VOICE2SERIES: REPROGRAMMING ACOUSTIC MODELS FOR TIME SERIES CLASSIFICATION

Chao-Han Huck Yang[1] * **Yun-Yun Tsai**[2] **Pin-Yu Chen**[3]
[1]Georgia Institute of Technology, USA; *corresponding author
[2]National Tsing Hua University, Taiwan
[3]MIT-IBM Watson AI Lab, IBM Research, USA

ABSTRACT

Learning to classify time series with limited data is a practical yet challenging problem. Current methods are primarily based on hand-designed feature extraction rules or domain-specific data augmentation. Motivated by the advances in deep speech processing models and the fact that voice data are univariate temporal signals, in this paper we propose *Voice2Series* (V2S), a novel end-to-end approach that reprograms acoustic models for univariate time series classification, through input transformation learning and output label mapping. Leveraging the representation power of a large-scale pre-trained speech model, on 31 different time series tasks we show that V2S outperforms or is tied with state-of-the-art methods on 22 tasks, and improves their average accuracy by 1.72%. We further provide theoretical justification of V2S by proving its population risk is upper bounded by the source risk and a Wasserstein distance accounting for feature alignment via reprogramming. Our results offer new and effective means to time series classification.

1 INTRODUCTION

Machine learning for time series data has rich applications in a variety of domains, ranging from medical diagnosis (e.g., physiological signals such as electrocardiogram (ECG) (Kampouraki et al., 2008)), finance/weather forecasting, to industrial measurements (e.g., sensors and Internet of Things (IoT)). It is worth noting that one common practical challenge that prevents time series learning tasks from using modern large-scale deep learning models is data scarcity. While many efforts (Fawaz et al., 2018; Ye & Dai, 2018; Kashiparekh et al., 2019) have been made to advance transfer learning and model adaptation for time series classification, a principled approach is lacking and its performance may not be comparable to conventional statistical learning benchmarks (Langkvist et al., 2014).

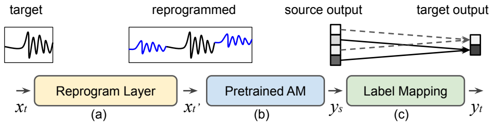


Figure 1: Schematic illustration of the proposed Voice2Series (V2S) framework: (a) trainable reprogram layer; (b) pre-trained acoustic model (AM); (c) source-target label mapping function.

To bridge this gap, we propose a novel approach, named **voice to series (V2S)**, for time series classification by *reprogramming* a pre-trained acoustic model (AM), such as a spoken-terms recognition model. Unlike general time series tasks, modern AMs are trained on massive human voice datasets and are considered as a mature technology widely deployed in intelligent electronic devices. The rationale of V2S lies in the fact that voice data can be viewed as univariate temporal signals, and therefore a well-trained AM is likely to be reprogrammed as a powerful feature extractor for solving time series classification tasks. Figure 1 shows a schematic illustration of the proposed V2S framework, including (a) a trainable reprogram layer, (b) a pre-trained AM, and (c), a specified label mapping function between source (human voice) and target (time series) labels.

Model reprogramming was firstly introduced in (Elsayed et al., 2019). The authors show that one can learn a universal input transformation function to reprogram a pre-trained ImageNet models (without changing the model weights) for solving MNIST/CIFAR-10 image classification and simple vision-based counting tasks with high accuracy. It can be viewed as an efficient approach for transfer learning with limited data, and it has achieved state-of-the-art (SOTA) results on biomedical image classification tasks (Tsai et al., 2020). Despite empirical success, little is known on how and why reprogramming can be successful. Different from existing works, this paper aims to address the following three fundamental questions: (i) Can acoustic models be reprogrammed for time series classification? (ii) Can V2S outperform SOTA time-series classification results? (iii) Is there any theoretical justification on why reprogramming works?

Our main contributions in this paper provide affirmative answers to the aforementioned fundamental questions.

1. We propose V2S, a novel and unified approach to reprogram large-scale pre-trained acoustic models for different time series classification tasks. To the best of our knowledge, V2S is the first framework that enables reprogramming for time series tasks.
2. Tested on a standard UCR time series classification benchmark (Dau et al., 2019) with 31 different tasks, V2S outperforms or is tied with the best reported results on 22 datasets and improves their average accuracy by 1.72%, suggesting that V2S is a principled and effective approach for time series classification.
3. In Section 3, we develop a theoretical risk analysis to characterize the performance of reprogramming on the target task via source risk and representation alignment loss. In Section 4, we also show how our theoretical results can be used to assess the performance of reprogramming. Moreover, we provide interpretation on V2S through auditory neural saliency map and embedding visualization.

