Deep Spatio-Temporal Feature Learning using Autoencoders

Yao Jia and **Mehul Motani**
Department of Electrical and Computer Engineering
National University of Singapore
eleyaj@nus.edu.sg, motani@nus.edu.sg

Abstract

In modern medicine, patient vital sign information is often collected as high-dimensional multi-channel time series data, which contains both spatial and temporal information. In this paper, we propose a hybrid feature learning model containing both spatial and temporal autoencoders to learn deep feature representations of time series data. We use a publicly available electroencephalograph (EEG) dataset to evaluate our model's classification performance and compare the results to: (i) using raw data as features, and (ii) features learned from various combinations of spatial and temporal autoencoders. Our findings highlight that the way in which we exploit spatial and temporal correlations makes a significant difference and demonstrate the effectiveness of our model in processing multivariate time series patient data.

1 Introduction

In modern medicine, patient data is often collected as high-dimensional multi-channel time series data, which contain both spatial and temporal information. An important example of such data are electroencephalograph (EEG) signals, which measure the electrical activity of the brain from electrodes placed along the scalp (see Figure 1). The records are usually time series data with large dimensionality and high sampling frequency. This makes the use of EEG signals for disease classification and predictive diagnosis challenging as large-scale data processing suffers from the curse of dimensionality [1]. Therefore, extracting a compact feature representation from the raw data facilitates practical and (hopefully) better clinical decision making.

Interestingly, EEG signals typically have spatial and temporal correlations due to the placement of the electrodes on the scalp (see Figure 1). However, extracting spatio-temporal correlations and useful features from EEG signals requires expert knowledge in conventional feature extraction approaches. These supervised approaches are domain-driven and sometimes do not generalize well. To overcome the pitfalls of supervised methods, the autoencoder as a data-driven unsupervised approach has been proposed to directly learn feature representations from large-scale data [2]. However, recent work in [3], [4] and [5] does not specifically cater for learning compact representations and dependencies from both the spatial and temporal dimensions of the data.

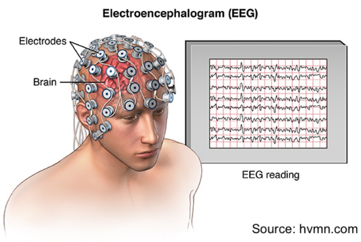

Figure 1: Example of EEG signal measurement and sample reading.

This research was supported in part by the Singapore Ministry of Education under grant R-263-000-D35-114.

Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

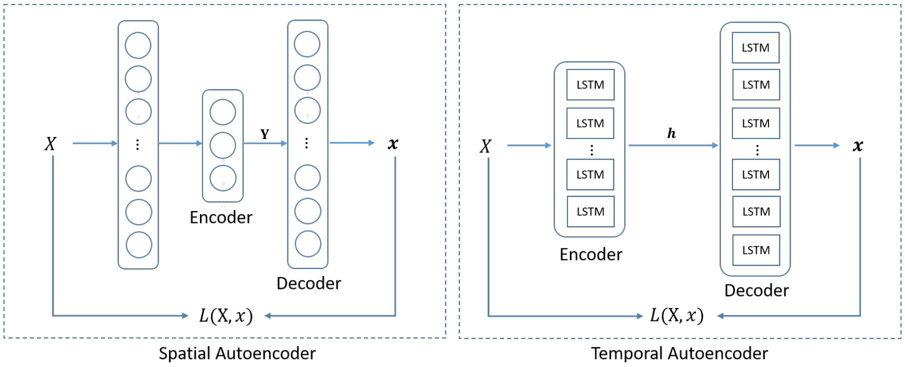

Figure 2: Exploiting spatial and temporal information with Autoencoders.

