sensors

MDPI

Article

Variational Regression for Multi-Target Energy Disaggregation

Nikolaos Virtsionis Gkalinikis*, Christoforos Nalmpantis and Dimitris Vrakas

School of Informatics, Aristotle University of Thessaloniki, 54124 Thesssaloniki, Greece; nvirtsion@csd.auth.gr (N.V.G); christofn@csd.auth.gr (C.N.); dvrakas@csd.auth.gr (D.V.)
* Correspondence: nvirtsion@csd.auth.gr (N.V.G.)

Abstract: Non-intrusive load monitoring systems that are based on deep learning methods produce high accuracy end-use detection, but are mainly designed with the one vs one strategy. This strategy dictates that one model is trained to disaggregate only one appliance, which is sub-optimal in production. Due to the high number of parameters and the different models, training and inference can be very costly. A promising solution to this problem is the design of a NILM system where all the target appliances can be recognized by only one model. This paper suggests a novel multi-appliance power disaggregation model. The proposed architecture is a multi-target regression neural network which consists of two main parts. The first part is a variational encoder with convolutional layers and the second part has multiple regression heads, which share the encoder's parameters. Given the total consumption of an installation, the multi-regressor outputs the individual consumption of all the target appliances simultaneously. The experimental setup includes a comparative analysis against other multi and single-target state-of-the-art models.

Keywords: non-intrusive load monitoring; energy disaggregation; nilm; deep learning; variational inference; multi-target regression; kl divergence; convolution neural networks

Citation: Virtsionis Gkalinikis, N.; Nalmpantis, C.; Vrakas, D. Variational Regression for Multi-Target Energy Disaggregation. *Sensors* **2023**, *1*, 0. https://doi.org/

Academic Editors

Received: 17 December 2022
Accepted:
Published:

1. Introduction

Disaggregation is the process of breaking down a quantity into its separate elements. Specifically, the term energy disaggregation is a synonym of non-intrusive load monitoring (NILM) [1], a set of methods that aim to estimate the consumptions of the electrical appliances that compose the aggregate consumption of an installation. NILM can be thought as a blind source separation task [2], where only the mains consumption signal is provided as input and could be an essential tool for both the individual consumers and the distribution system operators (DSOs). From the consumers' side, NILM constitutes a vital part of intelligent home systems providing insights in order to reduce energy waste, raise energy awareness [3,4], improve operational efficiency of installations [5–7] or create smart alerting mechanisms for the residents in need [8–10]. On the other hand, the DSOs can use NILM as a building block for various applications regarding the management and efficient monitoring of the grid [11,12] in combination with more accurate energy consumption forecasts [13,14]. In a similar fashion, disaggregation can be performed in other quantities that are also used in residential buildings as natural gas [15] and potable water [16–18] in order to preserve the resources and reduce the overall living costs of the habitats.

Early attempts to confront the task of energy disaggregation used combinatorial methods to estimate the on/off events of each appliance [1] and Factorial Hidden Markov Models [19–21] to derive the appliance consumption. In FHMMS a model consists of a multiple independent HMMs and the output is essentially a combination of all the hidden states. During the last decade though deep learning solutions seem to have conquered the energy disaggregation research field, producing state-of-the-art solutions [22]. Kelly and Knottenbelt [23] were the first to apply deep learning in order to tackle the problem of NILM, by introducing three novel architectures. Since then researchers in the field have devised a variety of solutions using different types of networks as convolutional [24,25],

recurrent [26–29] or combination of the two [30–32]. Some of them claim to have achieved state-of-the-art performance [24,27,33,34]. A popular technique from Natural Language Processing that produces good disaggregation models is the concept of attention [32,35–37]. The main idea is that the model can detect the most important parts of a sequence and learns to take them into consideration. Based on the results in the literature this type of models shows great generalization capabilities. Moreover, the concept of data generation have been successfully applied to the problem of NILM either as detection systems [37–41] or as data generators [42,43].

Mostly, the research concentrates on designing one network per appliance. Since a usual household contains a great number of devices connected to the electric sockets, successful disaggregation in this framework dictates the use of the same number of models in parallel. This approach may be easier to implement, but has its downsides. For starters, regarding practical systems, it is not a viable solution in terms of cost. Disaggregation applications are usually deployed on cloud services, where the charge depends on the training duration and the space requirements. Furthermore, a method of combining the results of all the models and provide a final answer should be designed. This is more demanding than it seems, since many problems and difficult decisions must be made. For example, two or more models could detect the same end-use (or part of it) as their target appliance. The choice of the correct output should considerate a number of parameters regarding the total consumption at the time of the specific end-use, the general accuracy of the models and the uncertainty of their answers etc. However, in order to make the one vs one strategy easier to be applied in practical solution Kukunuri *et al.* [44] proposed a set of compression approaches suitable for deployment on hardware with limited computation capacity.

In an effort to provide more solid and deployable solutions towards practical NILM applications multi-appliance approaches can be utilised. In this framework one model accurately detects the electrical signatures of multiple targets. This results to the estimation of the corresponding individual energy usages simultaneously. Ideally, successful training should provide a model that automatically takes into account the energy allocation of all the targets and provide the right answer without any extra work. Towards this direction Basu *et al.* [45,46] were the first to apply popular multi-label classification algorithms in the problem of NILM in order to detect on/off events of multiple targets. In recent studies regarding multi appliance detection, the general approach was composed of two steps; to first detect the on/off events of the devices and then estimate their energy usage [47–49]. During the experiments in the current paper it was found that providing the on/off states of the appliances as ground truth to the model boosts performance. Hence, a novel architecture was designed with the capacity to output the power consumption time series of many appliances directly, while keeping the computational costs low.

Apart from the deployment and practicalities issues though, the design of a NILM-oriented application should consider a set of parameters that are hidden from the naked eye. To begin with, a disaggregation algorithm heavily depends on the datasets that were used for the design and the evaluation. Even though the generalization capability of the methods is indeed a desirable property, the designer should take into account that NILM algorithms aim to detect appliance end-uses, which are closely affected from the habits and the culture of the residents. Of course this may apply on specific appliances and not all of them. Hence, the regionality of the data is an important factor to be taken account. Moreover, the sampling period of the data has a great effect on the detection limitations of the algorithms. Usually, disaggregation research revolves around sampling periods 1-10 seconds known as low sampling frequency in NILM. In lower granularities, e.g. 15 minutes, the unique features in electrical signatures are vanished. As a result, accurate appliance event detection is impossible with that data. Finally, the reproducibility of NILM experiments is not an easy task, since there is not a common bench-marking process among researchers. Thus, choosing a suitable NILM algorithm for an application is not always straightforward.

Towards the direction of reproducible and comparable results Symeonidis *et al.* [50] designed a framework composed of various stress test scenarios of evaluating energy disaggregation methods, whereas Batra *et al.* [51] implemented an easy-to-use API for rapid algorithm comparisons along with a set of baseline models. In an effort to standardize the way of how NILM experiments are conducted by Virtsionis Gkalinikis *et al.* [52] created Torch-NILM, the first Pytorch-based deep learning library for energy disaggregation. Torch-NILM contains tools to process time series data, build neural networks and three APIs to design experiments that follow an integrated benchmark method. Even though the aforementioned works are in the direction of standardization of experiments for tackling the comparability issue, the NILM research community is still lacking a globally accepted comparison framework [53].

2. Contributions

The current research concentrates on designing a deep learning architecture capable of detecting the desired set of electrical household appliances simultaneously. Also, the network should be computationally efficient in order to be used in both commercial and research applications, with high training and inference speeds. Thus, a cutting edge model with low storage and computation requirements is developed, named Variational Multi-Target Regressor (V.M.Regressor). The proposed architecture outperforms other multi-target disaggregation models and competes with known state-of-the-art models that use the one vs one (or single-target) strategy. An ablation experiment shows how the key ingredients of the network increase its disaggregation capability comparing to a simpler implementation, whereas a comparison of three variants of the model indicates the best one. What is more, in order to identify any changes in performance of the models, the final experiment was designed using a different number of target appliances.

This article contributes to the energy disaggregation research in the following points. Firstly, with a novel neural network that is able to achieve multi-target disaggregation. The network is built upon a combination of usual artificial layers as CNN and fully connected [54] with the concept of variational inference in a similar way as used in [34,55]. The novel network is compared with a variation of UNet-NILM multi-target model [56] and a baseline. Additional experiments of known single-target models on the same data are included in order to measure the performance difference of the two strategies. Finally, with an ablation study in order to highlight the benefits of variational inference in the current task. The ablation is essentially a comparison between the proposed neural network and a vanilla version without variational inference.

3. Materials and Methods

3.1. Datasets

The training for all the experiments of this work is executed on data from the UK-DALE [57] public dataset, which contains data from five residential buildings in the UK. On the other hand, for the evaluation of the models data from UKDALE and REFIT [58] dataset were utilized. These datasets are very popular among NILM researchers due to the fact that they contain high quality measurements with limited missing values. For most of the experiments five household devices are chosen: dish washer (DW), fridge (FZ), kettle (KT), microwave (MW) and washing machine (WM). There are two main reasons for this. For starters, these appliances are widely used among residents. Hence, accurate disaggregation of those consumptions could interest the users and the DSOs. Secondly, these appliances have different operation characteristics resulting to quite different electrical signatures. Thus, the multi-target models have to extract the most useful and complicate characteristics in order to separate the individual sources and achieve high performance.

