On the suitability of Data Selection for Cross-building Knowledge Transfer

Mouna Labiadh, Christian Obrecht, Catarina Ferreira da Silva, Parisa Ghodous

On the suitability of Data Selection for Cross-building Knowledge Transfer

Mouna Labiadh*†, Christian Obrecht†, Catarina Ferreira da Silva*, Parisa Ghodous*
Univ Lyon, CNRS, INSA-Lyon, Université Claude Bernard Lyon 1,
* *LIRIS UMR5205, F-69100,*
† *CETHIL UMR5008, F-69621,*
Villeurbanne, France
{mouna.labiadh, catarina.ferreira, parisa.ghodous}@univ-lyon1.fr,
christian.obrecht@insa-lyon.fr

***Abstract*—Supervised deep learning has achieved remarkable success in various applications. Such advances were mainly attributed to the rise of computational powers and the amounts of training data made available. Therefore, accurate large-scale data collection services are often needed. Once representative data is retrieved, it becomes possible to train the supervised machine learning predictor. However, a model trained on existing data, that generally comes from multiple datasets, might generalize poorly on the unseen target data. This problem is referred to as *domain shift*. In this paper, we explore the suitability of data selection to tackle the domain shift challenge in the context of domain generalization. We perform our experimental study on the use case of building energy consumption prediction. Experimental results suggest that minimal building description is capable of improving cross-building generalization performances when used to select data.**

***Keywords*-Data selection; domain generalization; knowledge transfer; data-driven modeling; energy consumption modeling;**

I. Introduction

Machine learning is dramatically improving a wide range of fields, from image and natural language processing, to genomics and robotics. However, machine learning models performance depends heavily on the representation of data they are provided. This makes feature engineering the key determining factor and the most challenging step in the machine learning pipeline. Deep learning [1] alleviates this central challenge by automatically learning meaningful representations out of multiple simpler representations. The recent success of deep learning is mainly attributed to the rise of computational powers and the massive amount of available data used for training.

A powerful machine learning model should rely on insightful utilization of relevant data. Hence, scalable and accurate data collection techniques are considered as key success factors. Data collection for machine learning [2] generally consists in acquiring new relevant datasets [3], labeling acquired data samples, improving the quality of existing data or the training of an existing model. A classic data collection workflow as described in [2] would start by checking if there is enough training data. Otherwise, we must either acquire relevant datasets, or generate an appropriate dataset. Once data is collected, we can choose to either label individual data samples within these datasets, to improve the quality of existing labels that may be noisy and biased, or to select an existing model and improve it using transfer learning techniques.

We seek to explore a main challenge of using existing data, the domain shift problem. Domain shift [4] causes models trained on one source domain to generalize poorly when applied to a target domain with mismatching data distribution. Consequently, learning scenarios in which we do not have enough and exactly representative training data of the intended testing context are heavily penalized. Proposed approaches addressing this challenge are mainly classified into *domain adaptation* and *domain generalization*. Domain adaptation [5][6] utilizes labeled source data and unlabeled or sparsely labeled target data to obtain a well-performing model on the target domain. However, in several cases, the target data are not available. Domain Generalization (DG) [7][8] addresses such cases by utilizing multiple source domains. This paper considers the domain generalization area of research, which aims to train accurate models that perform well on unseen target domains, by leveraging knowledge from different but related source domains. For this purpose, we propose to explore the suitability of data selection in the context of domain generalization. This approach consists in selecting representative data from domains that are similar to the target domain on which we do not dispose of enough data. This similarity is based on available minimal domains descriptions.

Particularly, we consider the problem of data-driven building energy modeling, which aims to accurately predict future energy use from specific measures. Energy demand prediction plays an integral part in the efficient planning and operation of power systems. Prior studies in this framework require labeled data of the building in question, such as historical data, physical parameters of the building, meteorological conditions, or information about the building occupancy, in order to train a reliable building energy consumption model. Our approach goes beyond state-of-the-art methods and proposes to transfer knowledge across

multiple sources buildings while using minimal information about the target building, such as in the case of renovated or newly-built buildings. Our challenge is to build a model that accurately predicts the future energy consumption of a previously unseen building, given one or many training datasets.

To tackle these challenges, we investigate the suitability of data selection mechanism for cross-building domain generalization. To the best of our knowledge, our work is a first attempt to model a target building with incomplete or minimal information about it, and thus tackling the data unavailability problem by transferring knowledge from auxiliary buildings. Reported energy prediction approaches [9][10] usually rely on detailed information about the target building, e.g. historical consumption data, meteorological data, occupancy information, etc.