Inspired by [6] which uses the idea of spatio-temporal autoencoders to predict the next video frame, in this paper, we propose a hybrid feature learning model containing spatial and temporal autoencoders to learn deep feature representations from both the spatial and temporal domains. We evaluate its performance and compare the results to: (i) using raw data as features, and (ii) features learned from various combinations of spatial and temporal autoencoders. Our findings highlight that the way in which we exploit spatial and temporal correlations makes a significant difference and demonstrate the effectiveness of our model in processing multivariate time series patient data.

2 Exploiting Spatial and Temporal Information

It is well known that compact representations can be derived from raw data in an unsupervised manner using autoencoders [2]. In this paper, we explore an autoencoder type structure to capture both spatial and temporal dependencies in raw time series data in an unsupervised manner. We use a standard neural network based autoencoder [2] to extract the spatial information, and we choose an autoencoder constructed with long short term memory (LSTM) units [7] to learn the temporal information. These two autoencoders, called the spatial autoencoder (SAE) and temporal autoencoder (TAE) respectively, are shown in Figure 2. Not surprisingly, the SAE and TAE can be stacked in various ways to yield hybrid combinations. What is surprising is that the way in which we stack them leads to widely varying performance and suggests that, at least for the datasets we studied, stacking an SAE followed by a TAE is better for classification (as compared to stacking a TAE followed by an SAE).

2.1 Capturing Spatial Features

The SAE is a standard autoencoder constructed with multi-layer neural networks, consisting of an input layer, an encoder and a decoder. We use $X \in \mathbb{R}^{S \times T}$ to represent the raw time series patient data, where S is the number of channels (e.g., the number of electrodes placed on the scalp in EEG data, the number of vital signs in ICU data), and T is the length of the time series; and we use x to represent the reconstructed patient data from an autoencoder, which has the same dimension as X.

The encoder in SAE contains only one layer with S' number of neurons that compress the number of channels in the input data X from S to S' while learning a more compact representation $Y \in \mathbb{R}^{S' \times T}$ in the spatial domain. This encoding process does not necessarily break the temporal dependencies. The decoder in SAE reconstructs the input data x based on the encoded features Y. SAE can be optimized by stochastic gradient descent and trained by minimizing the reconstruction error between X and x through the mean squared error loss function $L(X, x)$.

2.2 Capturing Temporal Features

Similar to the SAE, the TAE also has an input layer, an encoder and a decoder. In this paper, we use LSTMs with standard structure to construct the encoder and the decoder of the TAE. A total of M LSTM units in the encoder are used to learn and extract temporal dependencies from the multivariate time series inputs X, which have dimension of $S \times T$. We take outputs from all the hidden states of every LSTM unit in the TAE encoder, and consider them as the temporal feature representations $\boldsymbol{h} \in \mathbb{R}^{M \times T}$ of the input data. The decoder contains another set of S LSTM units which are attached after the encoder to map the encoded temporal feature representations $\boldsymbol{h}$ back to the input of TAE, X. The TAE can be optimized by gradient descent and trained by minimizing the reconstruction error between X and x through the mean squared error loss function $L(X, x)$.

2.3 Hybrid Stacked Combinations

As motivated by [8], the two types of autoencoders described above can be stacked together to learn deeper feature representations from the spatial and temporal domains. For example, we use SSAE and TTAE to denote the autoencoders which stack two SAEs or two TAEs in a chain respectively. In these stacked autoencoders models, raw data is passed into the first stage to learn one set of feature representations and then passed into the next stage to continue learning another set of feature representations. The autoencoder at each stage is trained and optimized individually.

Unlike SAEs and TAEs, where information in only one domain is captured, it is natural to stack a SAE and a TAE together (first explored in [6] for video frame prediction). We explore various combinations of the SAE and TAE to learn compact spatio-temporal feature representations from multivariate time series medical patient data. We denote the combination of one SAE followed by one TAE as STAE, and one TAE followed by one SAE as TSAE. Taking STAE as an example, multivariate time series data is passed into the SAE first so as the spatial information in the raw data is extracted without touching the time domain. The outputs from the SAE, which contains compact spatial feature representations are then fed into a TAE to further learn representations capturing the temporal characteristics. The resultant feature set learned from STAE is the deep spatio-temporal feature representation of the raw multivariate time series patient data.

