Technical Report On Time Series Analysis Based On Knowledge Distillation

Muhao Wei[1]

Abstract

Although deep learning has achieved great success on time series forecasting, two issues are unsolved. First, existing methods mainly extract features based on the local model only, which means that useful information in the specific channel cannot be fully used. Second, learning from multi-channel usually leads to an increase in the size of the model which makes it difficult to deploy. To realize the trade-off between cost and accuracy, in this study, we introduce the transfer learning and knowledge distillation and design various experiments to compared the performance with that of the original forecasting task.

1. Introduction

Time series forecasting is of paramount importance in a wide range of contemporary applications across diverse domains such as climate analysis(Ravuri et al., 2021), energy production(Wang et al., 2022), traffic flows(Li et al., 2018), financial markets, and various industrial systems(Ba et al., 2012). In the realm of climate analysis, accurate prediction of temperature patterns, precipitation levels, and other meteorological variables holds immense value for researchers and policymakers alike. Similarly, in the realm of energy production, precise forecasting of electricity demand and renewable energy generation plays a critical role in efficient resource allocation, grid management, and ensuring a stable and sustainable power supply. Transportation systems heavily rely on accurate traffic flow predictions to optimize traffic management, alleviate congestion, and enhance commuter experiences. Financial markets are renowned for their volatility and complexity, making accurate predictions of stock prices, exchange rates, and other financial indicators exceedingly challenging.

The pervasive nature and criticality of time series data have recently garnered significant attention from researchers, leading to the development of numerous deep learning forecasting models that aim to enhance the accuracy and effectiveness of time series predictions(Lim & Zohren, 2020). Leveraging advanced techniques such as Recurrent Neural Networks (RNN) and Transformers(Vaswani et al., 2017), these state-of-the-art methods excel in capturing latent representations that encapsulate the intricate dynamics of the underlying signals at each time step. By employing sophisticated prediction mechanisms, these models have made remarkable strides in the field of time series forecasting.

As a machine learning method, deep learning(Goodfellow et al., 2016) plays an important role in time series forecasting tasks by constructing deep neural network models to learn more complex feature representations and patterns from large-scale data. Compared to traditional statistical models or machine learning methods, deep learning models are better able to capture nonlinear relationships, long-term dependencies, and hidden patterns in time series data. For example, deep learning models such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) can effectively handle sequential data with temporal dependencies and model dynamic changes in time series. In addition, Transformer models based on attention mechanisms have also achieved significant results in time series forecasting tasks by parallelizing the processing of long sequence data and capturing global dependencies. Furthermore, the advantage of deep learning models in time series forecasting is also reflected in their efficient processing of large-scale data. With the development of the Internet and sensor technologies, we can access increasingly more time series data, providing deep learning models with more abundant training samples and thus improving prediction accuracy and generalization capability.

As an important approach in deep learning, transfer learning(Pan & Yang, 2010) leverages learned knowledge and experience to improve learning performance on new tasks. While deep learning models have advantages in handling large-scale data and complex tasks, training a new deep learning model may require a large amount of annotated data and computational resources in certain cases. This is where transfer learning can be useful. Transfer learning applies the knowledge and parameters of a pre-trained model from one or multiple source tasks to a target task, accelerating the learning process and improving performance on the

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. Correspondence to: Muhao Wei <weimh@lamda.nju.edu.cn>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribut.

target task. In time series forecasting tasks, transfer learning can play an important role. As time series data typically exhibit certain correlations and patterns, the knowledge and model parameters learned from one time series task can be transferred to another related time series task to improve prediction performance. This application of transfer learning can help reduce the amount of annotated data required for the new task, speed up the convergence of the model, and improve prediction accuracy.

Transfer learning improves the learning efficiency and performance of the target task by transferring the knowledge and model parameters from the source task, while knowledge distillation(Gou et al., 2020) transfers the knowledge of a complex model to a lighter model in a simplified form, thus achieving better generalization performance. Knowledge distillation can be applied in time series forecasting tasks by combining it with transfer learning. A complex model is trained on the source task, and its knowledge and parameters are then transferred to the target task of time series forecasting. By transferring the knowledge of the complex model to a simpler model in a simplified form, a more efficient, lightweight, and well-generalized model can be obtained, thereby improving the accuracy and efficiency of time series forecasting.