3.2. Data Preprocessing

In order for the neural networks to extract complex features and patterns, a minimum preprocessing should be applied to the raw data. The preprocessing of the data in this study is comprised of the following steps:

- Mains and target time series are aligned in time.
- The empty or missing values are replaced with zeros.
- The time series are normalised using standardization, with the values transformed in a way that are centred around the zero mean with a unit standard deviation:

$$Z = \frac{x - mean}{std} \tag{1}$$

  where Z is the standard score, x the data point, $mean$ and std the average and the standard deviation of the time series.
- The data is transformed in order to follow the sliding window method [27].
- The on/off states of the target appliances are calculated in each window. An appliance is considered that is working when its power level at the time of interest is above a predefined threshold. The on power thresholds in this manuscript were drawn from the work of Kelly and Knottenbelt [23].

3.3. Methodology

The experiments of this work are summarized in Table 1. All the experiments were designed and performed using the Torch-NILM framework created by Virtsionis Gkalinikis *et al.* [52]. The same data pre-processing and model hyperparameters were used across all the experiments. All the models use the same input window of length 200 data points and the sliding window approach. The batch size was set to 1024 and the sampling period 6 seconds. Each experiment was executed 10 times on different random seeds on a Nvidia TitanX GPU.

Table 1. Summary of experiments.

Experiment	Environment Setup	Goal
Ablation study comparing the same network with and without variational inference.	Applied the first category of benchmarking [50], where training and inference happens on the same installation.	To highlight the performance boost due to the variational inference approach.
Performance comparison of three variations of the proposed network.	Applied the first two categories of benchmarking [50], where training and inference happens on installations of the same dataset.	To decide which combination method is the best.
Benchmark evaluation of performance of multi-target models.	Execute the first three categories of benchmarking [50] for four installations from two different datasets.	For performance comparison of the novel model versus the baseline.
Performance comparison between models that use the multi-target against those built based on the single-target strategy.	Execute the first two categories of benchmarking Symeonidis *et al.* [50], where training and inference happens on the same dataset.	For performance comparison of the novel multi-target model versus single-target baselines.
Performance comparison between multi-target models for different number of appliances.	Applied the first category of benchmarking [50], where training and inference happens on the same installation.	To highlight any performance boost or drop of the models.

In order to stress test the methods under examination the benchmark framework that is proposed by Symeonidis *et al.* [50] was loosely followed. This framework consists of four categories-scenarios of experiments that aim to highlight the strengths and weaknesses of the NILM detection algorithms. Category 1 of the benchmark involves training and inference on data from the same installation, whereas in category 2 the algorithms are trained and tested on different buildings from the same dataset. The third and fourth scenarios evaluate the learning capabilities of the models across many installations in combination with inference on the same and or different dataset.

In the current research the first two scenarios were executed as described in the original paper, whereas the third and the fourth were considered as one category and applied with a variation. To be specific, in the executed experiments the category 3 corresponds to training only on one dataset (UKDALE) and inference on the same (UKDALE) or another (REFIT). Hence, the ability of learning across many houses is not evaluated in this case.

Table 2 summarizes the installations used for the benchmark categories that are used for all the experiments in this study. The training period for UKDALE 1 is 4 months from 01 March 2013 to 01 Aug 2013, whereas one month of data was used for inference in all scenarios.

Table 2. Installations used for the current study. UKDALE installation 1 was used for training in all categories, whereas UKDAEL 1 and 2 were used for inference in categories 1-2. REFIT installations 2 & 5 were used for evaluation in category 3.

Appliance	**Category 1**		**Category 2**		**Category 3**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Dish Washer	1	1	1	2	1	2, 5
Microwave	1	1	1	2	1	2, 5
Fridge	1	1	1	2	1	2, 5
Kettle	1	1	1	2	1	2, 5
Washing Machine	1	1	1	2	1	2, 5

3.4. Evaluation Metrics

A good NILM algorithm should have two qualities. For starters, it should successfully detect the operation states of the appliances. This is a non trivial task since overlapping events of different appliances is a usual phenomenon and constitute the detection more difficult. Secondly, it should provide good power estimation of the detected end-uses. This is of high value since it concerns the users and the DSOs. As a result, the performance of NILM solutions should be evaluated with metrics that measure these properties.

In most of the NILM research two known machine learning metrics are used. The performance in operation states detection is measured with F_1 (2), the harmonic mean of *Precision* (3) and *Recall* (4). High *Precision* indicates low rate of false positives (FP), whereas high *Recall* means that the number of false negative (FN) is low. The F_1 is a combination of these two. In Equations 3 - 4, the number of true positives is notated as TP.

The ability to produce right power estimations is measured with the MAE. MAE is calculated in Watts and it is given by (5) where T is the length of the predicted sequence, $y_t{}'$ the estimated electrical power consumption and y_t the true value of active power consumption at moment t.

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$MAE = \frac{1}{T}\sum |y'_t - y_t| \quad (5)$$

4. Topology of Neural Networks

In order to verify the fact that the proposed solution has good performance the comparison with strong baseline known models is necessary. In the current case study two main cases of performance comparison involve baseline models; the benchmark of multi-target detection models and the comparison between multi-target and single-target models. For the first case two multi-target models were considered as baseline, while one of them is claimed to achieve highly accurate results. In the second scenario three known single-target architectures were chosen based on their performance and popularity.

An overview of the properties of the used models is depicted in Table 3. It is noticed that single-target models consist of a large number of more parameters in comparison to the multi-target architectures. Due to the fact that in this case one model corresponds to one appliance, the scenario of using many heavy algorithms for accurate disaggregation in production mode is unscalable. On the other hand, regarding the size of the networks, the multi-target are lighter with decent training and inference times for multiple appliance disaggregation simultaneously.

Table 3. Properties of the tested models. Number of parameters, size of the model, training speed (GPU), inference speed (GPU and CPU). For the Single-Target models the numbers are measured for experiments with one appliance.

Strategy	Appliances	Architecture	Params (Mil)	Size (MB)	Training GPU(it/s)	Inference GPU(it/s)
Single-Target	1	DAE	2.9	11.540	102.13	139.20
		S2P	10.3	41.160	18.390	78.16
		NFED	4.7	18.956	20.220	44.93
Multi-Target	5	CNN-base	2.2	8.650	30.960	82.74
		UNet-NILM	2.2	8.940	14.750	50.01
		V.M.Regressor (Linear)	2.1	8.376	18.901	59.40
		V.M.Regressor (Addition)	2.0	8.170	19.405	61.10
		V.M.Regressor (Attention)	2.0	8.171	19.290	60.20

4.1. Single-Target Baseline Models

For the comparison of the multi versus single-target strategies three known NILM architectures were chosen; the Denoising autoencoders, the Sequence-to-point and the Neural Fourier Energy Disaggregator. These models are different to each other since they were designed using different elements. Thus, the comparative study would not be too biased against similar architectures.

Denoising autoencoders is a family of neural networks that are designed to eliminate the noise from the input signal and output a clean one. In NILM the goal is to separate the appliance consumptions from the mains consumption of the installation. Hence, the mains time series plays the role of the noisy signal, whereas the individual energy consumption is the noiseless target. The original architecture of DAE was proposed by Vincent *et al.* [59] and later it was adapted in NILM by Kelly and Knottenbelt [23] as a series of fully connected / dense artificial layers. The architecture is depicted in Figure 1:

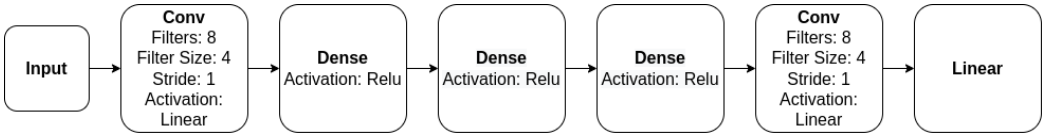


Figure 1. Architecture of DAE.

The model called Sequence-to-point (S2P) [24] is composed of a series of five convolutional layers that act as a feature extractor. These features are then passed through a dense layer with ReLU activation. S2P is considered a state-of-the-art and is used across many papers in the literature either as an inspiration and/or as baseline. The architecture of the network is summarized in Figure 2.

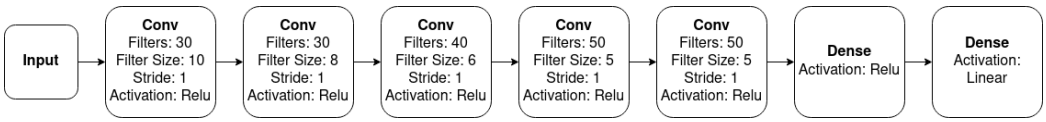


Figure 2. Architecture of S2P.

