Large-scale transfer learning for data-driven modelling of hot water systems

Hussain Kazmi[1,2], Johan Suykens[3], Johan Driesen[2]
[1]Enervalis, Belgium
[2]Department of Electrical Engineering (ESAT), KU Leuven, Belgium

Abstract

Hot water systems represent a substantial energy draw for most residential buildings. For design and operational optimization, they are usually either modelled by domain experts or through black-box models which makes use of sensor data. However, given the wide variability in hot water systems, it is impractical for a domain expert to individually model every hot water system. Likewise, black-box systems typically require an enormous amount of data to converge to a usable model. This paper makes use of transfer learning, a novel machine learning tool, to completely automate the learning process while substantially accelerating the performance of comparable black-box systems. Using real world data from 61 houses employing two different types of hot water systems, the proposed system is shown to work on both homogeneous and heterogeneous hot water systems. Convergence to a reliable model with transfer learning is on the order of a few weeks, as opposed to months or years without transfer. By presenting a detailed account of how transfer learning can be used in different contexts, we hope that it will become a widely used tool in the building modelling and simulation community.

Introduction

Hot water systems represent a substantial load in residential energy consumption (Pérez-Lombard et al. (2008)) and will also increasingly impact the electric grid with the electrification of heating systems (Baruah et al. (2014)). More recently, researchers have explored the possibility to use hot water systems as ubiquitous sources of flexibility. This flexibility can be leveraged to either improve operational efficiency (Kazmi et al. (2019)) or provide different services to the electric grid (Liu et al. (2018)). Such active control of hot water systems generally requires a dynamics model describing the behaviour of the hot water system. This model should include a characterization of both the storage element (i.e. the hot water vessel) and the heating element (e.g. an electric or gas boiler, a heat pump etc.), and can be used with a number of optimization schemes such as model predictive control and reinforcement learning based control (Kazmi et al. (2019)).

In addition to active control, a detailed dynamics model of the system can also enable simulation studies to study the effects of different variables on system performance (Fischer et al. (2017)). Other applications include providing recommendations to the users to improve some aspect of device or grid operational efficiency, and diagnosing or predicting faults during the operational phase (Chen and Lan (2009)).

A number of modelling techniques have been proposed in literature that aim to capture the behaviour of hot water systems. These include white-box modelling methods which utilize a human modellers domain expertise to characterize the system dynamics of the hot water system (Hensen and Lamberts (2012)). At the other end of the spectrum, lie black-box modelling techniques which remove the dependence on the human domain expert by learning the systems dynamics directly from sensor data. This can be done both offline (i.e. when a model is learned prior to operation) (Kazmi et al. (2016)) and online (i.e. when a model is learned during operation). Somewhere between these two extremes lie grey-box modelling methods which calibrate an existing model to observed data (Afram and Janabi-Sharifi (2014)).

Most of these methods suffer from a number of significant shortcomings. White-box methods are constrained by the expertise and availability of the human modeller. The sheer amount of hot water systems to be modelled makes it impractical to consider every single device individually. Furthermore, since these methods are typically employed in the design-phase, they seldom reflect operational performance of the modelled systems, often due to unexpected occupant behaviour. Black-box methods, while avoiding the costly dependence on human domain expertise, rely on extensive sensing of the system to model the system accurately. Where the data being gathered fails to adequately capture the internal state of the system, these methods break down. This is often the case for hot water systems where only minimal sensing is employed in the form of a solitary temperature sensor. As the temperature distribution inside

the storage vessel is not uniform because of stratification and other nonlinear dynamics, this sensory information is often insufficient to learn an accurate dynamics model. Additionally, since they rely on gathered data, black-box methods usually require large amounts of training data to converge to a reliable model of system dynamics (Kazmi et al. (2019)).

This paper presents a method which resolves these issues by leveraging transfer learning, a relatively recent development in machine learning (Pan et al. (2010), Mehrkanoon et al. (2018)). At its heart, the methodology provides a structured way of integrating information collected in a variety of settings to extract useful knowledge. Being data-driven, it is not limited to homogeneous devices, and can also accelerate learning in the context of heterogeneous devices (i.e. devices with different thermophysical characteristics). This paper presents the results of applying transfer learning to hot water systems in two different housing projects comprising of recently renovated net-zero energy buildings in The Netherlands. By successfully learning a reliable system dynamics model in an extremely limited time frame (on the order of days to weeks for both the storage and the heating element) the paper successfully demonstrates few-shot learning. Learning an accurate dynamics model quickly enables all the benefits of traditional black-box systems in a much more practicable manner. It is important to note that the methodology described here is not limited to hot water systems, and is generalizable to other types of energy systems.

