# DATSING: Data Augmented Time Series Forecasting with Adversarial Domain Adaptation

**Conference Paper** · October 2020

DOI: 10.1145/3340531.3412155

**3 authors**, including:

Hailin Hu
Tsinghua University
**27** PUBLICATIONS **697** CITATIONS

SEE PROFILE

Mingjian Tang
Atlassian
**30** PUBLICATIONS **368** CITATIONS

SEE PROFILE

# DATSING: Data Augmented Time Series Forecasting with Adversarial Domain Adaptation

Hailin Hu
AI Application Research Center,
Huawei Technologies
Shenzhen, China
huhailin2@huawei.com

MingJian Tang
AI Enablement, Huawei Technologies
Shenzhen, China
tang.ming.jian@huawei.com

Chengcheng Bai
John von Neumann Studio, Huawei
Technologies
Shenzhen, China
baichengcheng1@huawei.com

## ABSTRACT

Due to the high temporal uncertainty and low signal-to-noise ratio, transfer learning for univariate time series forecasting remains a challenging task. In addition, data scarcity, which is commonly encountered in business forecasting, further limits the application of conventional transfer learning protocols. In this work, we have developed, DATSING, a transfer learning-based framework that effectively leverages cross-domain time series latent representations to augment target domain forecasting. In particular, we aim to transfer domain-invariant feature representations from a pre-trained stacked deep residual network to the target domains, so as to assist the prediction of each target time series. To effectively avoid noisy feature representations, we propose a two-phased framework which first clusters similar mixed domains time series data and then performs a fine-tuning procedure with domain adversarial regularization to achieve better out-of-sample generalization. Extensive experiments with real-world datasets have demonstrated that our method significantly improves the forecasting performance of the pre-trained model. DATSING has the unique potential to empower forecasting practitioners to unleash the power of cross-domain time series data.

## CCS CONCEPTS

• **Computing methodologies → Machine learning algorithms**; **Learning paradigms**.

## KEYWORDS

time series prediction, neural networks, transfer learning, generalizability

**ACM Reference Format:**
Hailin Hu, MingJian Tang, and Chengcheng Bai. 2020. DATSING: Data Augmented Time Series Forecasting with Adversarial Domain Adaptation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412155

## 1 INTRODUCTION

Time series forecasting represents a valuable task across various fields, such as energy consumption estimation and financial forecasting [3, 5, 9]. Traditional methods, e.g., ARIMA [2] and exponential smoothing [7, 20], tend to outperform machine learning methods when time series are stationary or have systematic trend or seasonality. Recently, significant progress has been made in applying deep learning to time series forecasting [12, 13, 15–17, 19], showing encouraging indication in this direction.

The advantage of deep learning over statistical methods mainly lies in its ability to model complicated non-linear relationships and capture latent features. However, it is also prone to overfit to the training data, which calls for effective transfer learning strategies to adapt trained models to data from a novel domain, especially when the target domain has insufficient data to train an in-domain deep learning model from scratch. Unfortunately, in a realistic scenario, this transfer learning process is faced with two major challenges, i.e., dynamically changing patterns and data scarcity. While some promising results have shown the potential of training generalizable models for time series forecasting [11], the methods for how to effectively transferring general domain to a specific target domain are still lacking in the toolbox.

In this study, we develop a transfer learning framework, called DATSING (Data Augmented Time Series ForecastING with adversarial domain adaptation). In particular, we tackle the transferring learning in time series forecasting where the target domain only has scare data. To achieve this, we propose to first perform data augmentation to obtain a cluster of similar data from the general domain, and then further fine-tune the pre-trained model to enhance its performance in the target domain. With the adversarial learning-based domain adaptation procedure, we are able to systematically address the negative transfer issue embedded in the nature of time series data. Our experimental results on the real-world time series data have shown the effectiveness of our proposed framework by improving the in-domain forecasting performance.

