Medical Time Series Classification with Hierarchical Attention-based Temporal Convolutional Networks: A Case Study of Myotonic Dystrophy Diagnosis

Lei Lin, Beilei Xu, Wencheng Wu, Trevor W. Richardson, Edgar A. Bernal
Rochester Data Science Consortium
1219 Wegmans Hall, University of Rochester
250 Hutchison Road, Rochester, NY
{Lei.Lin, Beilei.Xu, Wencheng.Wu, Trevor.Richardson, Edgar.Bernal}@rochester.edu

Abstract

Myotonia, which refers to delayed muscle relaxation after contraction, is the main symptom of myotonic dystrophy patients. We propose a hierarchical attention-based temporal convolutional network (HA-TCN) architecture for myotonic dystrohpy diagnosis from handgrip force time series data, and introduce mechanisms that enable model explainability. We compare the performance of the HA-TCN model against that of benchmark TCN models, LSTM models with and without attention mechanisms, and SVM approaches with handcrafted features. In terms of classification accuracy and F1 score, we found deep learning models have similar levels of performance, and they all outperform SVM. Further, the HA-TCN model outperforms its TCN counterpart with regards to computational efficiency regardless of network depth, and in terms of performance particularly when the number of hidden layers is small. Lastly, HA-TCN models can consistently identify relevant time series segments in the relaxation phase of the handgrip force time series, and exhibit increased robustness to noise when compared to attention-based LSTM models.

1. Introduction

Artificial intelligence (AI) techniques, along with the ever-increasing availability of healthcare data, have opened numerous new avenues of research. Deep learning models are currently preferred as they avoid handcrafted features required by traditional machine learning approaches [3, 6] and have shown promising results and state-of-the-art performance. In spite of their superior performance, deep learning models have been criticized for their lack of interpretability, particularly in the healthcare community, for which model explainability is as important as accuracy. The certainty needs to exist that the diagnosis of a disease is made based on real causes instead of systemic biases in the data. Doctors and patients want to understand the reasoning process that leads to suggested treatment avenues.

In this study, we focus on interpretable classification of medical time series data, more specifically, diagnosis of myotonic dystrophy from handgrip force time series data. Myotonic dystrophy is an autosomal dominant, progressive neuromuscular disorder caused by gene mutation. Its core feature, myotonia, consists in delayed muscle relaxation following contraction, which heavily impacts a patient's daily life. The diagnosis of Myotonic dystrophy (Type 1) relies on handgrip time series data collected from standardized quantitative myotonia assessment (QMA) hardware equipped with a force transducer. Reasoning based on handcrafted features can be inaccurate, and medical experts are often required to verify resulting diagnoses. Our main contributions are as follows:

- We propose an end-to-end hierarchical attention-based temporal convolutional network (HA-TCN) to automate the handgrip time series data analysis task.
- We empirically show that the proposed HA-TCN framework performs similarly to Temporal Convolutional Network (TCN) [2] and Long Short-term Memory (LSTM) approaches, and that the deep-learning-based methods outperform support vector machines (SVM) that rely on handcrafted features.
- We demonstrate through experimental validation that HA-TCN models outperform TCNs in shallow architectural regimes because their hierarchical attention mechanisms enable them to better summarize relevant information across a wider range of time steps.
- We empirically show that the HA-TCN model can highlight key time series segments in the relaxation phase of an individual handgrip sample that differentiate patients from healthy individuals.

The remainder of this paper is organized as follows: Sec. 2 delves into technical details regarding the proposed HA-TCN framework; Sec. 3 presents the experimental setup and results; lastly, Sec. 4 concludes the paper.

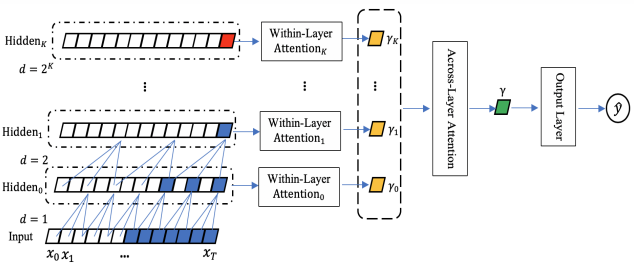


Figure 1: HA-TCN Architecture. This example network has one input layer and K hidden layers. The size of the kernel filter at each layer is 3. Dilation factor d increases exponentially with depth.