Experimental setup

We consider two different housing projects in the Netherlands in this case study. All the houses considered (in both projects) are net-zero energy buildings and are insulated to a very high degree. Furthermore, all the houses considered in both projects employ air-source heat pumps which are used to provide both hot water and space heating. The storage vessel installed in each house in both projects is likewise 200 litres. However, the hot water system is identical only for houses belonging to the same project. There are considerable differences in the make of the hot water system across the two projects (for instance, the vessel orientation and dynamics of the storage vessel, as well as the way the heat pump interacts with it differ considerably). In subsequent sections, we make this distinction clear by referring to households (and devices) belonging to the same project as homogeneous, and those belonging to different projects as heterogeneous. This setting is summarized in Fig. 1. As the paper focuses on data-driven modelling of the hot water system, it is important to enumerate the data streams it uses. These include:

1. Temperature measurement in the storage vessel: for project A, this was at the halfway point in the storage; for project B, it was at one third of the storage vessel height
2. Hot water flow in litres from the storage vessel
3. Ambient temperature
4. Electricity consumed by the heat pump for hot water production

Figure 1: Households considered in the different projects

Making use of this sensor data, the objective is to learn an accurate system dynamics model for the hot water system, which further comprises of a storage model and a heating model. The purpose of the storage model is to estimate the state of the vessel (i.e. its state of charge) at any given instant. On the other hand, the purpose of the heating model is to estimate the amount of energy required by the heating element (heat pump in this case) to reheat the storage vessel from an initial to a final state of charge. Finally, it is important to note here that while data from 53 houses was available for analysis in the first project, there were only eight houses in the second project.

Methodology

This section presents a typical black-box learning work flow and, using it as a benchmark, motivates the need for a transfer learning framework to improve the modelling process. It then presents two different methods to use transfer to allow accelerated learning in black-box settings. The modelling technique used in all of these cases is a deep neural network implemented using Keras (Chollet et al. (2015)), and its architecture is determined through an extensive grid search over hyperparameters which includes the number of layers, number of neurons in each layer, choice of activation function, regularization and learning rate (Goodfellow et al. (2016)).

Numerous metrics have been used in literature to evaluate the performance of black-box systems. In this paper, we focus on two such measures: the R^2 metric (or the explained variance in observation data by the fitted model) and the mean absolute error, or MAE (which quantifies prediction error in absolute terms in the measurement units). Additionally, specific thermodynamic tests were designed as general purpose checks to ensure the generalization potential of model predictions to test a variety of different situations which might arise in real-world situations. These include tests for the following three thermodynamic principles of heat pump operation, keeping all other factors constant:

1. As ambient temperature (T_{out}) increases, energy consumption of the heat pump decreases ($\hat{E}$)
2. As temperature difference between the start and end of the reheat cycle (ΔT) increases, energy consumption of the heat pump ($\hat{E}$) increases
3. As the target temperature (T_{end}) increases, energy consumption ($\hat{E}$) increases

Benchmark black-box method

Typical black-box models learn system behaviour directly from time series data. Historically, this has been in the form of using raw time series to predict future system states. In this case, the only question to consider is which sensor streams to include, and their temporal extents (i.e. how much historic data should be included) as input features. On the one hand, increasing the temporal window allows the neural network to detect longer term trends (i.e. low frequency events). On the other hand, increasing the temporal window length can overwhelm the neural network by providing it with unnecessary inputs. This latter is especially a concern in low data availability settings, where the dimensionality of the training vector can far surpass the amount of training samples collected. With powerful modelling techniques such as deep learning, this opens the door to overfitting, a commonly observed phenomenon in which the model simply memorizes training data, rather than generalizing to unseen test data. This also links with the curse of dimensionality where increasing the input feature vector considerably increases the exploration required by the neural network to learn an accurate representation of the hot water system Verleysen and François (2005). In the case considered in this paper, the length of the window was chosen by evaluating model performance for different window lengths. The best performance was observed with using an entire historic day for all sensors under observation (although the model improvements were marginal, when compared with other comparable window lengths).