## 2 RELATED WORK

### 2.1 Time series forecasting methods

Traditionally, the theoretical foundations of time series forecasting were deeply rooted in the statistics community. Some of the most commonly used models are ARIMA/SARIMA and Exponential Smoothing. They usually have explicit forms to encapsulate subtle temporal dynamics commonly found in time series data such as auto-correlation, seasonality and long-term trends. Since they are designed for fitting individual datasets, their extrapolation is

*CIKM '20, October 19–23, 2020, Virtual Event, Ireland*
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412155

constrained to the same time series. Furthermore, their expressive nature also limits their complexity. The recent wave of big data applications has nurtured another line of researches combining statistical methods with deep neural network architectures. For example, DeepAR [15] incorporates likelihood models into a recurrent neural network to perform probabilistic forecasting. Deep State Space [13] empowers the state space model with jointly-learned recurrent neural networks. The recent winner of M4 competition [16], employs a residual LSTM model to enhance the traditional Holt-Winters method. More recently, a pure deep learning-based network, called N-BEATS [12], has shown that stacked ultra-deep residual networks are capable of providing state-of-the-art forecasting performance. In particular, the N-BEATS model uses two residual stack branches to perform backcast and forecast at the same time, providing an intra-model dis-assembly for the complicated forecasting problem.

### 2.2 Domain adaptation methods

Domain adaptation aims to effectively transfer the model learned in the source domain to perform the same task in a target domain that has different feature distribution. Among the strategies for performing domain adaptation, adversarial training [6, 18] has gained a lot of attention, as it jointly learns a domain invariant representation scheme and domain-specific label predictors, and is applicable in both supervised (generally referred as "fine-tuning") and unsupervised (no target domain label is provided) settings.

Given the dynamic nature of time series forecasting, only a few domain adaptation studies have been conducted in this field. Specifically, [8] proposed fine-tuning CNN with layer freezing to enhance energy prediction, while [14] proposed to transfer trend factor, seasonal index and normalization statistics to new datasets with scare data. In contrast to these works, this study emphasizes on finding 1) a practical method to perform data augmentation to accomplish transfer learning with as few as one target time series sample; 2) an effective regularization method that prevents negative transfer caused by out-of-distribution (OOD) issue.

## 3 PRELIMINARIES

This work tackles the univariate time series forecasting problem. The input of our model, denoted as $[x_1, ..., x_T]$ is a sequence of historical values that are observed with an equal interval. Then a forecasting model aims to predict the future values $[x_{T+1}, ...x_{T+H}]$, where $H$ denotes the length of the forecasting horizon.

## 4 METHODOLOGY

Figure 1 shows the general pipeline of DATSING. Starting from a shared pre-trained deep learning model, our framework provides a personalized data augmentation and fine-tuning process to promote the prediction of each target time series.

### 4.1 Pre-training with general domain samples

The pre-training stage of DATSING framework aims to generate a starting point for the transfer learning process. In particular, it should provide a good representation scheme that can facilitate the forecasting across a broad domain. Here, we adopt the recently proposed deep residual network, N-BEATS, to accomplish this task. Mathematically, we can view the N-BEATS model as an end-to-end forecasting pipeline composed of three parts, i.e., the feature encoder $F$, the backcast decoder $M_b$ and the forecast decoder $M_f$.

Firstly, using the "doubly residual" connections, the layers before the last output layer of the N-BEATS model actually act as a deep neural encoder $F$. By performing a stack-wise dissection of both the forecast and backcast tasks, it generates a series of embeddings $\boldsymbol{E_f} = [\boldsymbol{e}_f^{(1)}, ..., \boldsymbol{e}_f^{(S)}]$ and $\boldsymbol{E_b} = [\boldsymbol{e}_b^{(1)}, ..., \boldsymbol{e}_b^{(S)}]$ for the forecast and backcast decoding, respectively, where $S$ denotes the stack number. Then, based on these embeddings, we can further learn the forecast decoder $M^f$ and backcast decoder $M^b$ such that

$$\boldsymbol{y} = \sum_{i=1}^{S} M^f(\boldsymbol{e}_f^{(i)}), \boldsymbol{x} = \sum_{i=1}^{S} M^b(\boldsymbol{e}_b^{(i)}) \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{x}$ represents the backcast and forecast, respectively.

To encourage the model to capture subtle trends change over time, during the pre-training stage, we slice the general domain time series using a sliding window with a step size of one, which we found to perform better than the original training protocol in [12].