The Neural Fourier Energy Disaggregator (NFED) [60] could be considered as a member of the transformer family [61–64] due the fact that it was based on FNET [65], a transformer where the attention layer was replaced by Fourier transformation as an efficient alternative. In comparison to state-of-the-art-models, the NFED achieves similar performs on lower computational costs. NFED is composed of fully connected and normalised layers along with two main residual connections. The architecture is depicted in Figure 3.

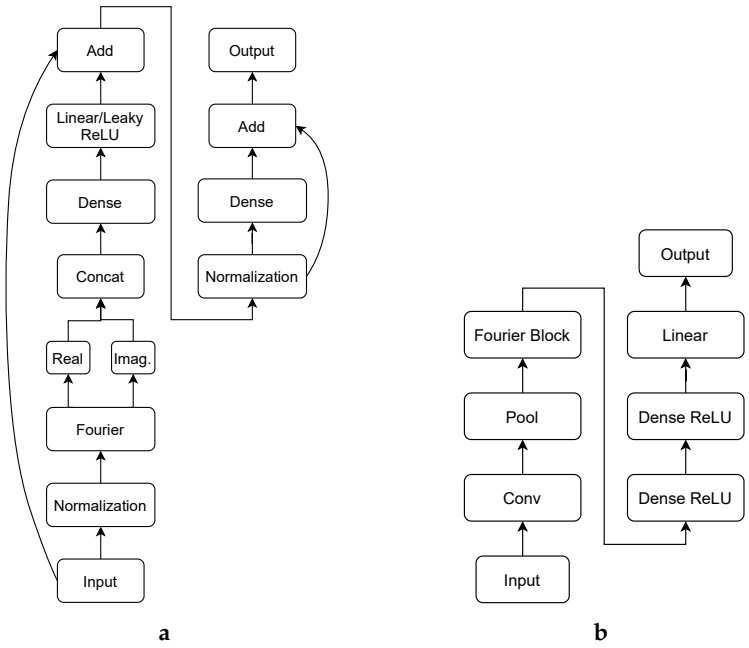


Figure 3. The NFED neural network. (**a**) Fourier block; (**b**) NFED architecture.

4.2. Multi-Target Baseline Models

In order to evaluate the proposed implementation, an adaption of the UNet NILM proposed by Faustine *et al.* [34] was used. As input the network receives the mains time series and it outputs both the appliance states and power time series. In the original paper, UNet NILM performs quartile multi target regression in a sequence-to-sequence fashion, where the length of the input is the same of the output. Quartile regression involves the smoothing of the mains and target time series with quartile filtering. This technique removes spikes and other features that may be valuable for successful disaggregation.

Our variation of the UNet NILM differs from the original in the following aspects. Firstly, the sliding window approach [27] was used instead of sequence-to-sequence. In this method, the input of the networks is a sequence and the output is the last disaggregate point of the sequence. Hence, almost real-time disaggregation could be achieved alongside with faster training and inference. Secondly, regular regression (without the quartile smoothing) was performed, in order to compare the two implementations on the same level. Due to

those changes, some parameters were adjusted for the model to perform at is best. Since the UNet NILM differs from the original implementation the CNN-base architecture described in Faustine *et al.* [34] was also used and adjusted accordingly in order to extract more insights about the performance of the models.

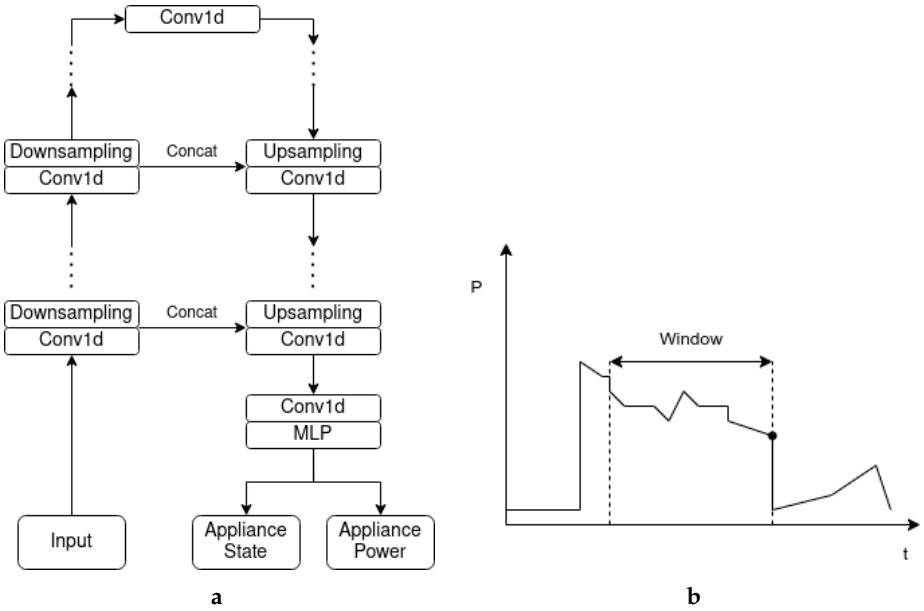


Figure 4. Architecture of UNet NILM. (**a**) UNet NILM; (**b**) Sliding window approach.

4.3. The Proposed Variational Multi-Target Regressor Architecture

As depicted in Figure 5, the novel implemetation is a combination of 4 basic modules; the ConvEncoder model (Figure 6), the ReparamTrick module (Figure 9), the Combination Mechanism (Figure 7) and the Shallow Regressor network (Figure 8). After training the model is supposed to output both the power consumption and the corresponding state of the target appliances.

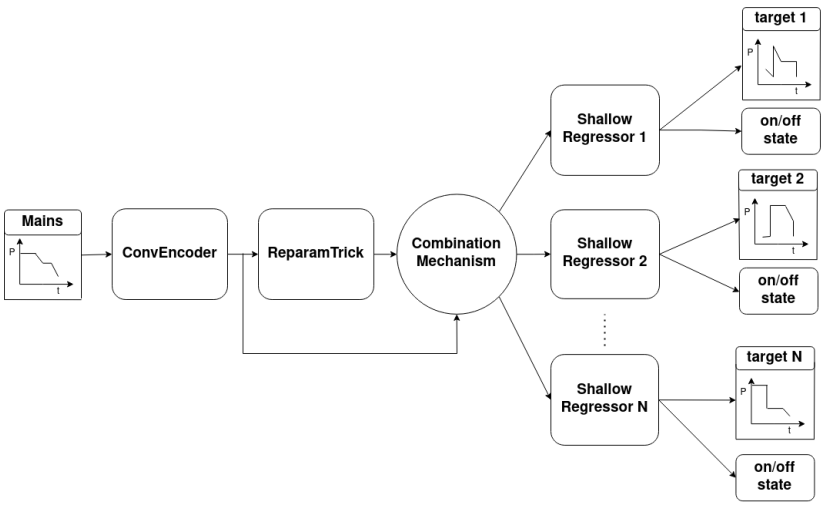


Figure 5. Architecture of Variational Multi-Target Regressor (V.M.Regressor).

The ConvEncoder is composed of a series of convolutional layers with the same kernel but with different number of filters that operate as a feature extractor. The output of the module goes through the ReparamTrick layer where the sampling through the reparametrization trick is executed. Then, the two vectors are combined producing a vector that contains the information from the extracted features and the encoding. The available

combination mechanisms are depicted in Figure 7 and essentially produce a vector with size equal to the size of ConvEncoder output. After observation during the designing of the architecture it was found out that the combination of the two vectors boosts the performance of the model.

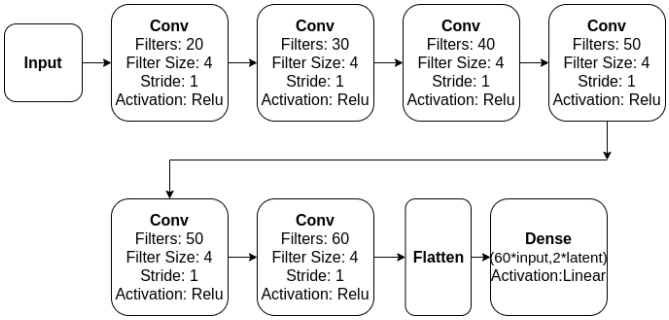


Figure 6. Architecture of ConvEncoder.

In the current study three lightweight and efficient combination methods were used. Firstly, a simple element wise addition of the two vectors is used. The addition at this case acts like a residual connection [66] between the input of the ReparamTrick module. The idea is to provide the model with some information extracted by the ConvEncoder in order to help the training and fight any degradation issues [67]. Secondly, a dot attention mechanism [68] in order to help the model focus on the most significant parts of the two vectors. In addition, a Dense neural layer with Linear activation was trained in order to learn how to combine the vectors automatically.