2 VOICE2SERIES (V2S)

Throughout this paper, we will denote a K-way acoustic classification model pre-trained on voice data as a *source* model, and use the term *target* data to denote the univariate time-series data to be reprogrammed. The notation P is reserved for denoting a probability function.

2.1 V2S REPROGRAMMING ON DATA INPUTS

Here we formulate the problem of V2S reprogramming on data inputs. Let $x_t \in \mathcal{X}_\mathcal{T} \subseteq \mathbb{R}^{d_\mathcal{T}}$ denote a univariate time series input from the target domain with $d_\mathcal{T}$ temporal features.

Our V2S aims to find a trainable input transformation function $\mathcal{H}$ that is universal to all target data inputs, which serves the purpose of reprogramming x_t into the source data space $\mathcal{X}_\mathcal{S} \subseteq \mathbb{R}^{d_\mathcal{S}}$, where $d_\mathcal{T} \leq d_\mathcal{S}$. Specifically, the reprogrammed sample x'_t is formulated as:

$$x'_t = \mathcal{H}(x_t; \theta) := \text{Pad}(x_t) + \underbrace{M \odot \theta}_{\triangleq\, \delta} \tag{1}$$

where Pad(x_t) is a zero padding function that outputs a zero-padded time series of dimension $d_\mathcal{S}$. The location of the segment x_t to be placed in x'_t is a design parameter and we defer the discussion to Section B.1. The term $M \in \{0, 1\}^{d_\mathcal{S}}$ is a binary mask that indicates the location of x_t in its zero-padded input Pad(x_t), where the i-th entry of M is 0 if x_t is present (indicating the entry is non-reprogrammable), and it is 1 otherwise (indicating the entry is not occupied and thus reprogrammable). The $\odot$ operator denotes element-wise product. Finally, $\theta \in \mathbb{R}^{d_\mathcal{S}}$ is a set of trainable parameters for aligning source and target domain data distributions. One can consider a more complex function $W(\theta)$ in our reprogramming function. But in practice we do not observe notable gain when compared to the simple function θ. In what follows, we will use the term $\delta \triangleq M \odot \theta$ to denote the trainable additive input transformation for V2S reprogramming. Moreover, for ease of representation we will omit the padding notation and simply use $x_t + \delta$ to denote the reprogrammed target data, by treating the "$+$" operation as a zero-padded broadcasting function.

2.2 V2S REPROGRAMMING ON ACOUSTIC MODELS (AMS)

We select a pre-trained deep acoustic classification model as the source model ($f_\mathcal{S}$) for model reprogramming. We assume the source model has softmax as the final layer and outputs nonnegative confidence score (prediction probability) for each source label. With the transformed data inputs $\mathcal{H}(x_t;\theta)$ described in (1), one can obtain the class prediction of the source model $f_{\mathcal{S}}$ on an reprogrammed target data sample x_t, denoted by:

$$P(y_s|f_{\mathcal{S}}(\mathcal{H}(x_t;\theta))),\ \text{for all } y_s \in \mathcal{Y}_{\mathcal{S}} \tag{2}$$

Next, as illustrated in Figure 1, we assign a (many-to-one) label mapping function h to map source labels to target labels. For a target label $y_t \in \mathcal{Y}_{\mathcal{T}}$, its class prediction will be the averaged class predictions over the set of source labels assigned to it. We use the term $P(h(\mathcal{Y}_{\mathcal{S}})|f_{\mathcal{S}}(\mathcal{H}(x_t;\theta)))$ to denote the prediction probability of the target task on the associated ground-truth target label $y_t = h(\mathcal{Y}_{\mathcal{S}})$. Finally, we learn the optimal parameters θ^* for data input reprogramming by optimizing the following objective:

$$\begin{aligned}\theta^* &= \arg\min_{\theta} \underbrace{-\log P(h(\mathcal{Y}_{\mathcal{S}})|f_{\mathcal{S}}(\mathcal{H}(x_t;\theta))}_{\text{V2S loss} \triangleq L}; \\ &\text{where} \quad h\left(\mathcal{Y}_{\mathcal{S}}\right) = y_t\end{aligned} \tag{3}$$

The optimization will be implemented by minimizing the empirical loss (V2S loss L) evaluated on all target-domain training data pairs $\{x_t, y_t\}$ for solving θ^*.

