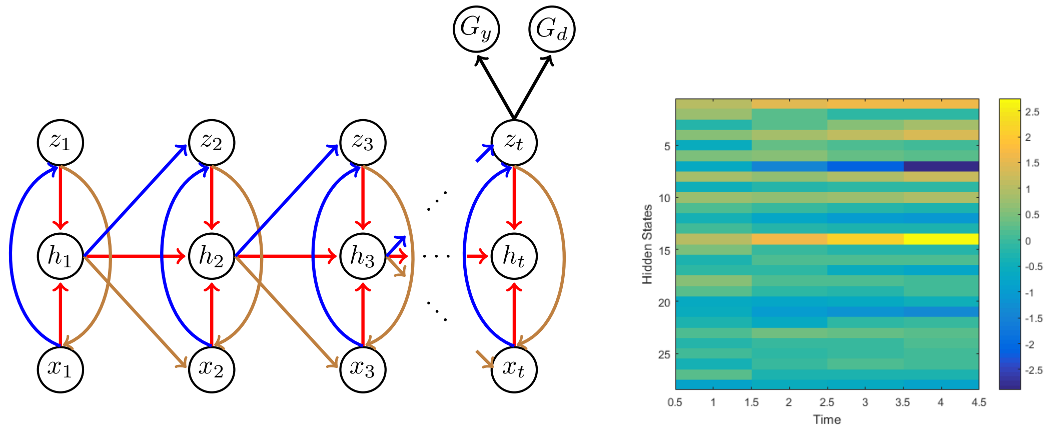

Figure 1: The left figure shows the VADDA model. Blue lines show the inference process where z_t(latent variable at time t) is inferred using both x_t and h_{t-1} (hidden state at $t-1$). Brown lines show the generation process where x_t is generated from both z_t and h_{t-1}. Red lines show the recurrence where h_t is informed by h_{t-1}, which is informed by z_{t-1} and x_{t-1}. Black lines indicate classification. G_d, G_y correspond to the domain and label classifier networks respectively. The right figure shows the plot of temporal dependencies captured by the VADDA model for a target domain data sample.

comes from source domain or target domain. We use the encoder of a VRNN to learn the latent representations $\mathbf{z^{i*}}$ for each $\mathbf{x^i}$. Afterwards, we use $\mathbf{z^{i*}}$ to learn to classify source data samples $\mathbf{x^i_{src}}$ and then adversarially to transfer this knowledge to target data samples $\mathbf{x^i_{tgt}}$. The process is as follows.

For each data sample $\mathbf{x^i}$ we learn a feature representation z^{i*} by minimizing the following reconstruction error of the VRNN at each time step:

$$\mathcal{L}_r(x^i_t;\theta_f) = E_{q(z^i_{\leq T^i}|x^i_{\leq T^i})}[\sum_{t=1}^{T^i}(-D(q(z^i_t|x^i_{\leq t}, z^i_{<t};\theta_f)||p(z^i_t|x^i_{<t}, z^i_{<t})) + \log p(x^i_t|z^i_{\leq t}, x^i_{<t}))])$$

where $q(z^i_t|x^i_{\leq t}, z^i_{<t})$ is the inference model, $p(z^i_t|x^i_{<t}, z^i_{<t})$ is the posterior, and $p(x^i_t|z^i_{\leq t}, x^i_{<t})$ is the generative model and θ_f is the parameters of the VRNN. For brevity, here we do not discuss the generation, inference and learning of the VRNN model [5].