### 4.2 Similarity-based data augmentation

In many forecasting tasks, the in-domain time series tend to be scare, which causes data insufficiency for conventional transfer learning protocols. Intuitively, we hypothesize that augmenting the target data with time series possessing similar input history will benefit the transfer learning process. Therefore, we here leverage a pre-computed pair-wise distance matrix to evaluate the similarity between samples. Specifically, for each time series in the target domain, we calculate its soft dynamic time warping (soft-DTW) distances [4] from all the samples in the general domain dataset. Then we select the nearest neighbors as the augmented data used for the fine-tuning towards the target domain and the original target time series is left for test. Here we do not apply other restrictions on the data selected and this can be a future direction to explore.

### 4.3 Domain adversarial transfer learning

When compared with other tasks, transfer learning in time series forecasting suffer more from the OOD issue as the underlying rule that governs time series dynamics may change over time. Therefore, merely conducting fine-tuning based on historically similar data points may cause the deep neural network to over-fitting a specific trend pattern and lose other learned rules that shall hold for time series in a domain-invariant manner.

On the other hand, we realize that the representations $\boldsymbol{E_b}$ and $\boldsymbol{E_f}$ produced by the feature encoder $F$ can reflect the hidden structures samples. Taking this into account and inspired by the previous studies on domain adaptation [6, 18], we further apply an adversarial regularization during the fine-tuning process. In particular, to encourage the feature encoder $F$ to leverage a domain-invariant representation space, we build a multi-layer perceptron (MLP) as the domain discriminator $D$ that is trained to distinguish the augmented in-domain data and randomly sampled general domain data based on the latent representations of each sample. During each iteration, three parts of loss are considered. In addition to updating the whole N-BEATS model with supervised fine-tuning, we also

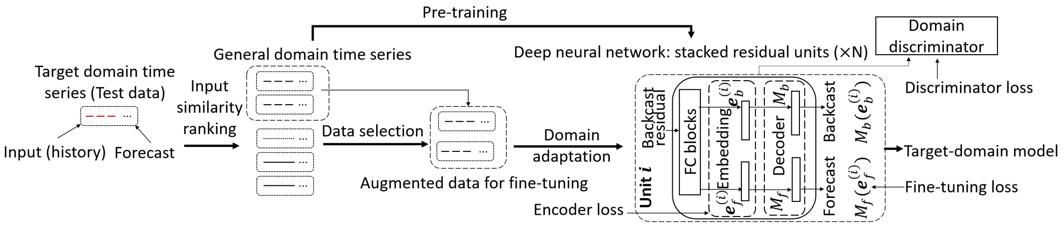

**Figure 1: A schematic illustration of the DATSING framework.**

update the domain discriminator $D$ as well as the feature encoder $F$ using the adversarial training procedure, i.e.,

$$\theta_D = \theta_D - \mu_A \frac{\partial L_D}{\partial \theta_D}, \theta_F = \theta_F - \mu_A \frac{\partial L_F}{\partial \theta_F}, \tag{2}$$

$$\theta_{F \cup M} = \theta_{F \cup M} - \mu_T \frac{\partial L_T}{\partial \theta_{F \cup M}}, \tag{3}$$

where $L_D$ denotes the cross entropy loss of the domain discriminator $D$, $L_F$ denotes the negative log likelihood loss of encoding augmented fine-tuning data, and $L_T$ is the fine-tuning loss provided by the forecasting discrepancy against the ground truth value. Note that in eq. (3) we denote the whole parameter set of the N-BEATS model as $\theta_{F \cup M}$ to reflect the model architecture.

## 5 EMPIRICAL RESULTS

### 5.1 Experiment Setup

DATSING is tested under a setting where the pre-training is conducted using a diverse general-domain time series dataset while the fine-tuning targets contain domain-specific characteristics. In particular, we use M4 competition dataset [10] as the general domain data and the tourism forecasting competition [1] dataset as the target domain data. We focus on the monthly frequency data ($N$ = 366) to mimic the most common business forecasting setting.