The remainder of this paper is structured as follows. Section II presents a classification of related works on domain generalization. Section III provides an overview on our proposed approach and the architecture of the model we utilize. Section IV depicts the experimental setup and summarizes results. Section V discusses experimental findings, and finally in Section VI, we draw conclusions and present an outlook and suggestions for future research.

II. Approaches to Domain Generalization

Proposed domain generalization approaches generally rely on the assumption that source domains and unseen target domains share common features that can be extracted. Hence, they seek to learn a domain agnostic representation or model. Domain generalization approaches proposed in literature may be roughly classified into four categories; (1) Data representation based techniques [8][11] that seek to learn domain agnostic representation that captures similarities across domains and where the domain discrepancy is minimized, (2) Model selection techniques [12] that aim to select the most relevant domain to a target sample and use its corresponding model, (3) Model based techniques [13][14] that aim to find a model architecture and algorithm that generalizes well on unseen target domains, and (4) Meta-learning based technique [15] that relies on a model agnostic training procedure that trains any given model so that it mitigates domain shift between domains.

Muandet et al. [8] propose to learn new domain invariant feature representations by minimizing the dissimilarity across domains via domain-invariant component analysis and a kernel-based optimization algorithm. Ghifary et al. [11] propose a Multi-Task Auto-Encoder (MTAE) that extends auto-encoders into a model that jointly learns to perform self-domain data reconstruction and between-domain data reconstruction. Xu et al. [12] use learned low-rank exemplar-SVMs, which can be defined as a linear Support Vector Machine (SVM) classifier trained on a single positive training instance and all negative training instances, for both domain adaptation and domain generalization. For domain generalization, the authors propose to either equally fuse all exemplar classifiers, or use the exemplar classifiers in the latent domain which the target data more likely belongs to. Given multiple source datasets/domains, Khosla et al.[13] propose an SVM based approach, in which the learned weight vectors are common to all datasets. Li et al. [14] proposed a low-rank parameterized convolutional neural network model for end-to-end DG learning. Li et al. [15] propose a Meta-Learning Domain Generalization (MLDG) approach. It consists in a model agnostic training procedure that can improve the domain generality of a base learner. This procedure is based on synthesizing virtual training and virtual testing domains within each mini-batch. The meta-optimization objective consists in minimizing the loss in the training domains, while simultaneously improving the virtual testing loss.

Our work is more related to the model selection techniques. We borrow the per-domain model building idea described in [12]. However, we select domains rather than models and combine their respective data to form a representative training set. We assume in our case that we dispose of a minimal description of the target domain that will allow us to define our data selection criteria, such as building typology, and year of construction.

III. The Proposed System

We present an overview of our methodology pipeline in Figure 1. We propose a cloud-based system for relevant existing data collection and reuse in predictive modeling tasks. Our system main objective is to train a building energy model for an unseen target building based on its description. The training data is obtained through a data selection service-oriented workflow.

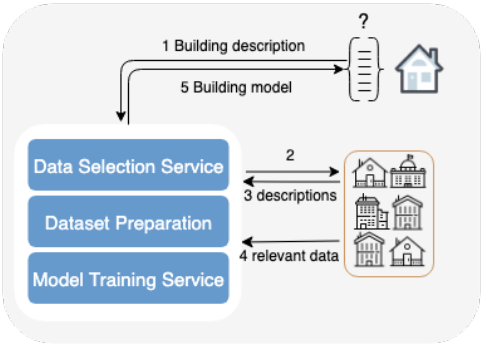

Figure 1. Data selection approach for domain generalization.

Our approach consists in training a model that simultaneously minimizes the prediction error and the domain discrepancy between the collected source domains and the unseen target domain. The data selection service is responsible for preparing a training dataset that is relevant against the target domain description. In our special case, target

domain description concerns high-level information about the target building we seek to model, e.g. typology, year of construction, location, etc. Once data is selected, the final training dataset is prepared and provided to the model training component. In this study, we attempt to explore the suitability of similar data selection in the context of building energy consumption modeling. The detailed description of the data selection workflow is beyond the scope of this paper.