In practice, we find that many-to-one label mapping can improve the reprogramming accuracy when compared to one-to-one label mapping, similar to the findings in (Tsai et al., 2020). Below we make a concrete example on how many-to-one label mapping is used for V2S reprogramming. Consider the case of reprogramming spoken-term AM for ECG classification. One can choose to map multiple (but non-overlapping) classes from the source task (e.g., 'yes', 'no', 'up', 'down' in AM classes) to every class from the target task (e.g., 'Normal' or 'Ischemia' in ECG classes), leading to a specified mapping function h. Let $\mathcal{B} \subset \mathcal{Y}_{\mathcal{S}}$ denote the set of source labels mapping to the target label $y_t \in \mathcal{Y}_{\mathcal{T}}$. Then, the class prediction of y_t based on V2S reprogramming is the aggregated prediction over the assigned source labels, which is defined as:

$$P(y_t|f_{\mathcal{S}}(\mathcal{H}(x_t;\theta)) = \frac{1}{|\mathcal{B}|}\sum_{y_s \in \mathcal{B}} P(y_s|f_{\mathcal{S}}(\mathcal{H}(x_t;\theta)) \tag{4}$$

where $|\mathcal{B}|$ denotes the number of labels in $\mathcal{B}$. In our implementation we use random (but non-overlapping) many-to-one mapping between source and target labels. Each target label is assigned with the same number of source labels. We report our test accuracy based on hyperparameters determined by 10-fold cross-validation on the training data.

3 Population Risk via Reprogramming

Theorem 1: Let δ^* denote the learned additive input transformation for reprogramming. The population risk for the target task via reprogramming a K-way source neural network classifier $f_{\mathcal{S}}(\cdot) = \eta(z_{\mathcal{S}}(\cdot))$, denoted by $\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[\ell_{\mathcal{T}}(x_t+\delta^*, y_t)]$, is upper bounded by:

$$\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[\ell_{\mathcal{T}}(x_t+\delta^*, y_t)] \leq \underbrace{\epsilon_{\mathcal{S}}}_{\text{source risk}} + 2\sqrt{K}\cdot \underbrace{\mathcal{W}_1(\mu(z_{\mathcal{S}}(x_t+\delta^*)), \mu(z_{\mathcal{S}}(x_s)))_{x_t\sim\mathcal{D}_{\mathcal{T}},\ x_s\sim\mathcal{D}_{\mathcal{S}}}}_{\text{representation alignment loss via reprogramming}}$$

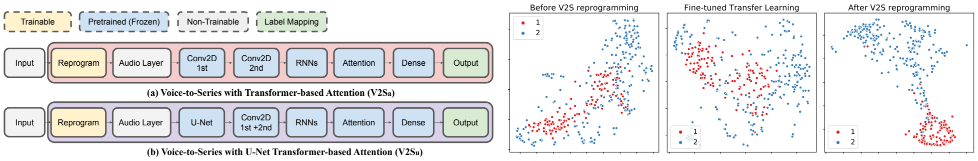
Theorem 1 shows that the target population risk via reprogramming is upper bounded by the summation of two terms: (i) the source population risk $\epsilon_{\mathcal{S}}$, and (ii) the representation alignment loss in the logit layer between the source data $z_{\mathcal{S}}(x_s)$ and the reprogrammed target data $z_{\mathcal{S}}(x_t+\delta^*)$ based on the same source neural network classifier $f_{\mathcal{S}}(\cdot) = \eta(z_{\mathcal{S}}(\cdot))$, measured by their Wasserstein-1 distance. The results suggest that reprogramming can perform better (lower risk) when the source model has a lower source loss and smaller representation loss. Proof will be given in the full version.

4 Performance Evaluation

Limited-vocabulary Voice Commends Dataset: To create a large-scale ($\sim$100k training samples) pre-trained acoustic model for our experiments, we select the Google Speech Commands V2 (Warden, 2018) dataset, which contains 105,829 utterances of 35 words with a sampling rate of 16 kHz..

Table 1: Performance comparison of validation (test) accuracy (%) on 31 UCR time series classification datasets (Dau et al., 2019). Our proposed V2S$_a$ outperforms or ties with some current SOTA results (Cabello et al., 2020; Wang et al., 2017; Lines et al., 2018) on 22 out of 31 datasets.