2. Related Work

2.1. Time Series Forecasting

Time series forecasting refers to the prediction of future values in a time series data based on patterns and trends observed in past observations. Traditional time series forecasting methods typically involve identifying the parameters of a time series model, solving for the model parameters, and using the solved model to make future predictions. One typical method is the AutoRegressive Integrated Moving Averages (ARIMA) model(Zhang, 2003), which is widely used in time series forecasting. In the process, the stability of the observed sequence needs to be tested first, followed by tests for white noise, and the computation of AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) for prediction.

Additionally, time series forecasting work is closely related to regression analysis in machine learning. Machine learning algorithms for time series forecasting can be categorized as follows:

- Support Vector Machines (SVM): SVM, based on statistical learning theory, has good generalization ability. SVM can effectively overcome the curse of dimensionality by using kernel function without increasing computational complexity. When SVM is used for time series forecasting, it is referred to as Support Vector Regression (SVR), which exhibits stable predictive capability for nonlinear time series.

- Gradient Boosting Regression Tree (GBRT): GBRT is a method that introduces gradient descent to solve regression problems. In order to minimize the loss function, GBRT utilizes negative gradients for computation and iteration, resulting in the optimal model(Elsayed et al., 2021).

- Hidden Markov Model (HMM): HMM is a statistical model that provides a probabilistic framework for modeling multivariate time series prediction(Zahari & Jaafar, 2015).

With the remarkable achievements of deep learning in computer vision and natural language processing, deep learning methods have been gradually introduced into time series forecasting applications. By constructing various network structures, deep neural networks can better represent high-dimensional data, reducing the need for manual feature engineering and model design. Through defining loss functions, end-to-end training becomes more convenient. Time Convolutional Network (TCN), proposed in literature(Chen et al., 2020), treats sequences as one-dimensional object frames and captures long-term relationships through iterative multi-layer convolutions. TCN utilizes causal convolutions, dilated convolutions, and skip connections of residual convolutions, adapting to the temporal nature of time series data and providing a wider temporal receptive field for time series modeling. DeepState Space model, based on recurrent networks, is proposed in literature(Rangapuram et al., 2018). This model estimates the relationship between two consecutive hidden states based on the idea of state space transformation, achieving predictions from the current hidden state to the current time step without requiring the input of the previous real or predicted value. This solves the problem of inconsistency between training and prediction. Transformer model, similar to GPT, was attempted for time series forecasting tasks and achieved good results in literature(Wu et al., 2020). Transformer models have the potential to improve predictive capabilities. However, Transformer models have limitations such as high computational complexity, high memory consumption, and encoder-decoder architecture, which make them unsuitable for direct application to longer time series forecasting problems. In literature(Li et al., 2019), Convolutional Self-Attention is introduced by generating queries and keys using causal convolutions in the self-attention layer. It incorporates Log-Sparse Mask in the self-attention model, reducing the computational complexity from $\mathcal{O}(L^2)$ to $\mathcal{O}(L \log L)$.

2.2. Knowledge Distillation

Knowledge distillation is a technique that aims to enhance the performance of a model while keeping the overall network architecture unchanged. The concept was initially introduced by Hinton et al.(Hinton et al., 2015), where a student model is supervised using both the hard labels and soft labels obtained from the teacher model's output. Since then, numerous studies have focused on exploring different strategies to effectively utilize soft labels for transferring knowledge between models.

One such approach is WSLD(Zhou et al., 2021), which analyzes the soft labels and assigns different weights to them based on the bias-variance trade-off. Another method called SRRL(Yang et al., 2021) enforces the output logits of both the teacher and student models to be the same, particularly after the teacher model's linear layer. DKD(Zhao et al., 2022), on the other hand, decouples the logit layer and assigns different weights to the target and non-target classes for better knowledge transfer. DIST(Huang et al., 2022) takes a different approach by using the Pearson correlation coefficient as a replacement for the traditional KL divergence, thereby transferring both inter-relation and intra-relation between the teacher and student models.