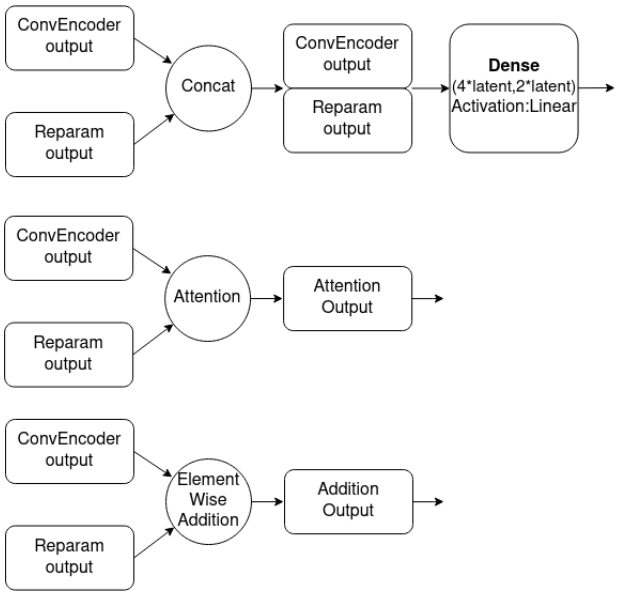


Figure 7. Overview of tested combination mechanisms and.

Finally, the product of the Combination mechanism is passed to all the ShallowRegressors to output the power and on/off estimation points for each target. Each ShallowRegressor is a series of lightweight fully connected layers that aim to filter out the unnecesary information and keep the valuable regarding the corresponding target appliance.

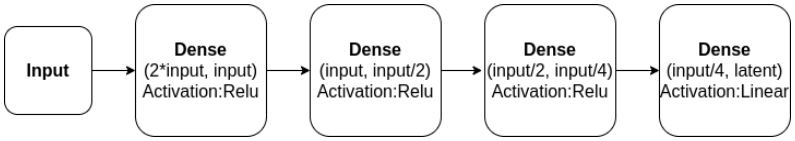


Figure 8. Architecture of Shallow Regressor.

The novel architecture is called Variational Multi-Target Regressor since it uses the concept of variational inference [55,69] in order to boost the performance of a multi-target regression network. The intuition is that the network learns a posterior distribution instead of point estimates. The posterior describes the target data more naturally that point estimates. Hence, the model is granted with the ability of dealing with unseen data points resulting to more generalised predictions. In order to learn the posterior, prior information is necessary as described by the Bayes rule 6, where given an input $x \epsilon R$ the unknown posterior $p(\mathbf{z} \mid \mathbf{x})$ equals to the likelihood $p(\mathbf{x} \mid \mathbf{z})$ times the prior $p(\mathbf{z})$ divided by the evidence $p(\mathbf{x})$. This prior information is inserted as hyper parameter and aims to direct the model towards the right answer.

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z})}{p(\mathbf{x})} \tag{6}$$

In NILM these posterior distributions are based on many parameters such as the electrical signature of each appliance, the frequency of operation, the duration of the end uses etc and they are generally hard to compute. As a result, an approximation of the posterior should take place. The idea of variational inference dictates that the unknown posterior can be approximated with another distribution q that comes from the same family as the prior's. Usually the steps for this process are; (a) choose a distribution family, (b) discover the member of the family that is closer to the original data distribution. The distance between the posterior and the approximation is measured with the KL-divergence [70].

In order to successfully estimate the target distributions, the output of the encoder is divided by the number of targets into equal vectors as shown in Figure 9. Then, mean and std are learned for each target vector and using the reparametrization trick the corresponding encoded vectors are sampled. Then, with the same statistics the KL-divergence for each target is computed. It should be noted that the proposed architecture and the various versions are implemented using Torch-NILM and the code is availabe at https://github.com/Virtsionis/torch-nilm, accessed on 5 January 2023.

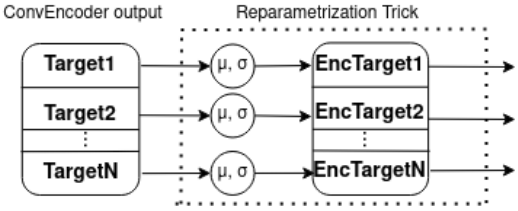


Figure 9. Reparametrization trick technique.

4.4. Loss function

As described earlier, the model approximates the posterior distribution, estimates the power consumption and on/off appliance states of each target appliance. In order to insert all this information to the training, a new loss function was designed and is presented in 7. This function consists of three different losses; the information loss, the regression loss and the classification loss. Additionally, three normalization factors were used to scale each loss separately for best performance. For all the experiments the values of beta, gamma and delta are 0.001, 1 and 10 correspondingly.

The information loss 8 is the sum of the KL-divergence between the posterior approximation q and the prior p for each target divided by the number of appliances N and is responsible for the posterior approximation. As regression loss 9 the sum of all the mean square errors between the targets and the ground truths is used, scaled by the number of appliances. Similarly, the binary cross entropy was calculated as the classification of the on/off states 10.

$$Loss = beta * info\_loss + gamma * class\_loss + delta * reg\_loss \tag{7}$$

$$info\_loss = \frac{1}{N}\sum_{n=1}^{N} KL(q_i(z|x)||p_i(z)) \tag{8}$$

$$reg\_loss = \frac{1}{N}\sum_{n=1}^{N} MSE(power_i, power_i') \tag{9}$$

$$class\_loss = \frac{1}{N}\sum_{n=1}^{N} binary\_cross\_entropy(state_i, state_i') \tag{10}$$

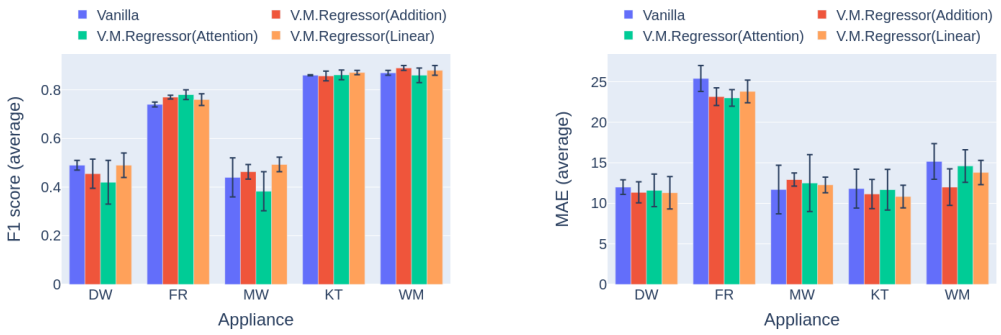
5. Experimental Results and Discussion

This section contains five experiments. To begin with, an ablation study is executed in order to verify that the variational inference approach boosts the performance of a vanilla multi-target regression model. In addition, a performance comparison between three variations of the proposed network is conducted to determine the best one. Next, in an effort to highlight the capabilities of the proposed network, benchmarking comparisons with two multi-target and three single-target architectures are conducted. Finally, an investigation regarding the relation between the model performance and the number of target appliances is performed.

5.1. Ablation Study - Variational Inference

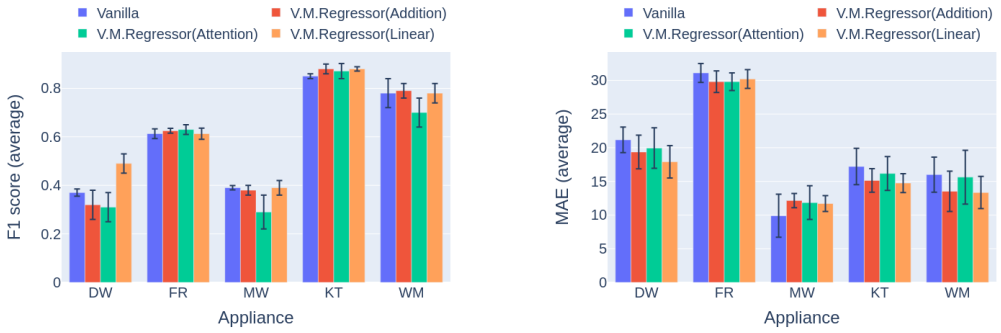
The goal of this investigation is to discover whether the variational inference approach assists the learning process of the proposed model. As a consequence, all the variations of the proposed model were compared side by side with the same network without the variational inference part. This model is called Vanilla and misses the ReparamTrick and Combination Mechanism modules. Thus the output of the ConvEncoder is directly passed to the ShallowRegressors. For this experiment the first two categories of the benchmark were applied, where data from UKDALE was used for training and inference.

Regarding the event detection the results in Figures 10a and 11a indicate that the proposed solution outperforms the vanilla variation in almost every case. In addition, the networks following the variational inference show better generalization capability on the unseen data 11a, reaching up to 27.9% higher performance in comparison to the vanilla implementation. The only cases where the vanilla achieves similar performance with the proposed counterpart is the dishwasher in category 1 and the microwave in category 2. What is more, in terms of the power estimation the results in Figures 10a and 11a dictate that the proposed solution achieves lower MAE errors in 9 out of 10 cases, meaning better estimation than the vanilla.



a **b**

Figure 10. Experiment 1: Ablation study to highlight the effect of variational inference on the performance of the model: (**a**) F1 in category 1: single building NILM; (**b**) MAE category 1: single building NILM.



a **b**

Figure 11. Experiment 1: Ablation study to highlight the effect of variational inference on the performance of the model: (**a**) F1 in category 2: Train and inference on different buildings of UKDALE; (**b**) MAE category 2: Train and inference on different buildings of UKDALE.