Dataset	Type	Input size	Train. Data	Class	SOTA	V2S$_a$	V2S$_u$	TF$_a$
Beef	SPECTRO	470	30	5	**93.33**	90.33	90.00	20
Coffee	SPECTRO	286	28	2	**100**	**100**	**100**	53.57
DistalPhalanxTW	IMAGE	80	400	6	**79.28**	**79.28**	75.34	70.21
ECG 200	ECG	96	100	2	90.9	**100**	**100**	**100**
ECG 5000	ECG	140	500	5	**94.62**	93.96	93.11	58.37
Earthquakes	SENSOR	512	322	2	76.91	**78.42**	76.45	74.82
FordA	SENSOR	500	2500	2	96.44	**100**	**100**	**100**
FordB	SENSOR	500	3636	2	92.86	**100**	**100**	**100**
GunPoint	MOTION	150	50	2	**100**	96.67	93.33	49.33
HAM	SPECTROM	431	109	2	**83.6**	78.1	71.43	51.42
HandOutlines	IMAGE	2709	1000	2	**93.24**	**93.24**	91.08	64.05
Haptics	MOTION	1092	155	5	51.95	**52.27**	50.32	21.75
Herring	IMAGE	512	64	2	**68.75**	**68.75**	64.06	59.37
ItalyPowerDemand	SENSOR	24	67	2	97.06	**97.08**	96.31	97
Lightning2	SENSOR	637	60	2	86.89	**100**	**100**	**100**
MiddlePhalanxOutlineCorrect	IMAGE	80	600	2	72.23	**83.51**	81.79	57.04
MiddlePhalanxTW	IMAGE	80	399	6	58.69	**65.58**	63.64	27.27
Plane	SENSOR	144	105	7	**100**	**100**	**100**	9.52
ProximalPhalanxOutlineAgeGroup	IMAGE	80	400	3	88.09	**88.78**	87.8	48.78
ProximalPhalanxOutlineCorrect	IMAGE	80	600	2	**92.1**	**92.1**	90.03	68.38
ProximalPhalanxTW	IMAGE	80	400	6	81.86	**84.88**	83.41	35.12
SmallKitchenAppliances	DEVICE	720	375	3	**85.33**	83.47	74.93	33.33
SonyAIBORobotSurface	SENSOR	70	20	2	**96.02**	**96.02**	91.71	34.23
Strawberry	SPECTRO	235	613	2	**98.1**	97.57	91.89	64.32
SyntheticControl	SIMULATED	60	300	6	**100**	98	99	49.33
Trace	SENSOR	271	100	4	**100**	**100**	**100**	18.99
TwoLeadECG	ECG	82	23	2	**100**	96.66	97.81	49.95
Wafer	SENSOR	152	1000	2	99.98	**100**	**100**	100
WormsTwoClass	MOTION	900	181	2	83.12	**98.7**	90.91	57.14
Worms	MOTION	900	181	5	80.17	**83.12**	80.34	42.85
Wine	SPECTRO	234	57	2	**92.61**	90.74	90.74	50
Mean accuracy (↑)	-	-	-	-	88.19	**89.91**	87.92	56.97
Median accuracy (↑)	-	-	-	-	92.61	**93.96**	91.08	53.57



(a) Voice-to-Series with Transformer-based Attention (V2Sa)

(b) Voice-to-Series with U-Net Transformer-based Attention (V2Su)

Figure 2: Left: V2S architectures. (a) V2S$_a$ (de Andrade et al., 2018) & (b) V2S$_u$ (Yang et al., 2020). Right: tSNE plots of the logit representations using Strawberry training set (Holland et al., 1998) and V2S$_a$, for the cases of before and after V2S reprogramming, and fine-tuned transfer learning (TF$_a$).

Transformer-based AMs: For training the source model, we use a popular transformer based single-head self-attention architecture (de Andrade et al., 2018) for V2S reprogramming, denoted as V2S$_a$ (Figure 2 (a)). We also train a similar architecture with U-Net (Long et al., 2015), denoted as V2S$_u$ (Figure 2 (b)), which is designed to enhance feature extraction in acoustic tasks (Yang et al., 2020).