Apart from distillation based on logits, some research(Shu et al., 2021) has focused on transferring knowledge from intermediate features. FitNet (Romero et al., 2015) directly distills semantic information from the intermediate features. AT(Zagoruyko & Komodakis, 2017) transfers the attention mechanism of the teacher's feature maps to the student model. OFD(Heo et al., 2019) introduces a modified measurement for the distance between the student and teacher models, along with designing margin ReLU activation. RKD(Park et al., 2019) extracts relational information from the feature maps, while CRD(Tian et al., 2020) successfully applies contrastive learning to the distillation process. KR(Chen et al., 2021) utilizes multi-level features for knowledge transfer, and TaT(Lin et al., 2022) helps the student model learn each spatial component of the teacher model. Additionally, MGD(Yang et al., 2022) incorporates a masking technique to force the student model to generate the same features as the teacher model.

These various approaches to knowledge distillation demonstrate the extensive efforts made to improve the transfer of knowledge between models, whether through logits or intermediate features. Each technique brings its unique contribution to the field, providing valuable insights into the design and optimization of knowledge distillation methods.

3. Experiments

Applying Transformer in time series prediction can provide more accurate and efficient prediction results, which can be deployed to multi-domains as Figure 1. However, directly applying Transformer to time series forecasting problems has several limitations.Firstly, the computation complexity of the attention mechanism is high, and the resulting weights only capture a small portion of useful information. The attention mechanism establishes relationships only between individual time points, limiting its ability to extract relevant information effectively. Besides, the modeling representation of time or position is not comprehensive enough. Moreover, there is no dedicated mechanism to achieve a suitable balance between data stationarity and non-stationarity.

On this basis, PatchTST addresses these limitations by dividing the time series data into multiple contiguous time segments, referred to as patches. Each patch is then modeled and predicted individually, and the predictions from all patches are integrated to obtain the final time series forecast. The advantages of the PatchTST model lie in its ability to capture local patterns and trends in time series data. By dividing the time series into multiple patches, the model can flexibly model and predict different segments of the data, thereby improving prediction accuracy and stability. Furthermore, the PatchTST model offers good interpretability, as the contributions and effects of different patches on the overall forecast can be analyzed.

In this experiment, based on knowledge distillation, we employ PatchTST as the teacher model and Transformer as the student model, aiming to simultaneously reduce model complexity and improve model performance. We conducted several experiments on the ETT dataset, during which we fixed learning rate as 0.01 and epochs as 100 with seed as 2023.The detailed experiment results are as follows.

In the distillation model, we constructed the loss function for the training process by interpolating the mean squared error (MSE) loss between the student model's output and the ground truth, and the MSE loss between the teacher model's output and the student model's output. The interpolation was performed using a fixed hyper-parameter alpha, set to 0.1. We compared the performance of PatchTST, Transformer, and the Distillation Model on the dataset, and the experimental results are presented in the Table 1 below.

Besides, we also explore various distillation methods, such as utilizing data augmentation strategies to leverage the capabilities of the teacher model, aligning different network layers, modifying the structure of Transformer and using different distill loss functions, among others.

To enhance the robustness of the models, we performed data augmentation by adding Gaussian noise with a mean of 1 and a variance of 0 to the original data. The experimental results of this data augmentation approach are presented in the Table 2 below.

