Topic 2- Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline

Shelja Agarwal
IIIT Delhi
shelja20470@iiitd.ac.in

Zhiguang Wan
GE Global Research
zhiguang.wang, yan@ge.com

Weizhong Yan
GE Global Research
Email-id

Tim Oates
Computer science and
Electric Engineering University of Maryland Baltimore County oates@umbc.edu

October 28, 2021

***Abstract*—Deep neural networks provide a simple and effective baseline for time series categorization. They're completely end-to-end and don't include any significant data. When compared against other state-of-the-art techniques, the Fully Convolutional Network wins out, while ResNet is a close second. Misuse of the Class Activation Map to discover specific labels in raw data is now permissible thanks to global average pooling. This is an excellent starting point for further investigation. Generalization of capacity, characteristics, structures, and classification are all subject to study.**

Keywords-Neural Networks, Time Series, Flly Convolutional Network, ResNet, Class Activation Map, Global Average Pooling

I. Introduction

Time series data is fairly widespread and may be found almost anywhere. Univariate time series are a nice place to start since they are straightforward. Time classification and representation learning are gaining popularity. Distance-based approaches, such as Euclidean distance or Dynamic time warping [1], work directly on raw time series with similarities. Combining DTW with k-nearest neighbours is a good way to go. The features of feature-based techniques differ. To minimise time complexity and increase efficiency, the BOSSVS method [2] combines the BOSS model with a vector space model. To improve accuracy, different classifiers are merged. For the same, ensemble-based techniques are used.All of this necessitates a significant amount of data and feature engineering. For multivariate time series classification, a multi-channel CNN was used in [3]. The author also started MCNN for the same. We present a single baseline for using deep neural networks to classify time series from beginning to end. ResNet and FCN outperform COTE and MCNN in terms of pure end-to-end training.

II. Network Architectures

A. Multiplayer perceptrons

There are 500 neurons in the completely connected layers, which obey two design rules: The non-linearity is filled by

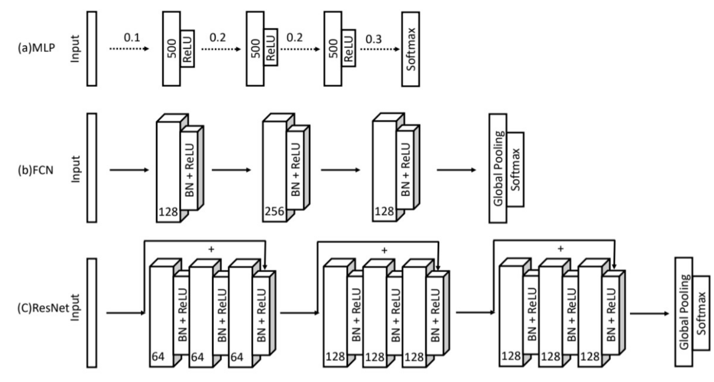

Fig. 1. The network structure of three tested neural networks. Dash line indicates the operation of dropout.

Fig. 1. The network structure of three tested neural networks. Dash line indicates the operation of dropout.

the rectified linear unit [4] utilising dropout [5] at each layer's input. A softmax layer completes the network. A fundamental layer is defined as

$$x = f_{dropout,p}(x)$$
$$y = W.x + b$$
$$h = ReLU(y)$$

Different layers have dropout rates of 0.1, 0.2, and 0.3 (Figure 1(a)).

B. Fully Convolutional Networks

For semantic segmentation on pictures, it has demonstrated astonishing quality and efficiency [6]. FCN fulfils the function of a feature extractor. However, the softmax layer is still responsible for the final result. Three 1-D kernels of size 8, 5, 3 are used in the convolution procedure without striding. The basic convolution block is as follows:

$$y = Wx + b$$
$$s = BN(y)$$
$$h = ReLU(s)$$

The final network is made up of three convolution blocks with different filter sizes. Rather of a fully linked layer, features are routed into a global average pooling layer [7] after the convolution blocks. A softmax layer creates the final label. (Figure 1(b))

C. Residual Network

By adding shortcut connections in each residual block, ResNet grows neural networks to deep structures. It performs at a high level in terms of performance and vision-related activities [8]. Each residual block is constructed using the convolutional blocks from Equation 2. The residual block is defined as follows:

$$h_1 = Block_{k1}(x)$$
$$h_2 = Block_{k2}(h_1)$$
$$h_3 = Block_{k3}(h_2)$$
$$y = h_3 + x$$
$$h = ReLU(y)$$

After a global average pooling layer and a softmax layer, the final ResNet stacks three residual blocks. As a baseline, our current structures are sufficient to deliver a qualified demonstration.(Figure 1(c))

III. Experiments and Results

A. Experiment Settings

From the beginning, the dataset is divided into training and testing sections. Adadelta [9] is used to train the MLP. The best model with the lowest training loss is chosen and its performance on the test set is revealed. The complexity of training and deploying deep learning models is greatly reduced in some scenarios.