At this point it is useful to compare the models based on their computational cost. The properties of all the networks under investigation are depicted In Table 4. Comparing the Vanilla and the V.M.Regressor(Addition) it is obvious that the cost of integrating variational inference into the model is negligible. Additionally, the V.M.Regressor(Linear) is slightly heavier than the vanilla with the addition of 100K parameters, which slightly affects the training speed of the model.

Table 4. Properties of the ablation study on combination methods for five appliances. Number of parameters in millions, size of the model on the disk, training speed (GPU), inference speed (GPU).

Architecture	Parameters (millions)	Size on the disk (MB)	Training GPU (it/s)	Inference GPU (it/s)
Vanilla	2.0	8.170	19.8	62.2
V.M.Regressor (Addition)	2.0	8.170	19.4	61.1
V.M.Regressor (Attention)	2.0	8.171	19.3	60.2
V.M.Regressor (Linear)	2.1	8.376	18.9	59.4

5.2. Combination Mechanism Selection

A crucial point in the novel architecture is the way that the output of the ReparamTrick module is used. After experimentation the novel architecture is implemented into three variations depending on the combination mechanism; the simple addition mode, the attention implementation and the combination with a linear layer. In order to decide which mechanism is the best, the macro averages of the F1 score and MAE error are compute for the 3 categories of the benchmark. The macro averaging is essentially the simple average of the evaluation metrics across all the appliances. The results shown in Table 5 indicate that for the first two categories the model with the Linear combination mechanism achieves the best performance. Regarding the scenarios in category 3, the variant with the simple Addition out performs the other two. On the contrary, the architecture with attention showed the lowest performance in all scenarios. To highlight how close the overall performance of the variations with the Linear and the Addition mode are, the percentage differences between the averages of the two metrics per category were calculated and depicted in Table 6.

Table 5. Experiment 2: Performance comparison between the available combination methods. The macro averaging is the simple average of a metric across the five appliances.

Category	Train	Test	Combination	F1 macro	MAE macro
1	UKDALE 1	UKDALE 1	Addition	0.687	**14.118**
			Attention	0.661	14.676
			Linear	**0.699**	14.402
2	UKDALE 1	UKDALE 2	Addition	0.599	17.993
			Attention	0.56	18.68
			Linear	**0.631**	**17.578**
3	UKDALE 1	REFIT 2	Addition	**0.506**	**25.678**
			Attention	0.446	29.402
			Linear	0.481	26.098
3	UKDALE 1	REFIT 5	Addition	**0.43**	**32.829**
			Attention	0.404	36.334
			Linear	0.412	**32.86**

Table 6. Percentage differences between the average macro scores per category of Experiment 2.

Category	Addition (F1)	Linear (F1)	F1 diff(%)	Addition (MAE)	Linear (MAE)	MAE diff(%)
1 & 2	0.643	**0.665**	3.364	16.055	**15.99**	0.405
3	**0.493**	0.484	1.843	**29.254**	29.479	0.766

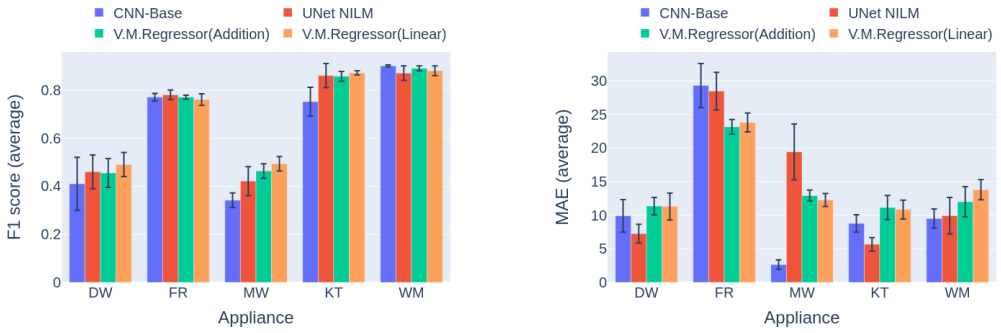
5.3. Comparison to Multi-Target baseline

The third experiment is a direct comparison of the novel deep learning solution versus two multi-target architectures that were introduced by Faustine *et al.* [56]; the UNet NILM and a CNN-Base network. The UNet NILM in the original implementation achieved high performance in comparison to the CNN-Base model on experiments with the UKDALE dataset and it is considered a strong opponent. In the current work the UNet NILM is adjusted to perform regular instead of quartile regression following the sliding window approach [27] shown in Figure 4b. In this comparison the best of the two variations of V.M.Regressor were used, the versions with the Addition and the Linear combination mechanism.

In the first category of experiments the house 1 of UKDALE is used for training and inference. This category is the most usual case in real world applications, where a dedicated disaggregation model per residence is trained. As can be seen in Figure 12, the V.M.Regressor (Linear) architecture reaches the maximum F1 score for 3 out of 5 target appliances, whereas for the remaining 2 there is an negligible difference between the state-of-the-art. On the other hand, in terms of the MAE error there is not a clear winner, with the UNet NILM winning in 3 occasions and the rest of the models winning on 1 appliance each. Even though the V.M.Regressor is behind the state-of-the-art in power estimation it should be noted that the maximum absolute difference in MAE error is observed during the kettle disaggregation and is under 6 Watts. Considering the fact that a regular kettle operates at a maximum power level of around 2000 Watts this difference is not very significant.

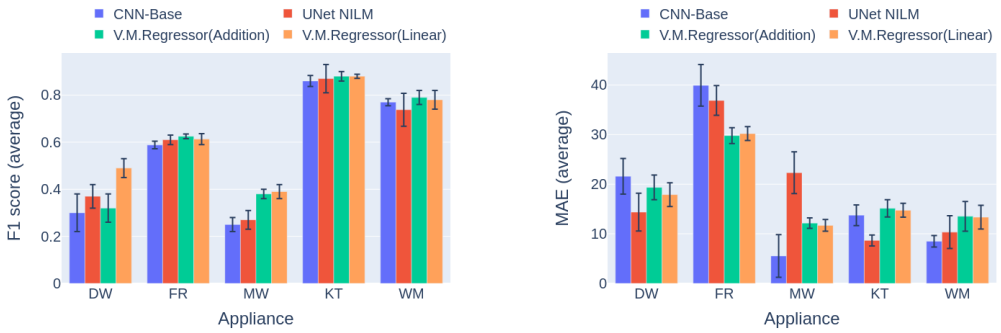
The results in category 2 of experiments are pictured in Figure 13 and show that the V.M.Regressor (Linear) is the clear winner in terms of F1 score with 3.2% average difference across the five target appliances. On appliance level, the largest differences in F1 score occur in dishwasher and microwave detection, with 6.3% and 9.5% accordingly. It should be noted that this category uses measurements from different installations of the same dataset for train and inference. Thus, an overall drop in performance is expected for many reasons such as the fact that there is is a great possibility that there are different appliance models in the houses or the routine and habits of the residents may differ significantly. This may explain the large drop performance of all models in the washing machine and fridge disaggregation. The promising fact here is that the proposed model retained similar performance on the rest

of the appliances. This highlights that the generalization capability of the V.M.Regressor is better that the models in comparison. Regarding the power estimation and the MAE error there are mixed results. Specifically, the UNet NILM model achieves lower MAE errors for 3 appliances in comparison to the V.M.Regressors that performed better only on the fridge appliance. In this case the simple CNN-Base model performed better than the others in the microwave and the washing machine.



a **b**

Figure 12. Experiment 3: Performance comparison of multi-target models in category 1 where training and inference on happen on UKDALE 1 house: (**a**) F1 in category 1, higher is better; (**b**) MAE in category 1, lower is better.



a **b**

Figure 13. Experiment 3: Performance comparison of multi-target models in category 2, where training happens on UKDALE 1 house and inference on UKDALE 2 house: (**a**) F1 in category 2, higher is better; (**b**) MAE in category 2, lower is better.

Category 3 of experiments is a greater challenge for the models. As the previous category the training and inference are applied to data from different installations. The difference here is that the installations are part of different datasets. This fact introduces many more possibilities and reasons for the models to under-perform due to regionality, every day habits, culture etc. In the current comparison two use cases were explored. The first one concerns training on UKDALE 1 and inference on REFIT 2 houses. The results of this scenario are depicted in Figure 14. It is notable that the V.M.Regressor variations are the clear winners in 4 out of 5 appliances regarding both the F1 score and the MAE error. Yet again, the V.M.Regressor shows good generlization capacity, being able to out perform the competition.

The second use case involves the same datasets but the REFIT 5 house for inference. In this case the novel neural network achieves the best event detection in 3 appliances with the simple CNN-BASE winning in the disaggregation of the fridge and the microwave.