A taxonomy of transfer learning

While black-box learning in the manner presented above is quite common in practice, it means learning different models for each household under consideration - an extremely data-inefficient practice. Transfer learning offers three key benefits when compared to traditional data-driven (i.e. black box) methods. These include a higher initial performance, a higher asymptotic performance and a faster rate of learning (Torrey and Shavlik (2010)). This is highlighted in Fig. 2. To achieve this, transfer learning leverages two key concepts which may be shared: a domain and a task (Pan et al. (2010)).

The **domain** $\mathcal{D}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ over the feature space, where $X = \{x_1, x_2, ..., x_n\} \in \mathcal{X}$. Here $\mathcal{X}$ includes the space of all possible feature vectors, whereas x_i is a particular feature vector corresponding to some input, and X is a particular learning sample. Thus, in the context of learning a representation for a hot water systems, an example of the input feature space $\mathcal{X}$ can be all possible combinations of the sensor data (or features extracted from this sensor data). The marginal distribution $P(X)$ over this feature space quantifies the probability of observing a specific feature vector, and depends also on the occupant behaviour and ambient conditions.

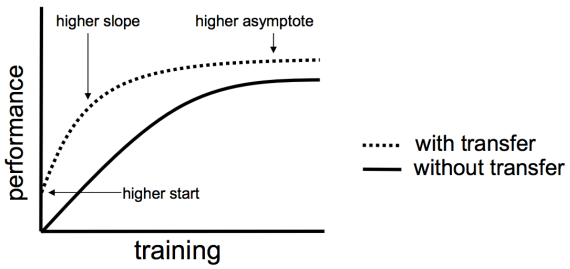

Figure 2: A stylistic representation of modelling performance with and without transfer learning, with increasing amounts of training data (Torrey and Shavlik (2010))

Given a domain, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a **task** $\mathcal{T}$ consists of a label space $\mathcal{Y}$ and a conditional probability distribution $P(Y|X)$, which is typically to be learned from the training data in the form of pairs $x_i \in X$ and $y_i \in \mathcal{Y}$. The task $\mathcal{T}$ is then given by $\{\mathcal{Y}, P(Y|X)\}$. In the hot water system context, $\mathcal{Y}$ is the set of all possible labels which are the state of charge for the storage element and the energy consumption of the heat pump. The conditional distribution $P(Y|X)$ is the dynamics model that we are interested in learning from historic behaviour, which is again influenced by both user and environment.

Permutations of transfer learning

There are four possibilities in transfer learning settings, given the domain and task definitions presented above. We list them briefly in this section.

1. When the **feature space** is different between the source and target domain, i.e. $\mathcal{X}_s \neq \mathcal{X}_t$. This can happen when the instrumentation on the source and target device are completely dissimilar. This case is not considered further in this paper.
2. When the **marginal probability distribution** differs between the source and target domain, i.e. $P(X_s) \neq P(X_t)$. This takes place when identical (or homogeneous) hot water systems are operated in different households, causing the different devices (which share the same system dynamics) to operate in different regions of the state-space.
3. When the **label space** differs across the source and target domain, i.e. $\mathcal{Y}_s \neq \mathcal{Y}_t$. As we are interested in uniform label spaces (i.e. the state of charge for the vessel and an estimation of energy

consumption for the heat pump), this case is also not considered further in the paper.

4. When the **conditional probability distribution** varies between the source and target task, i.e. $P(Y_s|X_s) \neq P(Y_t|X_t)$. This implies different device dynamics and is the case where heterogeneous devices are considered for the transfer task.

We refer to case 2 specifically as **transductive transfer** and case 4 as **inductive transfer**, following the terminology introduced in (Pan et al. (2010)). Transductive transfer learning refers to the case where the source and target tasks are the same, while the source and target domains are different. Inductive transfer learning, on the other hand, is the case where the target task differs from the source task. These two conditions are not mutually exclusive, and it is possible for transfer learning to take place using samples drawn from instances where both domain and task differ for the source and target, a case we refer to as **joint transductive-inductive transfer**.