For adversarial training of the VADDA, we use the gradient reversal trick proposed in the DANN model [7]. Let $G_y(z^{i*};\theta_y)$ with parameters θ_y and $G_d(z^{i*};\theta_d)$ with parameters θ_d denote the deep neural networks used to classify clinically relevant outcome labels y_i and domain labels d_i respectively. Let $\mathcal{L}_y(\mathbf{x^i};\theta_y)$ and $\mathcal{L}_d(\mathbf{x^i};\theta_d)$ denote the loss functions of these label and domain classifiers respectively. G_y neural network propagates the gradients calculated on label loss $\mathcal{L}_y$ to the VRNN encoder of our VADDA, while G_d neural network adversarially propagates the gradients (by reversing the gradients) to achieve the domain adaptation. Our VADDA combines all the above loss functions in an objective function to minimize

$$E(\theta_f,\theta_y,\theta_d) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T^i}\mathcal{L}_r(\mathbf{x^i};\theta_f)+\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_y(\mathbf{x^i_{src}};\theta_y)-\lambda(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_d(\mathbf{x^i_{src}};\theta_d)+\frac{1}{n'}\sum_{i=n+1}^{N}\mathcal{L}_d(\mathbf{x^i_{tgt}};\theta_d)))$$

where λ is a "trade-off" between optimizing on making domain-invariant representations and optimizing label classifier accuracy. We learn the parameters of the VADDA by iteratively solving the following equations:

$$\hat{\theta}_f = \arg\min_{\theta_f} E(\theta_f,\theta_y,\theta_d)$$

$$\hat{\theta}_y = \arg\min_{\theta_y} E(\theta_f,\theta_y,\theta_d)$$

$$\hat{\theta}_d = \arg\min_{\theta_d} E(\theta_f,\theta_y,\theta_d)$$

Minimizing $\mathcal{L}_y$ helps the VADDA classify example $\mathbf{x^i}$ based off its latent representation z_i^*; minimizing $\mathcal{L}_r$ helps the VADDA generating z_i^* to capture temporal patterns, and maximizing $\mathcal{L}_d$ helps the VADDA in creating z_i^* which are domain-invariant. Together, the three components allow the VADDA to capture temporal latent dependencies and transfer this knowledge across domains.

4 Experiments

We use the MIMIC-III dataset [12] - which is a public dataset with deidentified clinical care data collected at Beth Israel Deaconess Medical Center from 2001 to 2012. It contains over 58,000 hospital admission records of 38,645 adults and 7,875 neonates. For our work, we extracted 21 time series features (such as Base excess, pH value, Mean Air Pressure, PaO2, etc.) from 5527 admission records for acute hypoxemic respiratory failure (AHRF) based on [13]. We grouped the patients into 4 categories based on their age [1] - Group 2: working-age adult (20 to 45 yrs); Group 3: old working-age adult (46 to 65 yrs); Group 4: elderly (66 to 85 yrs); Group 5: old elderly (85 yrs and up). These 4 categories are treated as 4 domains in our work. All the time series are >96 hours of duration (4 days), and only the first 4 day (after admission) time series data is used for training and testing our models. We perform a mortality prediction task, where we predict whether the patient dies from AHRF during their hospital stay.

Table 1: AUC Comparison for Morality Prediction task with and without Domain Adaptation

Source-Target	SVM	LR	DT	AdaBoost	DANN	DDANN	VADDA
Group 3- Group 2	0.50	**0.5386**	0.5032	0.5154	0.5684	0.639	**<u>0.748</u>**
Group 4- Group 2	0.5	**0.6235**	0.6147	0.556	0.6809	0.630	**<u>0.759</u>**
Group 5- Group 2	0.5	0.5103	0.5049	0.5494	**<u>0.7078</u>**	0.634	0.703
Group 2- Group 3	0.5	0.5332	0.5321	**0.5364**	0.5362	0.617	**<u>0.732</u>**
Group 4- Group 3	**0.6005**	0.5258	0.5439	0.5287	0.6404	0.716	**<u>0.740</u>**
Group 5- Group 3	0.5	0.5018	**0.5249**	0.5188	0.6701	0.664	**<u>0.727</u>**
Group 2- Group 4	0.5	0.5018	**0.5087**	0.5042	0.5333	0.641	**<u>0.702</u>**
Group 3- Group 4	0.5	0.5122	**0.5410**	0.5035	0.6019	0.689	**<u>0.750</u>**
Group 5- Group 4	0.5	0.5	**0.5033**	0.5032	0.6672	0.627	**<u>0.735</u>**
Group 2- Group 5	0.5	**0.5868**	0.5580	0.5480	0.5216	0.674	**<u>0.725</u>**
Group 3- Group 5	0.5	0.5040	0.5003	**0.5141**	0.5776	0.697	**<u>0.734</u>**
Group 4- Group 5	0.5	0.5106	0.5390	**0.5507**	0.6278	0.684	**<u>0.744</u>**