In case of MAE metric, the V.M.Regressors reached the lowest values in the fridge and the microwave, whereas the UNet NILM performed better for the dishwasher and the washing machine. At the same time UNet performed worse than the baseline in the microwave power estimation producing the highest observed error across all experiments in the manuscript.

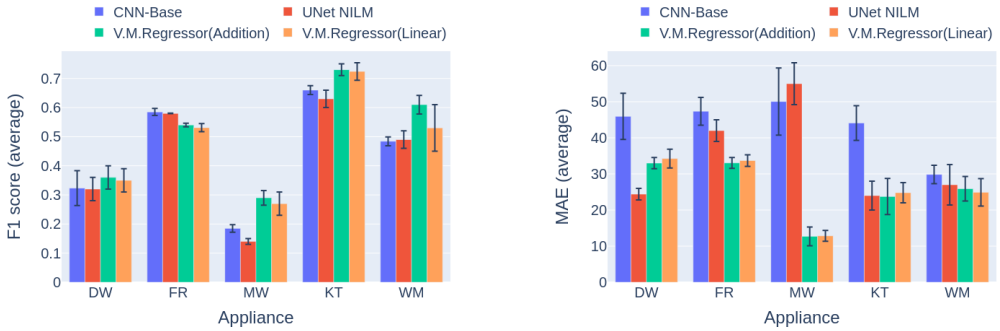


Figure 14. Experiment 3: Performance comparison of multi-target models in category 3, where training happens on UKDALE 1 house and inference on REFIT 2 house: (**a**)F1 in category 3, higher is better; (**b**) MAE in category 3, lower is better

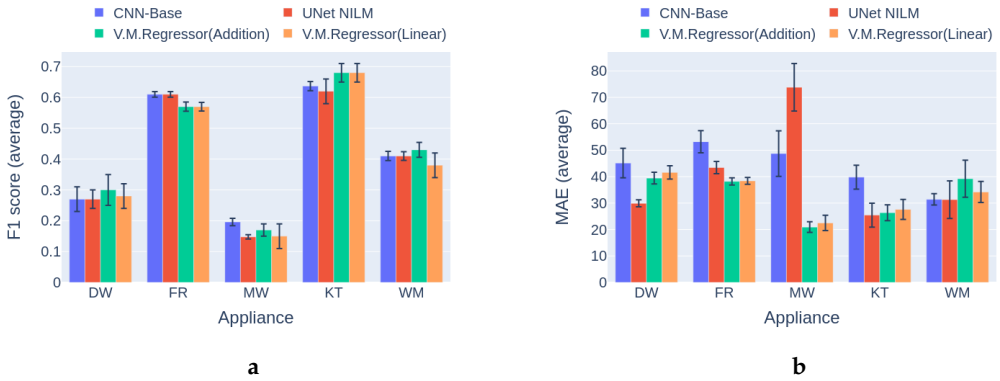


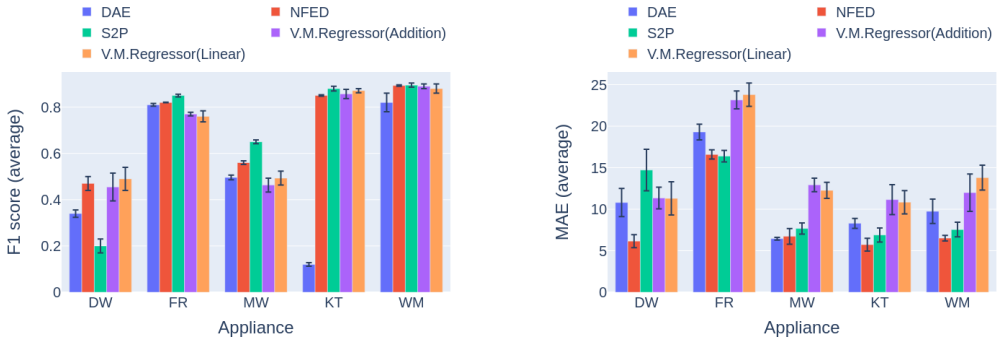
Figure 15. Experiment 3: Performance comparison of multi-target models in category 3, where training happens on UKDALE 1 house and inference on REFIT 2 house: (**a**) F1 in category 3, higher is better; (**b**) MAE in category 3, lower is better

5.4. Comparison to Single-Target models

Since the largest part of the NILM research revolves around single-target solutions it is found useful to compare some of them with the proposed network. Specifically, the S2P architecture proposed by Zhang *et al.* [24] is considered to produce state-of-the-art performance and it is used in almost every NILM paper as a strong baseline. The NFED [60] model is claimed to achieve similar performance using less computational resources. Finally, the DAE architecture [23] is one of the first architectures adjusted to solve the problem of energy disaggregation and it was included due to its popularity and high training and inference speeds. It should be noted that for this experiment the first two categories of the benchmark were used, with training and inference happening on data from the UKDALE dataset.

The results for the first category are presented in Figure 16a. Regarding the F1 score the V.M.Regressor outperforms the single-target models in dishwasher disaggregation, whereas it achieves similar performance in event detection of the kettle and the washing machine. For the fridge and the microwave the single-target models produce higher F1 measures. In terms of the MAE error, the proposed model produced the highest values for 4

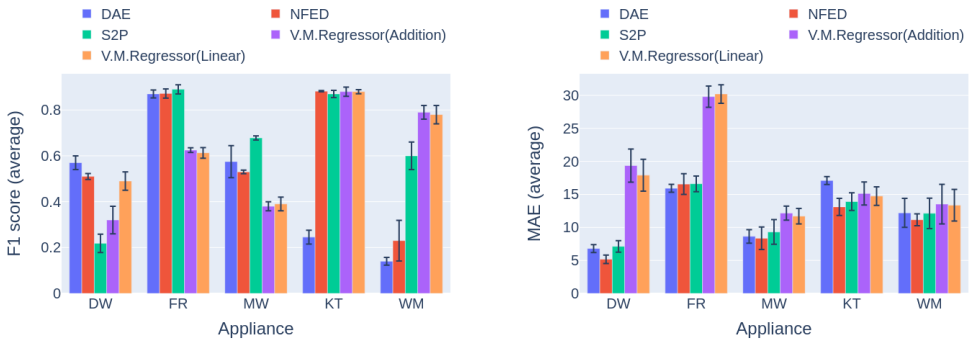
out of 5 of the appliances except the dishwasher. The clear winner in the MAE comparison is the NFED, achieving the lowest errors for 4 out of 5 appliances. Although the model shows higher errors, the absolute differences are under 5 Watts, except the case of the fridge. As a result, the novel network could be easily applied to a practical NILM providing results similar to single-target state-of-the art models with almost 25 times lower number of parameters and 5 times smaller training time in case of 5 target appliances, as shown in Table 3.



a **b**

Figure 16. Experiment 4: single-target vs multi-target models in category 1, where training and inference on happen on UKDALE 1 house: (**a**) F1 in category 1, higher is better; (**b**) MAE in category 1, lower is better

The final comparison in this experiment is based on the second category of benchmark [50]. After reviewing the results in Figure 17 it is obvious that the V.M.Regressor out performs the single-target models in terms of F1 score for the washing machine, with similar performance for the kettle and the dishwasher. For the rest of appliances, there is a large difference between the single-target models. It should be noted that there is not a single-target model that performs the same in disaggregating all the appliances. Hence, a different model could be more applicable for specific appliance disaggregation. On the aspect of power estimation all the models produce similar errors for 3 out of 5 appliances, except for the dishwasher and the fridge where the V.M.Regressor produces errors almost 15 Watts larger than the competition reaching to almost 30 Watts for the fridge. The new types of fridges usually operate around 80-150 Watts, meaning that 30 Watts of deviation in power estimation is almost 20-38% of the total power level. On the other hand 20 Watts miscalculation of an average dishwasher end use corresponds to 5-10% of the operation power level, which could be tolerated.



a **b**

Figure 17. Experiment 4: single-target vs multi-target models in category 2, where training happens on UKDALE 1 and inference on UKDALE 2 houses: (**a**) F1 score in category 2, higher is better; (**b**) MAE in category 2, lower is better.

5.5. Performance for different number of appliances

The number of appliances that the model can detect successfully is an important parameter of a practical NILM system. Thus, in the last experiment the performance of the proposed network and a baseline are compared on different sets of appliances. For the set of two appliances the kettle and the microwave are used. The set of four appliances contains the kettle, the microwave, the fridge and the washing machine. In the set of 6 appliances the dishwasher and the toaster are included. Finally, lights and electric boiler are adder in the set of 8 devices.

The experiment uses data only from UKDALE house 1, in ration 4/1 months for train/inference. The results are presented in Figure 18. It is notable from the results that the performance of the models follow a similar trajectory; both of them reach the maximum at four appliances and the minimum at eight, whereas in case of simultaneous disaggreation of six appliances the curves intersect. It should be mentioned that the proposed model outperforms the baseline at all cases.