Ways of achieving transfer

While much research on transfer learning has focused on computer vision and natural language processing problems, the same ideas hold for modelling energy systems. In general, two methods of achieving transfer with neural networks have been investigated:

1. **Feature sharing** is the form of transfer learning where source training data is directly used while learning the target model to improve learning performance. Both raw observations and extracted features can be used for this purpose.
2. **Parameter sharing** usually involves the training of a model (a neural network) with a large amount of source data. The weights (parameters) of this neural network are then used as initialization for the target; these weights are then fine-tuned using observed target data using backpropagation (the target data set is typically orders of magnitude smaller than the source data set). The fine-tuning is usually done with a much smaller learning rate, and it is also possible to completely freeze certain parts of the neural network to retain the representations already learned by the network (Yosinski et al. (2014)).

Sharing raw features is not guaranteed to work in heterogeneous settings, and can sometimes even lead to negative transfer. On the other hand, parameter sharing can lead to overfitting if the fine-tuning is not carried out properly. It is important to note here that both the source and target can draw data being collected by multiple agents, i.e. transfer can take place both synchronously and asynchronously depending on the nature of learning agents.

Towards few-shot learning

While transfer learning can improve performance of black-box methods in general, the way the benchmark black-box method is posed above is quite naive. The most obvious flaw in the formulation is to neglect the fact that the task is episodic. An episodic task refers to a problem which has a clearly defined initial and terminal state. Upon termination, the system state is reset and previous states do not affect future states. In other words, by defining a static temporal window, the black box method formulated above is forced to also consider data from previous episodes, which detracts from the learning process.

The realization of the episodic nature of the task allows for meaningful features to be extracted from the time series. More specifically, five features are extracted from the raw time series data: (1) the mid-point temperature in the storage vessel after a reheat cycle (this is a proxy for the initial state), (2) time elapsed since the last reheat cycle (episode duration), (3) hot water consumption since the last reheat cycle (human interaction during the episode)), (4) ambient temperature conditions, and (5) the mid-point temperature just before the reheat cycle (this is a proxy for the terminal state of the vessel). The last two features only influence the heat pump model, as the storage vessel is contained in a conditioned space, and thermodynamic losses remain relatively unaffected by ambient conditions. Extracting these features leads to a feature set whose dimensionality is roughly two orders of magnitude lower than the one used for raw time series learning, thereby circumventing the curse of dimensionality. Feature extraction in this manner also improves the interpretability of the learned model, another common problem in black-box methods.

It is important to keep in mind what the neural networks are actually learning. The storage model learns the temperature distribution in the vessel as a function of thermodynamic and mixing losses, given some initial conditions. This temperature distribution is then thresholded to obtain a state of charge (i.e. the amount of hot water above a certain temperature threshold reflects the state of charge (SoC)). The heating model, on the other hand, learns the amount of energy which would be required to reheat the storage vessel in a given state of charge and ambient conditions.

Results

In this section, we present results from applying the formulation presented above to the two different hot water systems. First, we discuss the application of the algorithm to the storage model, which is, in a way, an easier learning problem because of an abundance of data. The heating model is more difficult to learn accurately because the training examples available for

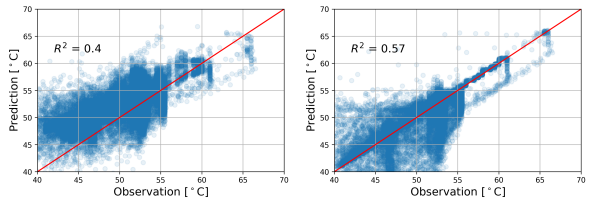

Figure 3: Storage vessel model accuracy with raw time series learning for increasing amounts of data (1 week, 32 weeks)

this are typically two orders of magnitude fewer than for the storage model. This is because, in a day, there are only a few (usually not more than two) reheat cycles, but the temperature data is collected every 5 or 15 minutes.

Storage model

Benchmark black-box: Fig. 3 presents the result of predicting the mid-point temperature in the storage vessel with a deep neural network with three hidden layers (chosen through hyperparameter search) trained on increasing amounts of gathered data in a household (1 week and 32 weeks). While the performance improves over time as more data becomes available to the neural network, the predictive accuracy continues to be quite low, as evidenced by the poor correlation between predicted and observed temperatures (and the correspondingly low R^2 values). One explanation for this poor performance was the high dimensionality of the input feature data when compared with the number of training examples.