2. HA-TCN Model

Let $X \in \mathbb{R}^T$, $X = \{x_0, x_1, ..., x_T\}$ denote a time series sequence. The decision-making task of interest consists in determining whether the subject whose time series is being analyzed suffers from myotonic dystrophy. Let $y \in (0, 1)$ denote the output of the network, and $f(\cdot)$ denote the input-output functional mapping effected by the network, i.e., $y = f(X)$. The architecture of HA-TCN is shown in Fig. 1.

2.1. Causal Convolutions and Dilated Convolutions

The basic convolutional operations in the proposed HA-TCN architecture are identical to those described in the original TCN publication [2]. Causal convolutions and dilated convolutions are applied. The former ensures that only present (i.e., at time t) and past (i.e., before time t) samples are involved in the computation of the output at time t. The latter introduces a fixed dilation factor d between every pair of adjacent filter taps. As in previous work, d is increased exponentially with the depth of the network, e.g., $d = O(2^i)$ for the i-th network layer.

2.2. Hierarchical Attention Mechanism

When TCNs are used for classification tasks, the last sequential activation from the deepest layer is used (see red cell in Fig. 1) [1]; as in RNNs, such activation condenses the information extracted from the entire input sequence into one vector. We posit this representation may be too abbreviated for complex sequential problems. With this in mind, we propose the addition of hierarchical attention mechanisms across network layers. As shown in Fig. 1, suppose the HA-TCN has K hidden layers, and H_i is the matrix consisting of convolutional activations at layer i, $i = 0, 1, ..., K$; $H_i = [h_0^i, h_1^i, ..., h_T^i]$, $H_i \in \mathbb{R}^{C \times T}$, where C is the number of kernel filters at each layer. The within-layer attention weight $\alpha_i \in \mathbb{R}^{1 \times T}$ is calculated as follows:

$$\alpha_i = softmax(tanh(w_i^T H_i)) \tag{1}$$

where $w_i \in \mathbb{R}^{C \times 1}$ is a trained parameter vector and $(\cdot)^T$ denotes the transpose operation.

The combination of convolutional activations for layer i is calculated as:

$$\gamma_i = ReLU(H_i \alpha_i^T) \tag{2}$$

where $\gamma_i \in \mathbb{R}^{C \times 1}$.

After executing each within-layer attention layer, the convolutional activations are transformed into $M = [\gamma_0, \gamma_1, ..., \gamma_i, ..., \gamma_K]$, $M \in \mathbb{R}^{C \times K}$. Similarly, the across-layer attention layer takes M as the input to calculate the final sequence representation used for classification:

$$\alpha = softmax(tanh(w^T M)) \tag{3}$$

$$\gamma = ReLU(M \alpha^T) \tag{4}$$

where $w \in \mathbb{R}^{C \times 1}$, $\alpha \in \mathbb{R}^{1 \times K}$, $\gamma \in \mathbb{R}^{C \times 1}$.

2.3. Relevant Time Series Segment Identification

In previous attention-based RNN studies, the relevant sequence segments corresponding to large attention weights were difficult to identify. This is due to the fact that the hidden state of a LSTM/RNN cell not only includes information from the current time step, but also from historical time steps, the receptive field (i.e., time span) of which is for the most part unknown [5].

In contrast, the HA-TCN model can track the origin of relevant segments once the within-layer and across-layer attention estimates are available. This is because the architecture of the HA-TCN mainly consists of feedforward convolutional blocks. Specifically, if $d = 2^i$ in the dilated causal convolution operation, then the start of the receptive field at the input layer covered by a filter at time t, layer i can be calculated as follows:

$$s = max(0, t - (2^{i+1} - 1) * (l - 1)) \tag{5}$$

where s is the start time step of the receptive field at the input layer and l is the size of the kernel filter.

The receptive field at the input layer for a filter at time step t and layer i can be then represented as $RF_{s \to t}^i$, $t = 0, 1, ..., T$, $i = 0, 1, ...K$. As an example, the receptive field for the filter at time step T, hidden layer 1 is highlighted in blue in Fig. 1.

Given a series of handgrip force data points, the within-layer attention α_i and across-layer attention α can be generated with a trained HA-TCN. We rank hidden layers based on their α value, and identify the relevant layers RL with larger attention weights (e.g., the top 10 percentile attention weights in α). Similarly, we identify the relevant time steps RT with larger attention weights based on α_i, $i \in RL$.