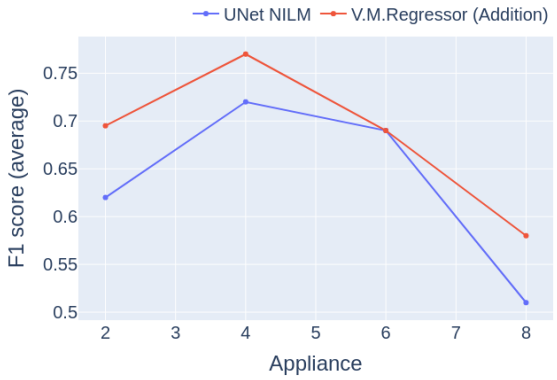


Figure 18. Experiment 4: Macro F1 score per number of appliances, for cat.1 Train and inference on UKDALE 1.

6. Conclusions and Future Work

Applying deep learning networks to practical Non Intrusive Load Monitoring applications is a non trivial task. The root cause is that state-of-the-art architectures usually consist of a great number of parameters. In combination with the fact that these networks are usually designed to disaggregate only one appliance at a time the training and inference speeds alongside the overall size of the solution are harshly affected. Since, this kind of systems are usually built and operate on the cloud, a high cost is introduced.

In this article V.M.Regressor, a cutting edge deep learning architecture, is proposed as a solution to real world NILM systems. V.M.Regressor is capable of high quality simultaneous multi-target disaggregation with the minimum computational requirements. The novel model outperforms a known state-of-the-art multi-target model with similar size and faster training and inference speeds, while it competes with state-of-the-art heavier single-target networks. The proposed model is build upon the principals of variational inference, an idea that boosts the performance and the generalization capability on unseen data.

For future work, the following suggestions could be made. To begin with, the concept of variational inference could be used to produce more multi-target solutions. Due to the fact that the integration of this concept does not increase the number of the model parameters it can also be applied to boost the performance of lightweight architectures capable of running on embedded appliances. Furthermore, in order to increase the generalization of this type of models training on many different datasets could be executed.

Author Contributions: Conceptualization, N.V.G. and C.N; methodology, N.V.G. and C.N; software, N.V.G. and C.N; validation, N.V.G., C.N and D.V.; formal analysis, N.V.G. and C.N.; investigation,

N.V.G.; resources, D.V.; data curation, N.V.G.; writing—original draft preparation, N.V.G.; writing—review and editing, N.V.G. and C.N; visualization, N.V.G.; supervision, D.V.; project administration, D.V.; funding acquisition, D.V. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hart, G.W. Nonintrusive appliance load monitoring. *Proceedings of the IEEE* **1992**, *80*, 1870–1891.
2. Pal, M.; Roy, R.; Basu, J.; Bepari, M.S. Blind source separation: A review and analysis. 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013, pp. 1–5. doi:10.1109/ICSDA.2013.6709849.
3. Naghibi, B.; Deilami, S. Non-intrusive load monitoring and supplementary techniques for home energy management. 2014 Australasian Universities Power Engineering Conference (AUPEC). IEEE, 2014, pp. 1–5.
4. Mahapatra, B.; Nayyar, A. Home energy management system (HEMS): concept, architecture, infrastructure, challenges and energy management schemes. *Energy Systems* **2019**, pp. 1–27.
5. Nalmpantis, C.; Vrakas, D. Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation. *Artificial Intelligence Review* **2019**, *52*, 217–243.
6. Lin, Y.H. A Parallel Evolutionary Computing-Embodied Artificial Neural Network Applied to Non-Intrusive Load Monitoring for Demand-Side Management in a Smart Home: Towards Deep Learning. *Sensors* **2020**, *20*. doi:10.3390/s20061649.
7. Angelis, G.F.; Timplalexis, C.; Krinidis, S.; Ioannidis, D.; Tzovaras, D. NILM Applications: Literature review of learning approaches, recent developments and challenges. *Energy and Buildings* **2022**, *261*, 111951. doi:10.1016/j.enbuild.2022.111951.
8. Alcalá, J.; Parson, O.; Rogers, A. Detecting Anomalies in Activities of Daily Living of Elderly Residents via Energy Disaggregation and Cox Processes. Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments; Association for Computing Machinery: New York, NY, USA, 2015; BuildSys '15, p. 225âĂŞ234. doi:10.1145/2821650.2821654.
9. Bousbiat, H.; Klemenjak, C.; Leitner, G.; Elmenreich, W. Augmenting an Assisted Living Lab with Non-Intrusive Load Monitoring. 2020, pp. 1–5. doi:10.1109/I2MTC43012.2020.9128406.
10. Athanasiadis, C.L.; Pippi, K.D.; Papadopoulos, T.A.; Korkas, C.; Tsaknakis, C.; Alexopoulou, V.; Nikolaidis, V.; Kosmatopoulos, E. A Smart Energy Management System for Elderly Households. 2022 57th International Universities Power Engineering Conference (UPEC), 2022, pp. 1–6. doi:10.1109/UPEC55022.2022.9917856.
11. Donato, P.; Carugati, I.; HernÃąndez, Ã.; Nieto, R.; Funes, M.; UreÃśa, J. Review of NILM applications in smart grids: power quality assessment and assisted independent living. 2020. doi:10.23919/AADECA49780.2020.9301641.
12. Bucci, G.; Ciancetta, F.; Fiorucci, E.; Mari, S.; Fioravanti, A. State of art overview of Non-Intrusive Load Monitoring applications in smart grids. *Measurement: Sensors* **2021**, *18*, 100145. doi:https://doi.org/10.1016/j.measen.2021.100145.
13. Massidda, L.; Marrocu, M. A Bayesian Approach to Unsupervised, Non-Intrusive Load Disaggregation. *Sensors* **2022**, *22*. doi:10.3390/s22124481.
14. Kaur, D.; Islam, S.; Mahmud, M.A.; Haque, M.; Dong, Z. Energy forecasting in smart grid systems: recent advancements in probabilistic deep learning. *IET Generation, Transmission Distribution* **2022**, *16*, n/a–n/a. doi:10.1049/gtd2.12603.
15. Alzaatreh, A.; Mahdjoubi, L.; Gething, B.; Sierra, F. Disaggregating high-resolution gas metering data using pattern recognition. *Energy and Buildings* **2018**, *176*, 17–32. doi:https://doi.org/10.1016/j.enbuild.2018.07.011.
16. Ellert, B.; Makonin, S.; Popowich, F. Appliance Water Disaggregation via Non-Intrusive Load Monitoring (NILM). 2015, Vol. 166. doi:10.1007/978-3-319-33681-7_38.
17. Pastor-Jabaloyes, L.; Arregui, F.J.; Cobacho, R. Water End Use Disaggregation Based on Soft Computing Techniques. *Water* **2018**, *10*. doi:10.3390/w10010046.
18. Gkalinikis, N.V.; Vrakas, D. Efficient Deep Learning Techniques for Water Disaggregation. 2022 2nd International Conference on Energy Transition in the Mediterranean Area (SyNERGY MED), 2022, pp. 1–6. doi:10.1109/SyNERGYMED55767.2022.9941424.
19. Kim, H.; Marwah, M.; Arlitt, M.; Lyon, G.; Han, J., Unsupervised Disaggregation of Low Frequency Power Measurements. In *Proceedings of the 2011 SIAM International Conference on Data Mining*; SIAM, 2011; pp. 747–758. doi:10.1137/1.9781611972818.64.
20. Kolter, J.Z.; Jaakkola, T. Approximate inference in additive factorial hmms with application to energy disaggregation. Artificial intelligence and statistics, 2012, pp. 1472–1482.