Benchmark black-box with transfer learning: The realization that all individual households are trying to learn the same dynamics model (especially within the same project) can be leveraged to apply transfer learning to accelerate the modelling process. In this case, the gathered features from individual households are combined together to form a single feature vector which is then used to learn the shared dynamics model for all households. As seen in Fig. 4, increasing the data weeks used for learning a model improves its accuracy (or the variance it can explain in the observed data) but only up to a certain extent before asymptoting. In this way, only one of the three benefits of transfer learning, as shown in Fig. 2, i.e. improved initial performance, is realized. The asymptotic performance remains largely unaffected.

Learning with extracted features: By reducing the dimensionality of the input feature vector from 96 or 288 (depending on sampling rate) to 3 (i.e. applying the feature transformations as explained in the previous section), the learning problem is simplified considerably. This is reflected in the improved accuracy of the learned storage model using extracted features, as shown in Fig. 4. This feature transformation also considerably simplifies the calculation of state of charge from the predicted temperature.

Demonstrating transfer: It is also instructive to

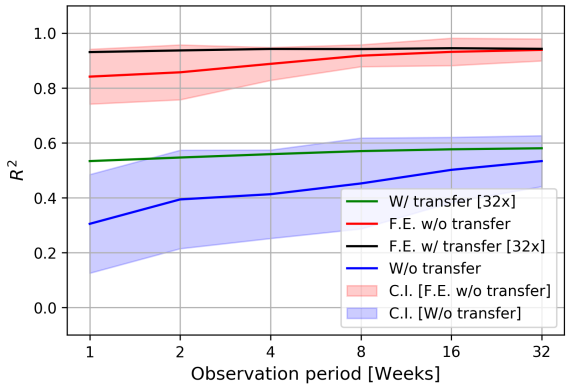

Figure 4: Storage vessel model accuracy with raw time series learning incorporating transfer learning data-weeks here represents amount of data in weeks used to train the neural network, the source of the data can be from different households achieving transfer

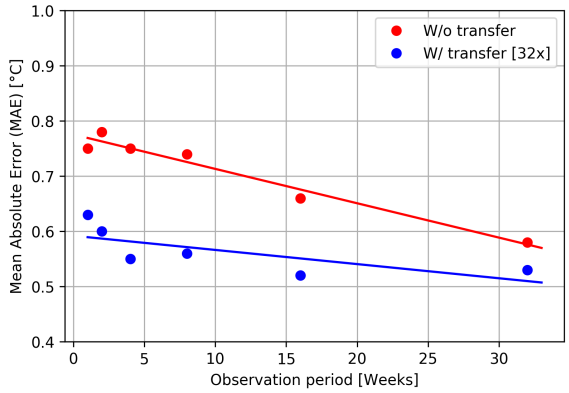

Figure 5: Mean Absolute Error [°C] as a function of increasing data collection (weeks) and agency (households)

summarize the effect of increasing agency and time on the learning model accuracy. This is highlighted in Fig. 5 where it is easy to see that increasing agency and data collection have largely the same effect, i.e. the initial performance of the system with transfer learning is close to the asymptotic performance of the learner without transfer. This means that gathering data for months in a single household can be replaced by collecting data in multiple households for a very brief amount of time. Of course this result holds only for homogeneous devices, but it can also be extended to heterogeneous devices, as we show in the next section. It is also fairly easy to see that while transfer learning allows for a much improved initial performance, the asymptotic performance is not too different for both with and without transfer learning.

Heating model

Benchmark block-box: As mentioned previously, the biggest challenge to model the heating element accurately arises from the very limited training dataset the learning algorithm has access to. Practically, this means that the learning algorithm has ten or fewer training examples after a week of interacting with the system for a single household. For data-intensive al-

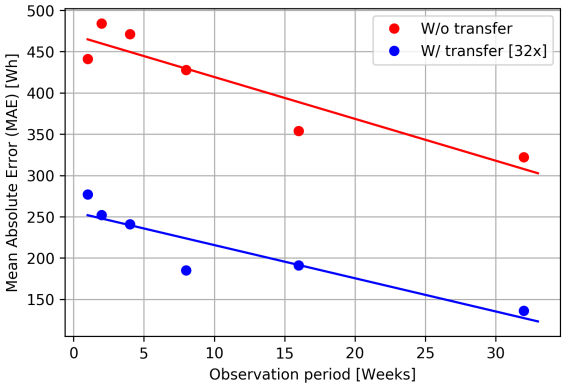

Figure 6: Mean Absolute Error [kWh] as a function of increasing data collection (weeks) and agency (households)