The dataset split and pre-processing are conducted similarly to the previous study. Specifically, for the M4 dataset, we use the official split of training and testing. The forecasting length is set consistent with the test horizon length of M4, i.e., 18 months, in both the pre-training and transfer learning process. Since the original tourism test split has a length of 24 months, we reconstruct the tourism dataset by first concatenating this original training and test part of the time series, and split the last 18 data points as the forecasting test horizon for the transfer target. Each time series is divided into a backcast part (i.e., the input feature) and the forecast part (i.e., the prediction target). To facilitate the transfer learning procedure, each time series is normalized by the maximum value of the backcast part. Following the previous protocol, we use mean absolute percentage error (MAPE) as the evaluation metric for the tourism dataset, i.e.,

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \frac{|A_i - F_i|}{|A_i|} \tag{4}$$

where $A_i$ and $F_i$ are the actual value and forecast value at time point $i$, respectively.

### 5.2 Implementation details

In our experiment, we build the transfer learning pipeline on the current state-of-the-arts deep learning model, N-BEATS [12]. To focus on the transfer procedure, we conduct our experiment mainly using one model configuration instead of the original ensemble setting. In particular, we set the backcast sequence length to be 5 times as long as the forecast length, and eliminate one tourism time series that does not possess sufficient data points in this setting. We set the training loss as MAPE and use the generic N-BEATS architecture. Other hyper-parameters for model configuration are set the same as the N-BEATS paper.

To calibrate the hyper-parameters involved in the data augmentation and transfer learning stages, we also construct a development dataset using the last horizon of the training set, i.e., the horizon right ahead of the test horizon. No test data are involved in the calibration process. Following a grid search, we set the fine-tuning batch size as 128, fine-tuning learning rate $\mu_T$ as 0.001, the size of augmented data for fine-tuning as 12,800, and the maximum iteration step as 300. For the adversarial training, we set the learning rate $\mu_A$ as $2 \times 10^{-5}$. We first perform a mean pooling with $\boldsymbol{E}_b$ and $\boldsymbol{E}_f$ and feed the results into the domain discriminator $D$, a two-branch MLP with one hidden layer of size 128. During fine-tuning, we perform model selection by choosing the iteration step showing the best symmetric mean absolute percentage error (sMAPE) performance on 10% of the domain-augmented data that are left out for validation to avoid the effect of extreme values. In our experiments, the training and evaluation were performed on one NVIDIA Tesla V100 GPU. The time cost of N-BEATS model pre-training on M4 dataset was 46 minutes and the fine-tuning using full DATSING protocol took 119.9 ±29.0 seconds per time series on average.

### 5.3 Forecasting performance

The comparison of forecasting performance between the proposed framework and baseline methods are listed in Table 1. To provide a better characterization of the experimental results, we report the average forecasting performance from two window settings, i.e., the first half of the test horizon (Points 1-9) and the whole test horizon (Points 1-18). We conduct the comparisons among zero-shot transfer learning of N-BEATS, fine-tuning without domain adversarial training (denoted as DATSING (-DA)) and the complete DATSING pipeline (denoted as DATSING (+DA)). Intriguingly, we find that DATSING (+DA) consistently performs better than the zero-shot learning under both window settings. To better characterize our improvement over the zero-shot setting, we further calculate the forecasting performance difference between N-BEATS zero-shot

**Table 1: Avergae forecasting performance (MAPE) on tourism monthly dataset. The total length of the test horizon is 18. Points 1-9 and 1-18 refer to the forecasting window that covers the first half and whole test horizon, respectively. Best performance are in bold.**

| Method | Points 1-9 | Points 1-18 |
|---|---|---|
| Auto ARIMA | 32.06 | 28.15 |
| N-BEATS zero-shot | 24.29 | 23.19 |
| DATSING (-DA) | **23.75** | 23.36 |
| DATSING (+DA) | **23.76** | **22.94** |

and DATSING (+DA) on the whole test horizon for each target time series and confirm the decrease of MAPE value is significant at 95% confidence level (mean = 0.24, median = 0.35, standard deviation = 5.9, $p = 0.015$ by one-sided Wilcoxon signed rank test). In contrast, without adversarial regularization, the fine-tuning process may suffer from negative transfer issues at later time points, showing an inferior average MAPE value on the whole test horizon. These results confirm that domain adversarial training can benefit the transfer learning process. To benchmark with a classical statistical model, we provide the performance of the ARIMA model. Though it performs well on some samples, the ARIMA model is less robust against steep trend shift compared with our proposed model, resulting in a higher average MAPE.