21. Parson, O.; Ghosh, S.; Weal, M.J.; Rogers, A.C. Non-Intrusive Load Monitoring Using Prior Models of General Appliance Types **2012**. *26*.
22. Fortuna, L.; Buscarino, A. Non-Intrusive Load Monitoring. *Sensors* **2022**, *22*. doi:10.3390/s22176675.
23. Kelly, J.; Knottenbelt, W. Neural nilm: Deep neural networks applied to energy disaggregation. Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments, 2015, pp. 55–64.
24. Zhang, C.; Zhong, M.; Wang, Z.; Goddard, N.; Sutton, C. Sequence-to-point learning with neural networks for nonintrusive load monitoring. *AAAI* **2018**.
25. Jia, Z.; Yang, L.; Zhang, Z.; Liu, H.; Kong, F. Sequence to point learning based on bidirectional dilated residual network for non-intrusive load monitoring. *International Journal of Electrical Power & Energy Systems* **2021**, *129*, 106837.
26. Mauch, L.; Yang, B. A new approach for supervised power disaggregation by using a deep recurrent LSTM network. 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, 2015, pp. 63–67.
27. Krystalakos, O.; Nalmpantis, C.; Vrakas, D. Sliding window approach for online energy disaggregation using artificial neural networks. Proceedings of the 10th Hellenic Conference on Artificial Intelligence, 2018, pp. 1–6.
28. Kaselimi, M.; Doulamis, N.; Doulamis, A.; Voulodimos, A.; Protopapadakis, E. Bayesian-optimized Bidirectional LSTM Regression Model for Non-intrusive Load Monitoring. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2747–2751. doi:10.1109/ICASSP.2019.8683110.
29. Fang, Z.; Zhao, D.; Chen, C.; Li, Y.; Tian, Y. Non-Intrusive Appliance Identification with Appliance-Specific Networks. 2019 IEEE Industry Applications Society Annual Meeting, 2019, pp. 1–8. doi:10.1109/IAS.2019.8912379.
30. Moradzadeh, A.; Mohammadi-Ivatloo, B.; Abapour, M.; Anvari-Moghaddam, A.; Farkoush, S.G.; Rhee, S.B. A practical solution based on convolutional neural network for non-intrusive load monitoring. *Journal of Ambient Intelligence and Humanized Computing* **2021**, pp. 1–15.
31. Faustine, A.; Pereira, L.; Bousbiat, H.; Kulkarni, S. UNet-NILM: A Deep Neural Network for Multi-Tasks Appliances State Detection and Power Estimation in NILM. Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring; Association for Computing Machinery: New York, NY, USA, 2020; NILM'20, p. 84âĂŞ88. doi:10.1145/3427771.3427859.
32. Virtsionis-Gkalinikis, N.; Nalmpantis, C.; Vrakas, D. SAED: self-attentive energy disaggregation. *Machine Learning* **2021**, pp. 1–20.
33. Langevin, A.; Carbonneau, M.A.; Cheriet, M.; Gagnon, G. Energy disaggregation using variational autoencoders. *Energy and Buildings* **2022**, *254*, 111623. doi:https://doi.org/10.1016/j.enbuild.2021.111623.
34. Faustine, A.; Pereira, L.; Bousbiat, H.; Kulkarni, S. UNet-NILM: A Deep Neural Network for Multi-Tasks Appliances State Detection and Power Estimation in NILM. Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring; Association for Computing Machinery: New York, NY, USA, 2020; NILM'20, p. 84âĂŞ88. doi:10.1145/3427771.3427859.
35. Piccialli, V.; Sudoso, A. Improving Non-Intrusive Load Disaggregation through an Attention-Based Deep Neural Network. *Energies* **2021**, *14*, 847. doi:10.3390/en14040847.
36. Gkalinikis, N.V.; Nalmpantis, C.; Vrakas, D. Attention in Recurrent Neural Networks for Energy Disaggregation. International Conference on Discovery Science. Springer, 2020, pp. 551–565.
37. Yue, Z.; Witzig, C.R.; Jorde, D.; Jacobsen, H.A. BERT4NILM: A Bidirectional Transformer Model for Non-Intrusive Load Monitoring. Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring; Association for Computing Machinery: New York, NY, USA, 2020; NILM'20, p. 89âĂŞ93. doi:10.1145/3427771.3429390.
38. Pan, Y.; Liu, K.; Shen, Z.; Cai, X.; Jia, Z. Sequence-To-Subsequence Learning With Conditional Gan For Power Disaggregation. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 3202–3206. doi:10.1109/ICASSP40776.2020.9053947.
39. Bejarano, G.; DeFazio, D.; Ramesh, A. Deep latent generative models for energy disaggregation. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 850–857.
40. Sirojan, T.; Phung, B.T.; Ambikairajah, E. Deep neural network based energy disaggregation. 2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE). IEEE, 2018, pp. 73–77.
41. Langevin, A.; Carbonneau, M.A.; Cheriet, M.; Gagnon, G. Energy disaggregation using variational autoencoders. *Energy and Buildings* **2022**, *254*, 111623.
42. Harell, A.; Jones, R.; Makonin, S.; Bajic, I.V. PowerGAN: Synthesizing Appliance Power Signatures Using Generative Adversarial Networks. *arXiv e-prints* **2020**, p. arXiv:2007.13645, [arXiv:eess.SP/2007.13645].
43. Ahmed, A.M.A.; Zhang, Y.; Eliassen, F. Generative Adversarial Networks and Transfer Learning for Non-Intrusive Load Monitoring in Smart Grids. 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), 2020, pp. 1–7. doi:10.1109/SmartGridComm47815.2020.9302933.
44. Kukunuri, R.; Aglawe, A.; Chauhan, J.; Bhagtani, K.; Patil, R.; Walia, S.; Batra, N. EdgeNILM: Towards NILM on Edge Devices. Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation; Association for Computing Machinery: New York, NY, USA, 2020; BuildSys '20, p. 90âĂŞ99. doi:10.1145/3408308.3427977.
45. Basu, K.; Debusschere, V.; Bacha, S. Load identification from power recordings at meter panel in residential households. 2012 XXth International Conference on Electrical Machines. IEEE, 2012, pp. 2098–2104.
46. Basu, K.; Debusschere, V.; Bacha, S. Residential appliance identification and future usage prediction from smart meter. IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2013, pp. 4994–4999.

47. Tabatabaei, S.M.; Dick, S.; Xu, W. Toward non-intrusive load monitoring via multi-label classification. *IEEE Transactions on Smart Grid* **2016**, *8*, 26–40.
48. Nalmpantis, C.; Vrakas, D. On time series representations for multi-label NILM. *NEURAL COMPUTING & APPLICATIONS* **2020**.
49. Athanasiadis, C.L.; Papadopoulos, T.A.; Doukas, D.I. Real-time non-intrusive load monitoring: A light-weight and scalable approach. *Energy and Buildings* **2021**, *253*, 111523. doi:https://doi.org/10.1016/j.enbuild.2021.111523.
50. Symeonidis, N.; Nalmpantis, C.; Vrakas, D. A Benchmark Framework to Evaluate Energy Disaggregation Solutions. International Conference on Engineering Applications of Neural Networks. Springer, 2019, pp. 19–30.
51. Batra, N.; Kukunuri, R.; Pandey, A.; Malakar, R.; Kumar, R.; Krystalakos, O.; Zhong, M.; Meira, P.; Parson, O. Towards reproducible state-of-the-art energy disaggregation. Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation, 2019, pp. 193–202.
52. Virtsionis Gkalinikis, N.; Nalmpantis, C.; Vrakas, D. Torch-NILM: An Effective Deep Learning Toolkit for Non-Intrusive Load Monitoring in Pytorch. *Energies* **2022**, *15*. doi:10.3390/en15072647.
53. Klemenjak, C.; Makonin, S.; Elmenreich, W. Towards comparability in non-intrusive load monitoring: on data and performance evaluation. 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). IEEE, 2020, pp. 1–5.
54. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **2014**, *61*. doi:10.1016/j.neunet.2014.09.003.
55. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014, [http://arxiv.org/abs/1312.6114v10].
56. Faustine, A.; Pereira, L.; Bousbiat, H.; Kulkarni, S. UNet-NILM: A Deep Neural Network for Multi-tasks Appliances State Detection and Power Estimation in NILM. 2020, pp. 84–88. doi:10.1145/3427771.3427859.
57. Jack, K.; William, K. The UK-DALE dataset domestic appliance-level electricity demand and whole-house demand from five UK homes. *Sci. Data* **2015**, *2*, 150007.
58. Firth, S.; Kane, T.; Dimitriou, V.; Hassan, T.; Fouchal, F.; Coleman, M.; Webb, L. REFIT Smart Home dataset, 2017. doi:10.17028/rd.lboro.2070091.v1.
59. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and Composing Robust Features with Denoising Autoencoders. Proceedings of the 25th International Conference on Machine Learning; Association for Computing Machinery: New York, NY, USA, 2008; ICML '08, p. 1096âĂŞ1103. doi:10.1145/1390156.1390294.
60. Nalmpantis, C.; Virtsionis Gkalinikis, N.; Vrakas, D. Neural Fourier Energy Disaggregation. *Sensors* **2022**, *22*. doi:10.3390/s22020473.
61. Choromanski, K.M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.Q.; Mohiuddin, A.; Kaiser, L.; et al. Rethinking Attention with Performers. International Conference on Learning Representations, 2020.
62. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. Proceedings of the International Conference on Machine Learning (ICML), 2020.
63. Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; Li, H. Efficient Attention: Attention with Linear Complexities. *CoRR* **2018**, *abs/1812.01243*.
64. Kitaev, N.; Kaiser, L.; Levskaya, A. Reformer: The Efficient Transformer. International Conference on Learning Representations, 2020.
65. Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; Ontanon, S. FNet: Mixing Tokens with Fourier Transforms. *arXiv preprint arXiv:2105.03824* **2021**.
66. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2015**, pp. 770–778.
67. Monti, R.P.; Tootoonian, S.; Cao, R. Avoiding Degradation in Deep Feed-Forward Networks by Phasing Out Skip-Connections. Artificial Neural Networks and Machine Learning – ICANN 2018; Kůrková, V.; Manolopoulos, Y.; Hammer, B.; Iliadis, L.; Maglogiannis, I., Eds.; Springer International Publishing: Cham, 2018; pp. 447–456.
68. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* **2015**, pp. 1412–1421.
69. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians **2018**.
70. Joyce, J.M., Kullback-Leibler Divergence. In *International Encyclopedia of Statistical Science*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp. 720–722. doi:10.1007/978-3-642-04898-2_327.