B. Evaluation

The results and a thorough comparison with eight other standards are shown in the table below. Some writings provide a ranking system for calculating performance and comparing results to provide overall averages. Mean Per-Class Error (MPCE) is used to calculate classification performance of certain models on numerous datasets in order to propose a simple evaluation measure.

TABLE I
TESTING ERROR AND THE MEAN PER-CLASS ERROR (MPCE) ON 44 UCR TIME SERIES DATASET

Err Rate	DTW	COTE	MCNN	BOSSVS	PROP	BOSS	SE1	TSBF	MLP	FCN	ResNet
50words	0.31	0.191	0.19	0.367	0.18	0.301	0.288	0.209	0.288	0.321	0.273
SonyAIBORobot	0.275	0.146	0.23	0.265	0.293	0.321	0.238	0.175	0.273	0.032	0.015
Adiac	0.396	0.233	0.231	0.302	0.353	0.22	0.373	0.245	0.248	0.143	0.174
fish	0.177	0.029	0.051	0.017	0.034	0.011	0.057	0.08	0.126	0.029	0.011
SonyAIBORobotII	0.169	0.076	0.07	0.188	0.124	0.098	0.066	0.196	0.161	0.038	0.038
AVG Arithmetic ranking	8.205	3.682	3.932	7.318	5.545	4.614	7.455	6.614	7.909	3.977	4.386
MPCE	0.0397	0.0226	0.0241	0.0330	0.0304	0.0256	0.0302	0.0335	0.0407	0.0219	0.0231

C. Results and Analysis

COTE is a model that aggregates the weighted votes from 35 separate classifiers to create an ensemble model. BOSSVS is a collection of BOSS models with varying window lengths. As a simple standard baseline, INN-DTW is included. We've included four measures in the table to help you compare alternative techniques. The authors proposed validating the efficiency of their models. The greatest and worst MPCE scores are FCN and MLP, respectively. ResNet is ranked third, just below COTE. When comparing FCN to ResNet, ResNet has a far easier time overfitting the data.

IV. Localize the contributing regions with class Activation map

The final softmax function's input is

$$g_c = sum_k$$
$$s = BN(y)$$
$$h = ReLU(s)$$

Various categories have different contributing regions. The CAM makes it simple to locate the contributing region in the raw data for specified labels. CAM presents a method for determining a probable explanation for how convolutional networks work in the classification situation.

V. Discussion

A. Overfitting and Generalization

Because of the enormous number of parameters, neural networks are a powerful universal approximator that is prone to overfitting. Given that the training accuracy is nearly 100 percent, models generalise pretty well. Batch normalisation aids in the speed and generality of training. The fully-connected layer is being replaced by a global average pooling layer, which minimises the number of parameters.

B. Feature Visualization and Analysis

Given a series X=x1, x2,... xn, we rescale X so that all values fall in the interval [0,1] GASF was chosen because it offers a simple way to evaluate multi-scale correlations in 1-D space. The efficiency of 1-D convolution may be seen in the visualisation and classification results.

C. Deep and shallow

Deep architectural research is fascinating and instructive. ResNet includes 11 layers, but it still performs well. Batch normalisation and global average pooling have significantly improved test data performance, but they still have a tendency to overfit. We encourage the examination of the ResNet structure when the data is greater and more complex.

D. Classification Semantics

Distance-based, feature-based, and neural network-based time series categorization are the three types. The combination of distance and feature-based performance improves performance. Figure 5 shows that the distance between three baseline models and other benchmarks is relatively significant, indicating that feature and classification are complementary. FCN and ResNet are in close proximity to one another.

VI. Conclusion

It's a basic and effective baseline for using deep neural networks to classify time series. Our basic models are completely end-to-end without any data processing or feature creation. The generality of our models, learning features, network architecture, and classification semantics are all discussed in detail. Apart from ranking, MPCE is unbiased when it comes to evaluating numerous model measurements.

References

[1] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," vol. 7, no. 3, pp. 358–386. [Online]. Available: http://link.springer.com/10.1007/s10115-004-0154-9

[2] P. Schäfer, "Scalable time series classification," vol. 30, no. 5, pp. 1273–1298. [Online]. Available: http://link.springer.com/10.1007/s10618-015-0441-y

[3] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Exploiting multi-channels deep convolutional neural networks for multivariate time series classification," vol. 10, no. 1, pp. 96–112. [Online]. Available: https://doi.org/10.1007/s11704-015-4478-2

[4] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," p. 8.

[5] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," vol. 15, no. 56, pp. 1929–1958. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[6] "Fully convolutional networks for semantic segmentation."

[7] M. Lin, Q. Chen, and S. Yan, "Network in network." [Online]. Available: http://arxiv.org/abs/1312.4400

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." [Online]. Available: http://arxiv.org/abs/1512.03385

[9] M. D. Zeiler, "ADADELTA: An adaptive learning rate method." [Online]. Available: http://arxiv.org/abs/1212.5701