## 5.4 The effects of transfer learning protocol on the latent representation of time series

To probe how the transfer learning process affects feature encoder $F$ during fine-tuning, we also visualize the feature distributions of samples from the general domain, augmented fine-tuning data as well as the target domain test data under different protocols.

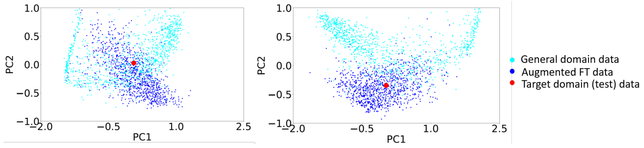

**Figure 2: Latent representations of time series for a representative example. Left: +DA; Right: -DA**

Specifically, we conduct a principal component analysis on the latent representations obtained from the fine-tuned model with and without adversarial training. A representative example using one time series in the tourism dataset is shown in Figure 2.

We find that using the conventional fine-tuning protocol (-DA), the final model tends to generate a representation scheme that drives the representation space of augmented data and test data away from the general domain, indicating a potential to over-fitting to the augmented data without retaining domain-invariant features learned in the pre-training process. On the other hand, the final representation from the +DA protocol shows much higher overlapping between the general domain data and in-domain data, showing the effectiveness of DA in regularizing the fine-tuning process.

# 6 CONCLUSION

In this work, we propose DATSING to better enable cross-domain time series forecasting. In our framework, representation learned from the general domain is further fine-tuned with regularization to strike a balance between fitting to a target domain and preserving invariant feature representation. We plan to extend the framework to the multivariate forecasting setting and more adaptive inclusion criteria to avoid noisy data in the augmentation process.

# REFERENCES

[1] George Athanasopoulos, Rob J Hyndman, Haiyan Song, and Doris C Wu. 2011. The tourism forecasting competition. *International Journal of Forecasting* 27, 3 (2011), 822–844.

[2] George EP Box, Gwilym M Jenkins, and John F MacGregor. 1974. Some recent advances in forecasting and control. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 23, 2 (1974), 158–179.

[3] Michael P Clements, Philip Hans Franses, and Norman R Swanson. 2004. Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting* 20, 2 (2004), 169–183.

[4] Marco Cuturi and Mathieu Blondel. 2017. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 894–903.

[5] Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. 2017. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews* 74 (2017), 902–924.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[7] Charles C Holt. 2004. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting* 20, 1 (2004), 5–10.

[8] Ali Hooshmand and Ratnesh Sharma. 2019. Energy Predictive Models with Limited Data using Transfer Learning. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems*. 12–16.

[9] Kenneth B Kahn. 2003. How to measure the impact of a forecast error on an enterprise? *The Journal of Business Forecasting* 22, 1 (2003), 21.

[10] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2020. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* 36, 1 (2020), 54–74.

[11] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2020. Meta-learning framework with applications to zero-shot time-series forecasting. *arXiv preprint arXiv:2002.02887* (2020).

[12] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2020. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1ecqn4YwB

[13] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. 2018. Deep state space models for time series forecasting. In *Advances in neural information processing systems*. 7785–7794.

[14] Mauro Ribeiro, Katarina Grolinger, Hany F ElYamany, Wilson A Higashino, and Miriam AM Capretz. 2018. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. *Energy and Buildings* 165 (2018), 352–363.

[15] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2019. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* (2019).

[16] Slawek Smyl. 2020. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting* 36, 1 (2020), 75–85.

[17] Qiangxing Tian, Jinxin Liu, Donglin Wang, and Ao Tang. 2019. Time Series Prediction with Interpretable Data Reconstruction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2133–2136.

[18] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7167–7176.

[19] Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. 2019. Deep Factors for Forecasting. In *International Conference on Machine Learning*. 6607–6617.

[20] Peter R Winters. 1960. Forecasting sales by exponentially weighted moving averages. *Management science* 6, 3 (1960), 324–342.