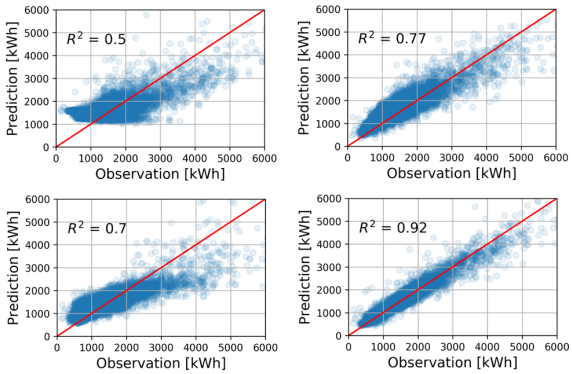

Figure 7: Scatter plot between observed and predicted electricity consumption for the heat pump as a function of increased data and agency: (top-left): 1 week of data for 1 agent; (top-right) 32 weeks of data for 1 agent; (bottom-left): 1 week of data for 32 agents; (bottom-right): 32 weeks of data for 32 agents

gorithms like deep neural networks, this leads to severe overfitting, especially when the deep neural network is using the raw time series as its input feature vector. In this case, the dimensionality of the input feature vector is multiple orders of magnitude higher than the number of examples available for learning. This seldom, if ever, works well in practice. Indeed, in this case the neural network failed to converge using raw time series data alone, with or without transfer learning.

Learning with extracted features: As before, to model the heating element, the extracted input feature vector is fed to the neural network which predicts the energy required to reheat the storage vessel given different ambient conditions. On average, this energy is between one and two kWhs (however it can vary considerably as a function of the vessels state of charge and ambient conditions). Unlike the raw learning case, the neural network successfully learns to predict the heating elements behavior given extracted features. This prediction grows progressively better as the agent observes more data, however the learning rate is much higher than for the case of the storage vessel.

Demonstrating transfer: The model improvement effect holds also as the number of agents (i.e. households involved in the learning process) increases. However, unlike the case of the storage element, the heating model continues to improve until all the gathered data has been used. In this case, transfer learning leads to both improved initial and asymptotic performance (as highlighted earlier in Fig. 2). It is important to note that without transfer, a single household would never have access to almost 20 years of operational data (which is the asymptotic amount of data used in the transfer learning case). This information is highlighted in Fig. 6 where it is easy to see that the error rate continues to drop as we increase the amount of data (either through observation period or the number of households).

An interesting caveat arises here as, unlike for the storage model, the model improves more significantly for a longer data gathering period with fewer households than it does with additional households with fewer data gathering (i.e. learning a model with data collected for one household over 32 weeks results in a better model than one learned with data collected over 32 households for one week). This makes intuitive sense and is because of better exploration of ambient conditions over 32 weeks (i.e. the model observes heat pump performance under different conditions) than is possible in only one week, even when multiple households are observed. This effect is highlighted in Fig. 7. This means that regardless of the amount of households involved in the initial transfer, learning will always continue to improve for a while as it takes stock of the effect of ambient conditions on heat pump performance. This is unlike the case of the storage vessel.

Induction

The heating model was eventually able to learn an extremely accurate representation of the heat pump (with a normally distributed relative mean error of less than 10%). However, it took almost 20 years of data to do so, implying that a more data-efficient representation can further improve real world learning performance. In the case of the storage model, this was not necessary as an accurate representation was learned in a week of data collection for the case of transductive transfer learning. This section considers inductive transfer learning to further accelerate heating model improvements, which can be achieved by making use of the data gathered in heterogeneous devices (i.e. from devices belonging to different projects in this case). In practice, inductive transfer learning can be achieved in one of two ways (parameter sharing or feature sharing), as explained earlier.