Subsequently, we can calculate the frequency of each time step j that belongs to the relevant field $RF_{s \to t}^i$, $i \in$

Model	Accuracy	F1 Score
SVM	88.40%	0.85
LSTM	$92.38\% \pm 2.49\%$	0.94 ± 0.01
Bi-LSTM	$93.26\% \pm 1.85\%$	0.94 ± 0.01
TCN [2]	$93.02\% \pm 2.38\%$	0.93 ± 0.01
LSTM + Attention	$93.58\% \pm 1.64\%$	0.94 ± 0.01
Bi-LSTM + Attention [6]	$94.00\% \pm 2.20\%$	0.94 ± 0.01
HA-TCN	$93.82\% \pm 2.30\%$	0.95 ± 0.01

Table 1: Performance comparison: Accuracy and F1 Score (mean $\pm$ standard deviation). SVM performs the worst. Deep learning models perform similarly. Attention mechanisms slightly improve model classification accuracy.

RL, and $t \in RT$:

$$Freq_j = \sum_{i \in RL, t \in RT} I_j(RF^i_{s \to t}), j = 0, 1, ..., T \qquad (6)$$

where $I_j(*)$ is the indicator function (i.e., valued 1 if j belongs to $RF^i_{s \to t}$ and 0 otherwise). Lastly, the relevant time series segment can be identified based on corresponding high frequencies (e.g., the top 10 percentile frequencies).

3. Experiments

3.1. Dataset and Experimental Setup

467 individual handgrip time series samples from 37 patients and 270 samples from 18 healthy subjects were acquired. The baseline performance benchmark is a SVM operating on handcrafted features. The relaxation time from the 90^{th} percentile to the 5^{th} percentile of the strength in the relaxation phase, i.e., RT90-5, is extracted and the SVM trained on the extracted feature.

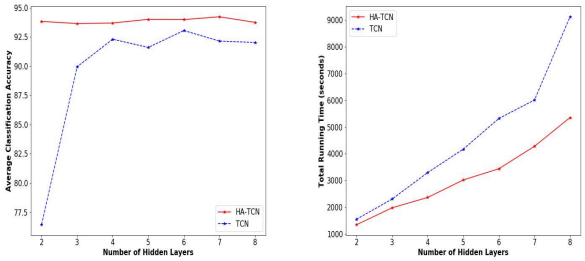
We also build LSTMs and Bidirectional LSTMs (Bi-LSTM), attention-based LSTMs [4] and attention-based Bi-LSTMs [6], as well as the traditional TCN [2] for comparison. The HA-TCN model consists of two hidden layers with dilated causal convolutions and dilation factors of 1 and 2, respectively, and a kernel size of 50. Ten-fold cross validation at the subject level is conducted, meaning that each subject is ensured to be included in the test set at least once.

3.2. Experimental Results

The average classification accuracy and F1 score of each model are shown in Table 1. For deep learning models, each fold is executed five times to account for model dependence on random weight initialization. The results show that the deep learning models outperform the SVM-based approach, which leverages handcrafted features resembling those used by doctors. All deep learning models have similar classification accuracies and F1 scores.

Next, we compared the performance of HA-TCNs with that of TCNs [2]. A kernel size of 50 was used in both models, as well as the same dilation factor of 2^i. The main difference between the models is that the former implements a hierarchical attention mechanism to combine information across time steps, while the latter only uses the activation from the last cell of the deepest hidden layer for classification. Five runs of 10-fold cross validation were conducted with network depths ranging from 2 to 8 hidden layers.

The results in Fig. 2 show that the proposed HA-TCN model achieves high classification accuracy with only two hidden layers, and that increasing network depth does not have a significant impact on model performance. In contrast, with two hidden layers, the classification accuracy of the TCN is relatively low. While its performance increases more significantly with network depth, it is lower than that of the HA-TCNs independently of network depth. We hypothesize this consistent difference in performance may be due to the reliance of the TCN on a single activation; this disadvantage can only be partially overcome with increasing network depth. Fig. 2(b) also shows that the HA-TCN model always takes less time for training than the TCN models.



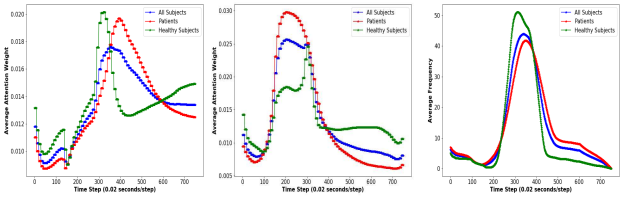
(a) (b)

Figure 2: Comparison of the HA-TCN and TCN models. (a) Average classification accuracy. (b) Total running time for five runs of 10-fold cross validation.