We note that [9] also considered a STAE-type structure for multivariate time series data and include a discussion and comparison in Sec. 4.

3 Classification Performance and Results

In this section, we present the classification problem and dataset used in this paper, and evaluate the performance of various feature representations learned from the dataset with four different classifiers.

The EEG dataset extracted from the UCI Machine Learning Repository [10] contains EEG time series data from 1200 different patients. Each patient's EEG time series data contains measurements from 64 channels for 256 time stamps. Therefore, one piece of EEG data has a dimension of 64×256. Each patient is labeled as either alcoholic (Class 0) or non-alcoholic (Class 1). In the EEG dataset used in our study, both Class 0 and Class 1 contain 600 patients equally. The classification problem considered in this study is to classify whether a patient is alcoholic (Class 0) or non-alcoholic (Class 1), based on her EEG measurements. The results presented here are averaged based on 5-fold cross validation repeated 10 times.

In Table 1, we compare the classification performance using the raw EEG data and the different feature sets learned by various autoencoder combinations, for four different classifiers: support vector machine (SVM), gradient boosting (GB), random forest (RF) and decision tree (DT) [1]. With the SVM classifier, it can be seen that the feature set learned by STAE achieves the highest classification accuracy among all different feature learning methods, and also outperforms the raw data. Using the STAE feature set, we can classify a patient to be alcoholic with 82.1% accuracy, which is about 2.4%, 4.9% and 5.3% higher than using the feature sets extracted from SAE, TAE, and TSAE for classification respectively. Similar observations in classification accuracy can also been seen in Table 1 with the GB, RF and DT classifiers. These observations imply that feature representations extracted from the spatial or temporal domain solely do not well represent the multivariate time series data.

Furthermore, Table 1 shows that feature representations learned from STAE achieve the highest scores in the AUC-ROC and F1-score with the SVM classifier. Similar patterns can be seen from the AUC-ROC and F1-score with the GB, RF and DT classifiers.

We also notice that with all four different classifiers, classification with feature sets extracted by SAE and TAE purely do not always achieve better accuracies and F1-scores as compared to the classification with the raw EEG data. This may be because the information carried by one domain

	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
		SVM			GB			RF			DT	
RAW	0.817	0.820	0.819	0.775	0.865	0.775	0.708	0.721	0.712	0.612	0.613	0.613
SAE	0.802	0.889	0.802	0.768	0.868	0.768	0.754	0.755	0.746	0.622	0.621	0.622
TAE	0.783	0.864	0.783	0.808	0.889	0.808	0.741	0.741	0.739	0.633	0.636	0.633
TSAE	0.780	0.860	0.780	0.744	0.817	0.743	0.683	0.683	0.671	0.600	0.590	0.589
STAE	**0.821**	**0.896**	**0.821**	**0.812**	**0.900**	**0.812**	**0.773**	**0.770**	**0.765**	**0.661**	**0.661**	**0.662**

Table 1: Classification performance on UCI EEG dataset shows that STAE consistently beats all other feature learning combinations. ACC=Accuracy, AUC=Area under the ROC Curve, and F1= F1 Score. SVM=Support Vector Machine, GB=Gradient Boosting, RF=Random Forest and DT=Decision Tree.

(i.e. spatial or temporal) is not a good representative for the entire dataset distribution. In other words, the dependencies between the spatial and temporal measurements are not fully captured.