A. Cross-building Knowledge Transfer

We explore the suitability of data selection for cross-building knowledge transfer. We start by training a model that captures energy use dynamics for each building, and then test its generalization performance across other unseen buildings. Our objective is to study the possibility to select representative buildings data based on available target building metadata. For this purpose, we start by identifying similar buildings based solely on their respective descriptions. We then perform cross-building knowledge transfer; we build one model for each building and study its transfer-ability across other unseen buildings. Our aim is to investigate whether cross-building performances are correlated with buildings' descriptions.

Our model learns to predict future building-level aggregate energy consumption based on energy consumption history and both past and future climate data. In reality, a wide range of factors may impact the energy use in buildings, such as occupants behavior, building typology, construction materials, etc. In this work, we focus on the meteorological data factor by feeding our model with past and future climate data along with the aggregate past energy consumption. The motivation behind utilizing both future and past climate data is to attempt to capture the relationship between climatic changes and building's energy load profile fluctuations.

B. Model Architecture

Figure 2 shows the architecture of our learning model. We consider a unidirectional Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) as our supervised predictor. RNNs [16] are a powerful class of supervised machine learning models that are capable of modeling sequential data. LSTM [17] is a RNN architecture that helps to prevent the effect of vanishing and exploding gradients [18] often encountered in recurrent networks. LSTM offers the ability to selectively pass information across sequence steps while processing sequential data one element at a time.

Our model is trained to predict daily energy consumption of subsequent week. As input, we provide our model with daily energy consumption of the previous week and climate time series of the subsequent week.

Our training set $\mathcal{X} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, x^{(2)}), ...\}$ is structured into time-based sequences of fixed length. Input sequences are denoted by $(x^{(1)}, x^{(2)}, ..., x^{(T)})$ where

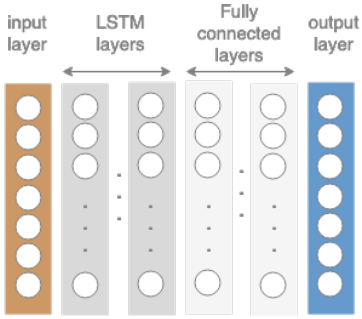

Figure 2. Architecture of the LSTM-RNN model used to predict daily energy consumption of a whole week.

T denotes the sequence length, and each feature vector $x^{(t)} \in \mathbb{R}^n \, \forall t = 1..T$ is of size n. Feature vectors are composed of current week's aggregate energy consumption, air temperature, average horizontal solar irradiance, wind speed, and the same weather features as aforementioned for subsequent week. Similarly, target sequences are denoted by $(y^{(1)}, y^{(2)}, ..., y^{(T)})$, where $y^{(t)} \in \mathbb{R}$ is a real vector denoting the energy consumption at future time steps. The goal of the model is to predict future energy consumption $y^{(t)}$ from the input feature vector $x^{(t)}$.

The architecture of the network is composed of several hidden layers. It consists of one or more LSTM layers followed by one or more fully-connected layers. The output layer is a fully-connected layer with a linear activation function. The model is trained using the Root Mean Squared Error (RMSE). We also use the batch normalization mechanism [19] to address the internal covariate shift problem usually encountered in deep neural networks training. Training phase were conducted using Backpropagation Through Time (BPTT) optimization algorithm in the context of LSTM networks.

During our experimental study, we explore variants of this architecture to fine-tune its hyperparameters, e.g. number of fully-connected layers, number of LSTM layers, etc. We retain the architecture variant that yields the best cross-domain and in-domain generalization results.

IV. Experimental Setup

We perform our experimental studies on the use case of building energy consumption prediction. Our system transfers knowledge from several buildings, to one target building on which we assume we are facing a data unavailability problem.

A. Dataset

The proposed solution is experimentally evaluated using REFIT Electrical Load Measurements dataset [20]. The dataset contains cleaned electrical consumption measurements for 20 UK households on aggregate and appliance level. For each household, the whole house aggregate loads

and nine individual appliance measurements at 8-second intervals were collected continuously over a period of approximately two years. During monitoring, the occupants were conducting their usual routines.

In addition, climate data was also collected from a nearby weather station. Figure 3 highlights the differences of energy load profiles across a subset of four buildings in the REFIT dataset. Descriptions about each building comprised information related to occupancy (number, ages, sex, etc.), size, construction year, typology, and total number of appliances owned.

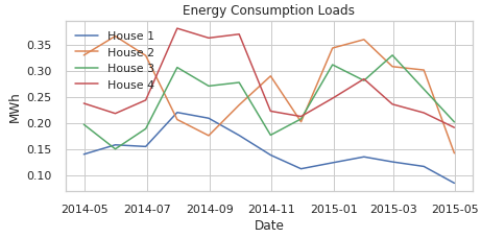

Figure 3. Monthly energy load profiles across buildings.