In this paper, the performance of both types of induction is compared. Project A is considered the source (because of greater data availability), while Project B

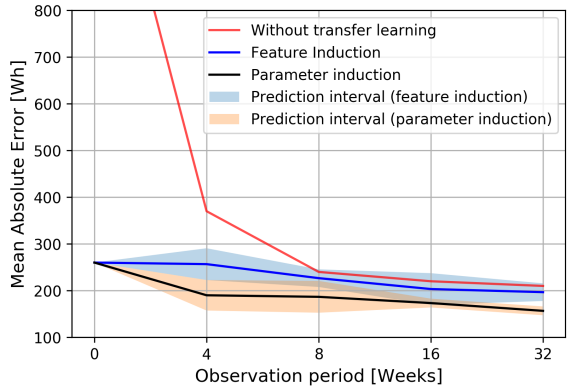

Figure 8: Mean Absolute Error [°C] as a function of increasing data collection (weeks) and the learning scheme (i.e. with or without transfer)

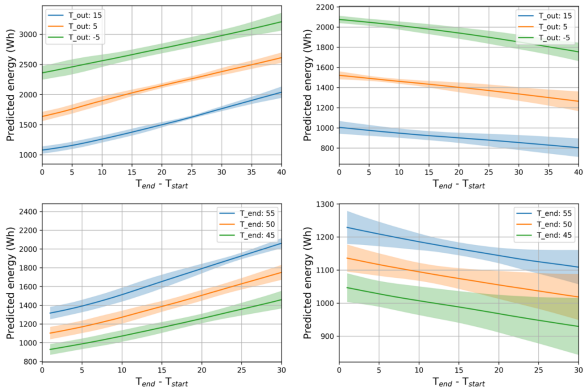

Figure 9: Results of learning with and without induction learning for the heating model; results shown here are to visualize the trends of the learned model for (left) with induction for 2 agents after 4 weeks, and (right) without induction with 2 agents after 32 weeks

is treated as the target which can make use of transfer learning to learn a reliable model quicker. From Fig. 8, it is obvious that parameter induction (i.e. initializing the neural network with previously trained data) outperforms naive feature induction. It is also important to note that both parameter and feature sharing perform substantially better than the model learned using just the target data (i.e. project B). This effect is especially pronounced in the early stages of data collection.

In this case, the workflow for feature sharing is as follows: the training data gathered from project A is aggregated with training data from project B, all of which is then used to train a single neural network. The workflow for parameter sharing is more involved as first a neural network is trained on the already available data from project A. Then the weights of this neural network are used as the initialization for project B where observed data is used to fine-tune the weights through backpropagation. Results of both these methods are compared with a neural network which is initialized randomly but then is trained using only the target data (i.e. for project B). It is obvious that pre-training the neural network drastically speeds up real world performance and reduces data requirements by over an order of magnitude making it realistic to model the heating element through sensor data alone.

Thermodynamic validation: While the prediction error with inductive transfer is much lower than the benchmark, it is not obvious whether the neural networks learned using data alone can generalize to beyond the training and test set. This is especially a concern because both training and test data are sampled from real world behaviour of hot water systems, which a controller is meant to affect. This controller has the potential to drive the system to different, unseen parts of the state-space. As evident from Fig. 9, the model learned without induction has been able to learn only two of the three fundamental properties tested correctly after 32 weeks of data collection, even when applying transductive transfer learning over two households. The model is unable to generalize well on arguably the most important property, i.e. a higher temperature difference between start and end temperature in the storage leads to higher energy consumption. On the other hand, agents making use of induction were able to learn (retain) all three properties correctly from the source task within four weeks while simultaneously far outperforming the case without transfer on the MAE and R^2 metrics. This case is a successful example of applying both transductive and inductive transfer learning to help accelerate model learning.

Discussion and Conclusion

This paper has presented results from using transfer learning to accelerate the real world performance of black-box systems. This is an important real world challenge because residential energy systems, while increasingly important from a demand side management perspective, are prohibitively expensive to be modelled by a human domain expert because of their wide variability. Likewise, existing black-box systems suffer from many shortcomings, and can take a long time (during which observational data has to be gathered) before converging to a reasonable model. This limits their real-world applicability.