Another interesting comparison is between the STAE and TSAE models. Since both feature learning models use the combination of SAE and TAE (albeit in different order), one would expect them to have similar performance. Surprisingly, the classification results of the two differ significantly. The results presented in Table 1 show that STAE-SVM outperforms TSAE-SVM by 5.3%, 4.2%, and 5.3% in Accuracy, AUC-ROC score, and F1-score, respectively. Similar results are obtained with the GB classifier, where STAE-GB outperforms TSAE-GB by around 9.1%, 10.2%, and 9.3% for the three metrics. The same pattern is found with the RF and DT classifiers: STAE-RF outperforms TSAE-RF by 13.2%, 12.7%, and 14.0, while STAE-DT outperforms by 10.2%, 12.0%, and 12.4% for the three metrics. A possible reason is that when data is passed into an LSTM unit, the multiple features are summarized, probably reducing the original spatial dependencies. When the spatial dependencies are captured first, the temporal dependencies are not necessarily being broken in the process. We also note that TSAE performs worse than the single autoencoder model (e.g., SAE, TAE) and worse than the raw EEG data, with all four classifiers.

Our finding that STAE consistently beats the SAE/TAE/TSAE combinations is robust in the sense that we have tested in with strong classifiers (such as SVM) and weak classifiers (such as DT), and on a variety of other time-series datasets, including the PTB Diagnostic ECG database on Physionet [11]. Our results suggest that the STAE stacking model is effective in spatio-temporal feature learning.

4 Reflections

In this paper, we propose to learn deep feature representations from multivariate time series patient data via spatial and temporal autoencoders. When stacking a single type of autoencoder in a sequence, feature representations from one domain can be learned. Feature representations that contain information in multiple domains can be obtained from a hybrid model which contains a stacked combination of different types of autoencoders. Our key finding is that the order of the autoencoders in the stacking sequence seems to make a significant difference in the performance over a wide range of classifiers. We contend that a more careful study and explanation of this phenomenon are excellent open research questions relevant to the community.

We note that [9] carries out a related study, in which the combination of a classic neural network based autoencoder and an LSTM autoencoder is used to learn features in patient data with missing observations. The main difference between this previous study and our work in this paper has to do with the feature representation output from the temporal autoencoder. The autoencoder in [9] uses the information at the last time stamp as the feature representation, which is similar to what is done for video prediction [6]. However, in our paper, information at each time stamp are used as the feature representations for the entire time series. We find that, for multivariate time series data, keeping only the information at the last time stamp is inferior to keeping information at every time stamp.

For comparison, we conduct a similar experiment to one done in [9] where 30% of the raw data is dropped to emulate missing observations. We apply linear imputation before using our hybrid combination of SAE and TAE. With our model, STAE-SVM with 30% missing observations can achieve 73.5% classification accuracy and an AUC-ROC score of 0.807. These numbers are slightly higher than the 72.3% accuracy and 0.790 AUC-ROC reported in [9]. The improvement in the results demonstrates the effectiveness of our model in learning spatio-temporal features from multivariate time series data.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer-Verlag, 2009.

[2] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[3] R. Miotto *et al.*, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, no. 26094, May 2016.

[4] C. Zhou *et al.*, "Learning deep representations from heterogeneous patient data for predictive diagnosis," in *ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics (ACM-BCB)*, Aug. 2017, pp. 115–123.

[5] S. Nitish, M. Elman, and R. Salakhutdinov, "Unsupervised leanring of video representations using LSTMs," in *International Conference on Machine Learning*, Jul. 2015, pp. 843–852.

[6] V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," *CoRR*, vol. abs/1511.06309, 2015. [Online]. Available: http://arxiv.org/abs/1511.06309

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] P. Vincent *et al.*, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[9] J. Yao, C. Zhou, and M. Motani, "Spatio-temporal autoencoder for feature learning in patient data with missing observations," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2017, pp. 886–890.

[10] UCI EEG Database, https://archive.ics.uci.edu/ml/datasets/eeg+database, Accessed 21 June 2018.

[11] PTB Diagnostic ECG Database, https://physionet.org/physiobank/database/ptbdb/, Accessed 21 June 2018.