In Figure 4, we illustrate the REFIT dataset description with a heatmap. We consider five descriptive features for each building; the number of occupants, the construction year, the number of appliances, the building type, and the size. The number of occupants in the REFIT dataset varies from one to four occupants. The construction years of buildings are grouped into eight classes based on year intervals spanning from 1850 to post 2002. Three house types are present in the REFIT dataset; detached, semi-detached, and mid-terrace. Building sizes are computed based on number of bedrooms.

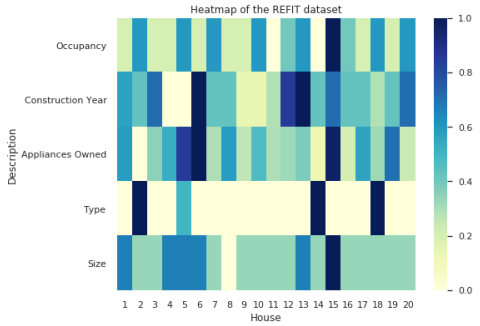

Figure 4. Heatmap of the REFIT dataset description after pre-processing; Missing data in one column were replaced with the most frequent value in that column, categorical features (type) were label encoded, ordinal features (construction year class) were converted to ordinal integers, resulted values were scaled between 0 and 1.

To depict similarities between buildings, we start by hierarchically clustering them based on the provided description vectors. Categorical data was one-hot encoded as a further pre-processing step. We use the Euclidean distance to compute pair-wise similarities. Clustering results are illustrated in Figure 5 by a dendrogram. The figure identifies a cluster of fourteen similar buildings, which is composed of the subset of the following buildings {1, 3, 4, 7, 8, 9, 10, 11, 13, 14, 16, 17, 18, 20}. Buildings 17 and 8 are identified as the most similar buildings in the dataset. Looking at their descriptions, they share the same number of occupants, building type, and construction year class. Building 17 also has only one more bedroom compared to building 8. Building pairs {9, 11}, and {16, 20} are also respectively identified as mutually similar.

B. Evaluation metrics

Our goal is to achieve a good generalization performance by accurately predicting short-term energy consumption of unseen buildings. Therefore, we assess our proposed model using the Root Mean Squared Error (RMSE). RMSE is defined as the square root of the average squared distance between prediction and ground truth, using the formula:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}\,,$$

where y_i and $\hat{y}_i$ respectively denote the true value and the predicted value of the i-th data sample, and N denotes the size of the dataset.

C. Model Training

For each building, we use data between April 2014 and May 2015 for training. For cross-building evaluations, we use data between April 22nd, 2014 and June 1st, 2014. The whole dataset was scaled so all values will be between 0 and 1, using min-max normalization algorithm. The input and the output sequences are of length 7. The input corresponds to a 7-dimensional feature vector. Our network is composed of two hidden layers; one LSTM layer of size 256, and one fully-connected layer of size 128. The Rectified Linear Unit (ReLU) is used as the non-linear activation function for hidden layers. The output layer consists of a fully-connected layer with linear activation function. The fine-tuning of weights is done using Gradient Descent algorithm with an exponentially decaying learning rate ranging between 10^{-3} and 10^{-5}. Weights initialization follows a normal distribution with zero mean and standard deviation $\sigma = 1$, whereas biases are initialized to zeroes. The gradients are back-propagated through timestep batches of length 80. For the training epochs number, we have fixed 1000 as the maximum number of epochs. To avoid over-fitting, we have implemented an early stopping mechanism which breaks the training loop when training cost does not improve on the training set after 20 epochs.