We also compared the impact of different alpha values on

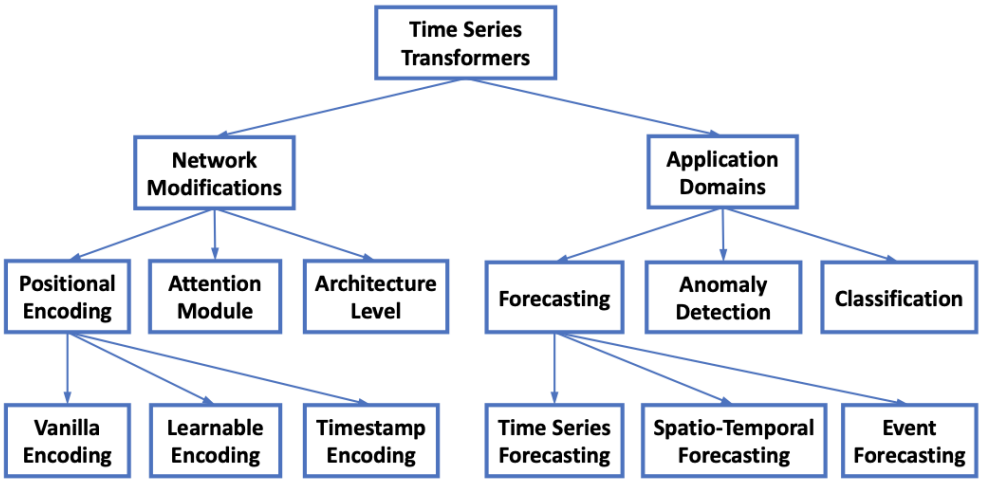

Figure 1. Tansformer Advances In the Application of Time Series

Table 1. MSE/MAE of three methods on four ETT datasets

DATASET	ETTH1	ETTH2	ETTM1	ETTM2
PATCHTST	0.7015/0.5592	0.3575/0.3901	0.7234/0.6551	0.8528/0.6624
TRANSFORMER	0.7014/0.5584	0.3518/0.3866	0.6075/0.5695	0.7596/0.6318
DISTILLATION	0.7011/0.5583	0.3577/0.3902	0.7024/0.6415	0.8529/0.6625

Table 2. Performance on Gaussian Noise Based on Distillation

DATASET	WITHOUT NOISE	WITH NOISE
ETTH1	0.7011/0.5583	0.7015/0.5594
ETTH2	0.3577/0.3902	0.3573/0.3900
ETTM1	0.7024/0.6415	0.7235/0.6550
ETTM2	0.8529/0.6625	0.8528/0.6624

Table 3. Performance on Different Alpha Based on Distillation

ALPHA	MSE	MAE
0.01	0.7015	0.5593
0.1	0.7011	0.5583
0.5	0.7010	0.5580
0.7	0.7017	0.5598
0.9	0.7018	0.5589

the performance of the distillation model. The experimental results are presented in the Table 3.

Besides, We explored the model performance by attempting modifications such as adding a linear layer in the outEmbedding. The results of these modifications are presented in the Table 4.

We compare the validation loss on the ETTh1 dataset as Figure 2 and find that the loss of the distillation model are smoother and more superior.In this experiment, we remove the earlystopping step and lr update step, and record the situation of 100 epochs.

Table 4. Performance on Adding Linear Layer Based on Distillation

DATASET	ORIGINAL	ADD LINEAR
ETTH1	0.7011/0.5583	0.7015/0.5595
ETTH2	0.3577/0.3902	0.3580/0.3905
ETTM1	0.7024/0.6415	0.7236/0.6548
ETTM2	0.8529/0.6625	0.8527/0.6624

4. Conclusion

In this technical report, we incorporate knowledge distillation into time series prediction tasks and conduct a series of experiments to explore the superiority of knowledge distillation. By combining this technique with time series prediction tasks, we can effectively improve the performance of time series regression tasks.