In the above table, we test classification without adaptation using Support Vector Machines (SVM) with RBF kernel, Logistic Regression (LR), Decision Trees (DT), and Adaboost with decision trees classifiers. We test classification with adaptation with shallow and deep Domain Adversarial Neural Networks (DANN and DDANN), and with our Variational Adversarial Domain Adaptation Model (VADDA).

Preliminary Results Table 1 shows the AUC comparison results of Mortality prediction task with and without domain adaptation for all source-target domains pairs. It is hard to determine which non-adaptive models performed best, as best results seem spread by age group. However, across all age groups, adaptive models outperformed non-adaptive models. The average increase in performance by using a shallow DANN was 9.54%; a deep DANN 16.91%; and our VADDA 26.23%, with it outperforming the deep DANN by 10.05%. As expected, domain adaptation by the VADDA from distant age groups (e.g between Group 5, the oldest group, to Group 2, the youngest group) showed the worst increase in performance with 0.99%. Despite, most results of distant age adaptation are promising: 3-5 adaptation (Group 3 to Group 5) shows an increase of 31.55% and 2-4 shows an increase of 30.16%. These results empirically demonstrate that domain adaptation across age-groups is beneficial for healthcare predictive tasks and our VADDA model achieves the state-of-the-art performance. Moreover, the complex latent temporal dependencies captured by VADDA model can be visualized (as shown in the Figure 1) and are useful for further interpretation and analysis.

5 Summary

With a multitude of difficulties to overcome like complex temporal relations or stark contrasts in the amount of labeled data across different distributions of data, healthcare data provides ample opportunity to showcase innovation in machine learning. In this paper, we proposed VADDA model to conquer the domain adaptation of complex healthcare time series data. We demonstrated results where, in all instances, domain adaptive models out-perform non-domain adaptive models, with our VADDA model showing significant success. In our future work, we will provide a theoretical understanding of the VADDA model and evaluate it on larger healthcare datasets.

References

[1] `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/`.

[2] B. Alemayehu and K. E. Warner. The lifetime distribution of health care costs. *Health services research*, 39(3):627–642, 2004.

[3] J. Blitzer, M. Dredze, F. Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007.

[4] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516. ACM, 2015.

[5] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio. A Recurrent Latent Variable Model for Sequential Data. *arXiv.org*, May 2016.

[6] H. Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1), 2016.

[8] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011.

[9] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012.

[10] J. Jiang. A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*, 2008.

[11] J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, volume 7, pages 264–271, 2007.

[12] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 2016.

[13] R. G. Khemani, D. Conti, T. A. Alonzo, R. D. Bart III, and C. J. Newth. Effect of tidal volume in children with acute hypoxemic respiratory failure. *Intensive care medicine*, 35(8):1428–1437, 2009.

[14] R. G. Khemani, N. J. Thomas, V. Venkatachalam, J. P. Scimeme, T. Berutti, J. B. Schneider, P. A. Ross, D. F. Willson, M. W. Hall, and C. J. Newth. Comparison of spo2 to pao2 based markers of lung disease severity for children with acute lung injury. *Critical care medicine*, 40(4):1309–1316, 2012.

[15] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1785–1792. IEEE, 2011.

[16] M. Long and J. Wang. Learning transferable features with deep adaptation networks. *CoRR, abs/1502.02791*, 1:2, 2015.

[17] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair auto encoder. *arXiv preprint arXiv:1511.00830*, 2015.

[18] S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

[19] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[20] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.