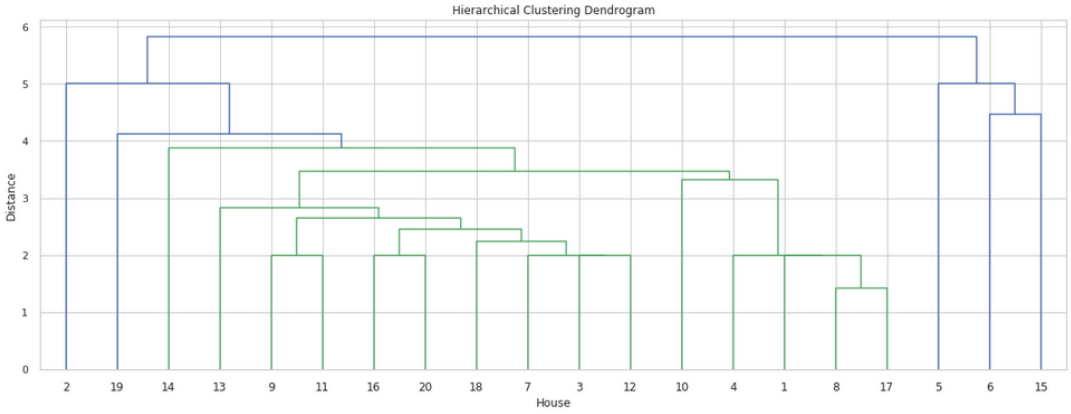

Figure 5. Dendrogram of the hierarchical clustering of REFIT households based on their descriptions. Clusters within which distance is below 70% of the maximal cluster-wise distance Categorical are colored in green. Features in the buildings feature vectors were one-hot encoded. The distance used was the Euclidean distance.

D. Experimental Results

We trained 19 models for each building following the same process. One building (number 12) was not considered due to insufficient training data. Each model was tested on the remaining unseen buildings in order to study its cross-building transfer-ability. Figure 6 depicts the predictions errors of cross-building model transfers as a heatmap. We can visually identify two clusters within each of them generalization performances are high. These clusters are respectively composed of the following subsets of buildings {2, 3, 18, 19} and {5, 6, 7}. We also notice that buildings 13 and 14 are mutually similar and that models trained on buildings 10 and 17 generalize well when applied to them during inference mode. Furthermore, we can visually conclude that all trained models perform poorly when applied to building 15. Model trained on building 15 also has poor generalization performances when applied to the remaining unseen buildings.

We now seek to examine similar buildings based on these results; our assumption is that similar buildings models are transferable among each other. Hence, a model that is trained on a building i will generalize well when applied to a building j if buildings i and j are similar. We start by processing the experimental results matrix (Figure 6) to transform it to a distance matrix. For this purpose, we simply compute pairwise averages between each element at row i and column j and its corresponding element at row j and column i. Drawn clusters from this distance matrix are illustrated in Figure 7 using a dendrogram. We use the Euclidean distance to compute pair-wise similarities. Figure 7 identifies two main clusters, which are respectively composed of the following subsets of buildings {5, 6, 7, 8, 10, 13, 14, 16, 17, 20, 16} and {1, 2, 3, 4, 9, 11, 15, 18, 19}.

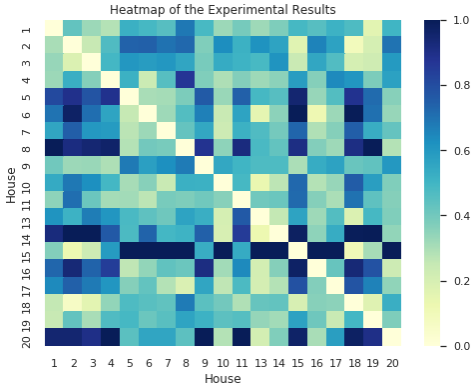

Figure 6. Heatmap of the experimental test errors; we trained 19 models, each of them on one single building. Each model was tested on each building. Th y-axis represents buildings on which each model was trained, the x-axis represents the buildings on which each model was tested. The evaluation metric was RMSE. Final results were scaled between 0 and 1. House number 12 was not considered due to insufficient training data.

V. Discussion

From Figure 5 and Figure 7, we can notice that buildings 8 and 17 which were the most similar based on their descriptions are clustered under the same cluster based on their cross-domain generalization errors. This means that models trained on building 8 will generalize well when applied to building 17 during inference mode, and vice versa. Similarly, the two sets of buildings {9, 11}, and {16, 20} are identified as similar in both clustering schemes; based on descriptions and cross-domain generalization errors. Furthermore, poor cross-domain generalization performances of building 15 (Figure 6) is explainable by its dissimilarity with the rest of buildings (Figure 5).