Reprogramming Performance: Table 1 summarizes the performance of each method on 31 datasets. Notably, our reprogrammed V2S$_a$ model attains better or equivalent results in 22 over the 31 univariate UCR datasets, suggesting that V2S as a single method is a competitive and promising approach for time series classification. The transfer learning baseline TF$_a$ has poor performance, which can be attributed to limited training data. V2S$_a$ has higher mean/median accuracy, increased by 1.72/1.35%. and lower MPCE (Wang et al., 2017) (relative error decreases by about 2.87%) than that of SOTA results, demonstrating the effectiveness of V2S. For most datasets, V2S$_a$ has better performance than V2S$_u$, which can be explained by Theorem 1 through a lower empirical target risk upper bound.

5 Conclusion

In this work, we proposed V2S, a novel approach to reprogram a pre-trained acoustic model for time series classification. We also developed a theoretical risk analysis to characterize the reprogramming performance. Experimental results on UCR benchmark showed superior performance of V2S, by achieving new (or equal) state-of-the-art accuracy on 22 out of 31 datasets. We further provide in-depth discussion on the success of V2S through representation alignment, acoustic saliency map, and embedding visualization in the full version for additional and future studies.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.

Anthony Bagnall, Jason Lines, Jon Hills, and Aaron Bostrom. Time-series classification with cote: the collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2522–2535, 2015.

Nestor Cabello, Elham Naghizade, Jianzhong Qi, and Lars Kulik. Fast and accurate time series classification through supervised interval search. *ICDM 2020*, 2020.

Keunwoo Choi, Deokjin Joo, and Juho Kim. Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with keras. In *Machine Learning for Music Discovery Workshop at 34th International Conference on Machine Learning*. ICML, 2017.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Luke M Davis. *Predictive modelling of bone ageing*. PhD thesis, University of East Anglia, 2013.

Douglas Coimbra de Andrade, Sabato Leo, Martin Loesener Da Silva Viana, and Christoph Bernkopf. A neural attention model for speech command recognition. *arXiv preprint arXiv:1808.08929*, 2018.

Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. In *International Conference on Learning Representations*, 2019.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Transfer learning for time series classification. In *2018 IEEE international conference on big data (Big Data)*, pp. 1367–1376. IEEE, 2018.

Jonathan B Fritz, Mounya Elhilali, Stephen V David, and Shihab A Shamma. Auditory attention—focusing the searchlight on sound. *Current opinion in neurobiology*, 17(4):437–455, 2007.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Jon Hills, Jason Lines, Edgaras Baranauskas, James Mapp, and Anthony Bagnall. Classification of time series by shapelet transformation. *Data mining and knowledge discovery*, 28(4):851–881, 2014.

JK Holland, EK Kemsley, and RH Wilson. Use of fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purees. *Journal of the Science of Food and Agriculture*, 76(2):263–269, 1998.

Argyro Kampouraki, George Manis, and Christophoros Nikou. Heartbeat time series classification with support vector machines. *IEEE Transactions on Information Technology in Biomedicine*, 13 (4):512–518, 2008.

L.V. Kantorovich and G. Rubinstein. On a space of completely additive functions. In *Vestnik Leningradskogo Universiteta*, volume 13 (7), pp. 52–59, 1958.

Kathan Kashiparekh, Jyoti Narwariya, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. Convtimenet: A pre-trained deep convolutional neural network for time series classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Emine Merve Kaya and Mounya Elhilali. Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160101, 2017.

Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. Sliced wasserstein distance for learning gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3427–3436, 2018.

Martin Langkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24, 2014.

Jason Lines, Sarah Taylor, and Anthony Bagnall. Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles. *ACM Transactions on Knowledge Discovery from Data*, 12(5), 2018.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

G Peyré and M Cuturi. Computational optimal transport. arxiv e-prints. *arXiv preprint arXiv:1803.00567*, 2018.

Chotirat Ann Ratanamahatana and Eamonn Keogh. Three myths about dynamic time warping data mining. In *Proceedings of the 2005 SIAM international conference on data mining*, pp. 506–510. SIAM, 2005.

Patrick Schäfer. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015.

Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning*, pp. 9614–9624, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Richard Veale, Ziad M Hafed, and Masatoshi Yoshida. How is visual salience computed in the brain? insights from behaviour, neurobiology and modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160113, 2017.

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pp. 1578–1585. IEEE, 2017.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Yuzhong Wu and Tan Lee. Enhancing sound texture in cnn-based acoustic scene classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 815–819. IEEE, 2019.

Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Pin-Yu Chen, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee. Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. *arXiv preprint arXiv:2010.13309*, 2020.

Rui Ye and Qun Dai. A novel transfer learning framework for time series forecasting. *Knowledge-Based Systems*, 156:74–99, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.