Figs. 3(a) and 3(b) show the average attention weights for each time step for the LSTM models. Fig. 3(c) shows the average frequency that a time step belongs to a specific receptive field. Because the HA-TCN model in this study only has two hidden layers, we select the one with the larger across-layer attention weight as the relevant layer RL, then choose those with the top 10 percentile attention weights as the relevant time steps RT. The frequency of a time step belonging to the receptive fields is calculated based on Eqs. (5) and (6). The plots in Fig. 3 also show the corresponding curves for patients and healthy subjects separately.

In Fig. 3(a), the maximum average attention weight for patients occurs at around time step 400. For healthy patients, that maximum occurs at time step 310. Similarly,

in Fig. 3(b), the peak of the average attention weight curve for the healthy group is at around the 300^{th} time step. The peak of the curve for the patients is flatter, and the time steps around the 200^{th} time step all have large attention weights. These inconsistencies indicate that the attention-based LSTMs cannot be used for interpretable diagnosis of myotonic dystrophy. In contrast, it can be seen in Fig. 3(c) that the curves for the HA-TCN model across subjects, including patients and healthy subjects, align extremely well, with peaks taking place at around the 350^{th} time step.



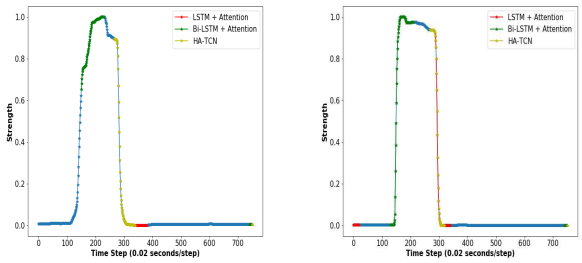
(a) LSTM+Attention (b) Bi-LSTM+Attention (c) HA-TCN

Figure 3: Explainability Comparison of Deep learning models: average attention weights per time step for (a) LSTM and (b) Bi-LSTM; (c) average frequency that the time step is within the receptive fields based on the HA-TCN.

Fig. 4 shows relevant time series segments identified by attention-based deep learning models for two examples, one belonging to a patient (Fig. 4(a)), and one sampled from a healthy subject (Fig. 4(b)). Segments not identified as relevant by any model are plotted in blue, while segments marked with colors correspond to those with the top 10 percentile attention weights or frequencies for their respective model as indicated in the legend. It can be seen that the one-directional LSTM model with attention highlights the most relevant segment at the end of the relaxation phase for the patient in Fig. 4(a), which is inconsistent with the definition of myotonia. Although the strength decreasing part for the healthy subject in Fig. 4(b) is identified by this model, it also identifies the beginning of the curve as important. For relevant segments identified by the Bi-LSTM model with attention in Figs. 4(a) and 4(b), the segments corresponding to the squeezing stage are assigned high weights, which again shows that such a model is not usable for interpretable diagnosis of myotonic dystrophy. Only the HA-TCN model can consistently identify the deceasing portion of the time series as the part most relevant for disease classification.

4. Conclusion

In this paper, we introduced the HA-TCN architecture for a case study of myotonic dystrophy diagnosis. The HA-TCN model achieves performance comparable with state-of-the-art deep learning models. All deep learning models outperform traditional machine learning approaches relying on handcrafted features. However, only the HA-TCN



(a) (b)

Figure 4: Relevant segment identification with attention-based deep learning models: (a) sample from a patient, and (b)sample from a healthy subject. Segments marked with different colors are identified as relevant segments from the deep learning models.

can highlight the relaxation phase in the handgrip time series data, which is consistent with the diagnosis criterion that clinicians have been using. Furthermore, we show that the HA-TCN outperforms the TCN in terms of performance particularly when the number of hidden layers is small, and is more computationally efficient in general.

References

[1] Sequential mnist and permuted sequential mnist. `https://github.com/locuslab/TCN/tree/master/TCN/mnist_pixel`. Accessed: 2019-03-07.

[2] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018.

[3] Y. Sha and M. D. Wang. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 233–240. ACM, 2017.

[4] S.-s. Shen and H.-y. Lee. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. *arXiv preprint arXiv:1604.00077*, 2016.

[5] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[6] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212, 2016.