We may therefore suggest that buildings, that are judged similar based solely on their descriptions, do yield to good prediction results when performing cross-building knowl-

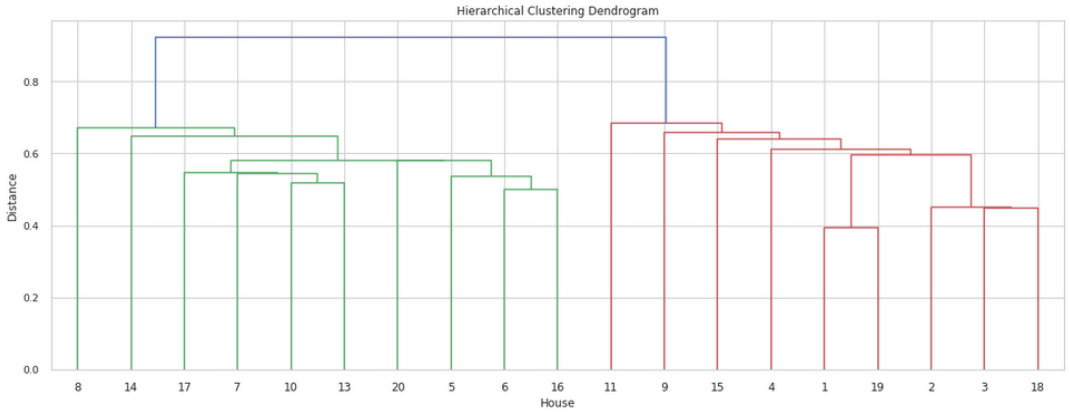

Figure 7. Dendrogram of the hierarchical clustering of REFIT households based on experimental cross-building prediction results. Clusters within which distance is below 70% of the maximal cluster-wise distance Categorical are colored in green and red. The distance used was the Euclidean distance.

edge transfer.

In the context of this study, we have leveraged a very restricted set of building descriptions, i.e. number of occupants, typology, size, etc. Therefore, we believe that richer and broader building description would help to select similar data more accurately and more reliably. Furthermore, and due to the large variety of building typologies and design, and uncertainties surrounding its environment and occupancy patterns, we consider that data selection approaches based on similarity metrics are important in order to perform large-scale and accurate cross-domain domain generalization.

VI. Conclusion and Perspectives

This paper discusses the suitability of the data selection approach for cross-building knowledge transfer. Evaluation work was conducted on the case study of building energy consumption modeling. For this purpose, we have trained per-building models and studied their transfer-ability across other unseen buildings. Experimental results show that minimal building descriptions are capable of guiding domain generalization applications in the context of energy modeling, by identifying similar buildings. Overall, we believe our results confirm the suitability of data selection mechanisms that are based on similarities of building minimal descriptions.

As future work, we will investigate large-scale data selection service approaches for domain generalization. We also intend to extend our system by automating the data selection algorithm based on user queries. User queries will contain the description of the target building to which we want to transfer knowledge.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[2] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data-ai integration perspective," *arXiv preprint arXiv:1811.03402*, 2018.

[3] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. D. Ibáñez-Gonzalez, E. Kacprzak, and P. T. Groth, "Dataset search: a survey," *CoRR*, vol. abs/1901.00735, 2019. [Online]. Available: http://arxiv.org/abs/1901.00735

[4] M. Sugiyama and A. J. Storkey, "Mixture regression for covariate shift," in *Advances in Neural Information Processing Systems*, 2007, pp. 1337–1344.

[5] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.

[6] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[7] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Advances in neural information processing systems*, 2011, pp. 2178–2186.

[8] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*, 2013, pp. 10–18.

[9] K. Amarasinghe, D. L. Marino, and M. Manic, "Deep neural networks for energy load forecasting," in *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*. IEEE, 2017, pp. 1483–1488.

[10] Y. Ding, M. A. Neumann, E. Stamm, M. Beigl, S. Inoue, and X. Pan, "A personalized load forecasting enhanced by activity information," in *2015 IEEE First International Smart Cities Conference (ISC2)*. IEEE, 2015, pp. 1–6.

[11] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2551–2559.

[12] Z. Xu, W. Li, L. Niu, and D. Xu, "Exploiting low-rank structure from latent domains for domain generalization," in *European Conference on Computer Vision*. Springer, 2014, pp. 628–643.

[13] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *European Conference on Computer Vision*. Springer, 2012, pp. 158–171.

[14] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5542–5550.

[15] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[16] Z. C. Lipton, "A critical review of recurrent neural networks for sequence learning," *CoRR*, vol. abs/1506.00019, 2015. [Online]. Available: http://arxiv.org/abs/1506.00019

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[18] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, pp. III–1310–III–1318. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042817.3043083

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of united kingdom households from a two-year longitudinal study," *Scientific data*, vol. 4, p. 160122, 2017.