Several important conclusions can be drawn from this work to improve on state-of-the-art in black-box modelling. Primarily, the paper demonstrates that transfer learning can improve the modelling accuracy of black-box systems in both homogeneous and heterogeneous device contexts. It shows that, depending on the quantity of observational data and homogeniety of devices, transductive or inductive transfer learning might yield the greatest performance gains. Furthermore, when applying inductive transfer, parameter sharing outperforms feature sharing for heterogeneous devices, while for homoegeneous devices feature sharing is arguably a better idea. Likewise, the

amount of data gathered also influences the gains possible with transfer learning: to illustrate this point, the paper shows how the storage vessel model behaves very differently from the heat pump model. The fundamental difference between the two models is a reliance on ambient conditions which means that longer observational periods help improve modelling performance for the heat pump. This also means that a model learned in a certain geographical location may not be directly usable in a different location, but might serve as a source model which could be fine-tuned for improved performance. The paper also demonstrates the importance of using multiple metrics for evaluating modelling performance, rather than relying on a single indicator, as this can yield misleading results.

It is important to note here that the initial predictions of the neural network before substantial amounts of data have been gathered can be completely incorrect. While transfer learning can address this to an extent, it is also possible to incorporate domain-specific knowledge into the learning process. However, as this detracts from the task-agnostic learning approach espoused in this paper, this was not considered in this paper. Regardless of the transfer mechanism employed, the paper also highlights the importance of extracting meaningful features to improve modelling performance, and shows that a naive black-box formulation is insufficient for hot water system modelling. Another challenge with transfer learning is the risk of negative transfer, which is an area of active research. It is therefore important to stress here that transfer learning, by itself, might not be the silver bullet to solve all of black-box modelling challenges.

While the focus of this paper has been on modelling hot water systems, the framework is generalizable to other energy systems. While the heterogeneous systems considered in this research belonged to the same family of devices (i.e. both were heat pump hot water systems), it is a possible future research direction to evaluate the framework for more diverse systems (such as heat pumps and resistance heaters). Given the potential gains and the limited cost of realizing them, we believe transfer learning should be a fundamental part of every modeller's repertoire. This paper provides a useful starting point in this direction.

References

Afram, A. and F. Janabi-Sharifi (2014). Review of modeling methods for hvac systems. *Applied Thermal Engineering 67*(1-2), 507–519.

Baruah, P. J., N. Eyre, M. Qadrdan, M. Chaudry, S. Blainey, J. W. Hall, N. Jenkins, and M. Tran (2014). Energy system impacts from heat and transport electrification. *Proceedings of the Institution of Civil Engineers-Energy 167*(3), 139–151.

Chen, Y. and L. Lan (2009). A fault detection technique for air-source heat pump water chiller/heaters. *Energy and Buildings 41*(8), 881–887.

Chollet, F. et al. (2015). Keras.

Fischer, D., T. Wolf, J. Wapler, R. Hollinger, and H. Madani (2017). Model-based flexibility assessment of a residential heat pump pool. *Energy 118*, 853–864.

Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio (2016). *Deep learning*, Volume 1. MIT press Cambridge.

Hensen, J. L. and R. Lamberts (2012). *Building performance simulation for design and operation*. Routledge.

Kazmi, H., S. D'Oca, C. Delmastro, S. Lodeweyckx, and S. P. Corgnati (2016). Generalizable occupant-driven optimization model for domestic hot water production in nzeb. *Applied Energy 175*, 1–15.

Kazmi, H., J. Suykens, A. Balint, and J. Driesen (2019). Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads. *Applied Energy 238*, 1022–1035.

Liu, M., S. Peeters, D. S. Callaway, and B. J. Claessens (2018). Trajectory tracking with an aggregation of domestic hot water heaters: Combining model-based and model-free control in a commercial deployment. *arXiv preprint arXiv:1805.04228*.

Mehrkanoon, S., M. B. Blaschko, and J. A. Suykens (2018). Shallow and deep models for domain adaptation problems. *Proceedings ESANN 2018*, 291–299.

Pan, S. J., Q. Yang, et al. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*(10), 1345–1359.

Pérez-Lombard, L., J. Ortiz, and C. Pout (2008). A review on buildings energy consumption information. *Energy and buildings 40*(3), 394–398.

Torrey, L. and J. Shavlik (2010). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 242–264. IGI Global.

Verleysen, M. and D. François (2005). The curse of dimensionality in data mining and time series prediction. In *International Work-Conference on Artificial Neural Networks*, pp. 758–770. Springer.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328.