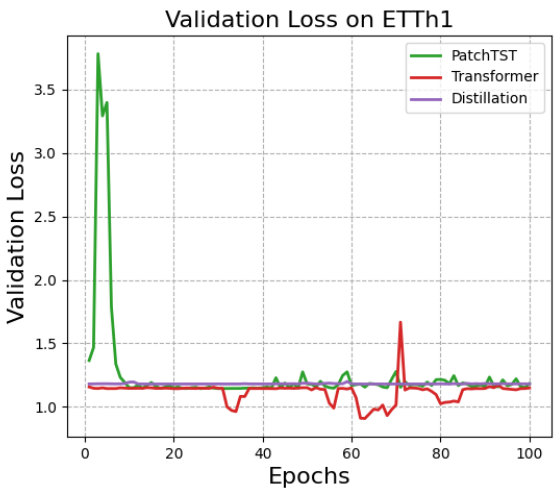

Figure 2. Validation Loss on ETTh1 Dataset with alpha=0.1

References

Ba, A., Sinn, M., Goude, Y., and Pompey, P. Adaptive learning of smoothing functions: Application to electricity load forecasting. In Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 2519–2527, 2012.

Chen, P., Liu, S., Zhao, H., and Jia, J. Distilling knowledge via knowledge review. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 5008–5017. Computer Vision Foundation / IEEE, 2021.

Chen, Y., Kang, Y., Chen, Y., and Wang, Z. Probabilistic forecasting with temporal convolutional neural network. *Neurocomputing*, 399:491–501, 2020.

Elsayed, S., Thyssens, D., Rashed, A., Schmidt-Thieme, L., and Jomaa, H. S. Do we really need deep learning models for time series forecasting? *CoRR*, abs/2101.02118, 2021.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. Knowledge distillation: A survey. *CoRR*, abs/2006.05525, 2020.

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., and Choi, J. Y. A comprehensive overhaul of feature distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 1921–1930. IEEE, 2019.

Hinton, G. E., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Huang, T., You, S., Wang, F., Qian, C., and Xu, C. Knowledge distillation from A stronger teacher. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y., and Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5244–5254, 2019.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Lim, B. and Zohren, S. Time series forecasting with deep learning: A survey. *CoRR*, abs/2004.13408, 2020.

Lin, S., Xie, H., Wang, B., Yu, K., Chang, X., Liang, X., and Wang, G. Knowledge distillation via the target-aware transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10905–10914. IEEE, 2022.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3967–3976. Computer Vision Foundation / IEEE, 2019.

Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., and Januschowski, T. Deep state space models for time series forecasting. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural*

Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp. 7796–7805, 2018.

Ravuri, S. V., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N. H., Clancy, E., Arribas, A., and Mohamed, S. Skilful precipitation nowcasting using deep generative models of radar. *Nat.*, 597(7878):672–677, 2021.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Shu, C., Liu, Y., Gao, J., Yan, Z., and Shen, C. Channel-wise knowledge distillation for dense prediction*. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 5291–5300. IEEE, 2021.

Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

Wang, Z., Xu, X., Trajcevski, G., Zhang, K., Zhong, T., and Zhou, F. Pref: Probabilistic electricity forecasting via copula-augmented state space model. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 12200–12207. AAAI Press, 2022.

Wu, N., Green, B., Ben, X., and O'Banion, S. Deep transformer models for time series forecasting: The influenza prevalence case. *CoRR*, abs/2001.08317, 2020.

Yang, J., Martínez, B., Bulat, A., and Tzimiropoulos, G. Knowledge distillation via softmax regression representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Yang, Z., Li, Z., Shao, M., Shi, D., Yuan, Z., and Yuan, C. Masked generative distillation. In Avidan, S., Brostow, G. J., Cissé, M., Farinella, G. M., and Hassner, T. (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XI*, volume 13671 of *Lecture Notes in Computer Science*, pp. 53–69. Springer, 2022.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Zahari, A. and Jaafar, J. A novel approach of hidden markov model for time series forecasting. In Kim, D. S., Kim, S., Lee, S., Hanzo, L., and Ismail, R. (eds.), *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, IMCOM 2015, Bali, Indonesia, January 08 - 10, 2015*, pp. 91:1–91:5. ACM, 2015.

Zhang, G. P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175, 2003.

Zhao, B., Cui, Q., Song, R., Qiu, Y., and Liang, J. Decoupled knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11943–11952. IEEE, 2022.

Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., and Zhang, Q. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *CoRR*, abs/2102.00650, 2021.