Università degli studi di
Milano-Bicocca

Advanced Machine Learning

Final Project

Electical Motor Temperature

Authors:
Federico Moiraghi - 799735 -
f.moiraghimotta@campus.unimib.it
Pranav Kasela - 846965 - p.kasela@campus.unimib.it
Roberto Berlucchi - 847939 - r.berlucchi@campus.unimib.it

2019 - 2020

Abstract

The present Report is a summary of methodologies used to predict the temperature of various part of a prototype electrical motor after some tests on a bench. The resulting model have to yield acceptable predictions and to be light enough to be used by the car itself during its daily use: autos can start cooling components before the temperature grows critically (first task) or can estimate temperature without a specific sensor (second task), due to its cost and weakness, knowing only basic environmental information.

1 Introduction

The data set comprises several sensor data collected from a permanent magnet synchronous motor (PMSM) deployed on a test bench. The PMSM represents a German OEM's prototype model and test bench measurements were collected by the LEA department at Paderborn University.

Recordings are sampled at a frequency of $2Hz$ and are divided in various profiles, with a total of 998070 observations. Each profile indicates a different test session, that last from one to six hours.

Input variables are:

- **Ambient temperature**: as measured by a thermal sensor located close to the stator;
- **Coolant temperature**: the motor is water cooled and the measurement is taken at outflow;
- The current and voltage are transformed through a $dq0$ transformation in a d-q coordinate system, it basically converts a three phase balanced reference system (in an AC system) into 2 coordinates, denoted by d and q, via a rotating reference frame with angle θ. The currents are denoted by **i_d** and **i_q** and the voltages are denoted by **u_d** and **u_q**;
- **Motor speed**.

Target variables are:

- **pm**: Permanent Magnet surface temperature, representing the rotor temperature (measured with an infrared thermography unit).
- **stator_yoke**, **stator_tooth**, **stator_winding**: temperatures of the corresponding components measured with a thermal sensor.

In some of the variables, Gaussian noise is introduced to simulate real world driving cycles. Being sensors data, missing values are replaced by the provider with the previous one, causing some flat areas when sensors fall offline for a long period.

The main objective is to create a lightweight model to predict the `pm` and `stator`s variables, minimizing the MSE (because the model needs to be deployed with best cost-precision ratio); a secondary goal is to predict more accurately higher temperature than the lower temperature using a modified loss.

2 Datasets

The data set can be found on Kaggle[1]. From the data set the `torque` feature is immediately excluded, as it is deemed unreliable from the data set provider itself.
The data set is divided into train, validation and test set: validation data consists of `profile_id`s $20, 31, 46, 54, 62, 70, 79, 72$; the test set profiles are 35 and 42 and the training set consists of all the other profiles. Their relative distributions are plotted in the Figure 1.

Figure 2 shows the correlation between the variables and it is seen that the target variables are highly correlated among themselves, in particular the `stator`s variables.

Data was already standardized by the provider (to anonymize data), but variables do not have a normal distribution, thus a further normalization between 0 and 1 (only on the training set) is applied.

3 The Methodological Approach

Different *Deep Learning* architectures are tried in order to compare them and choose the most suitable model to the problem: each model is built using `pytorch` and optimized with an Auto-ML algorithm using `sherpa` optimization library. In particular the *Gaussian Process* (GP) surrogate model is used with the *Expected Improvement* (EI) acquisition function to have a better exploration during the optimization.

[1] https://www.kaggle.com/wkirgsn/electric-motor-temperature

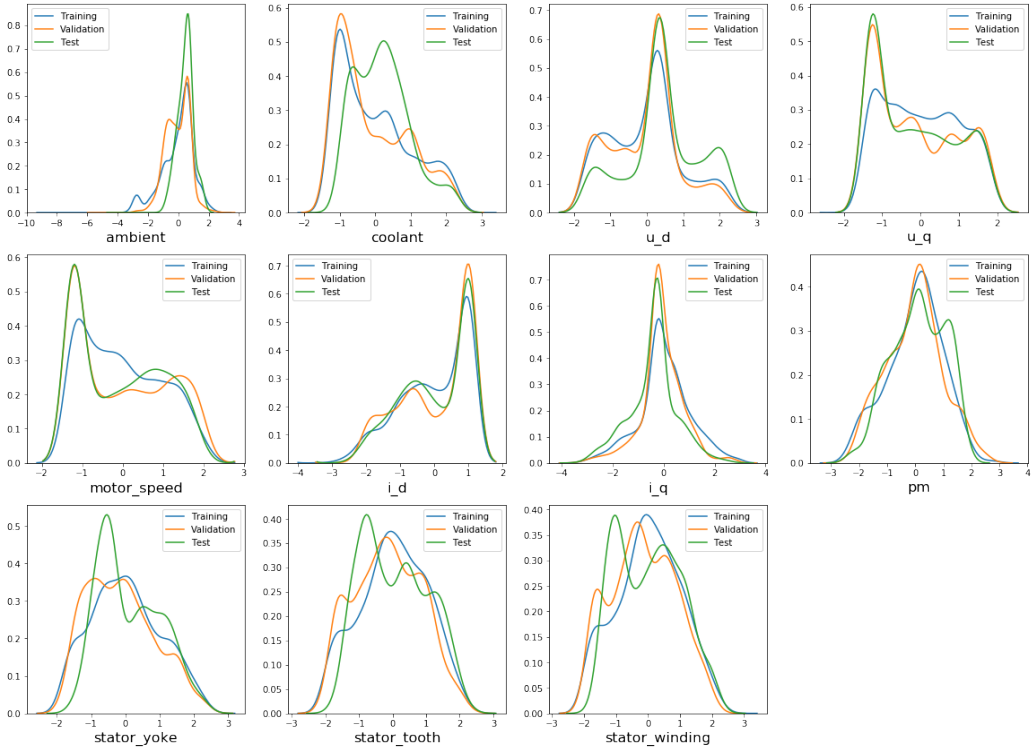

Figure 1: Distribution of the variables grouped by the division

Each model is optimized using Adam optimizer, using standard *Mean Squared Error* (MSE) as loss function in a first trial, and then re-optimized using a custom weighted-version that assigns a triple importance to temperatures that surpass the value of the median (computed on the train set).

3.1 Dataloaders

To avoid giving all input and output matrix to the model , two ad-hoc dataloaders are programmed to optimize the learning process. Belonging to various independent series, data is represented internally as a sequence of tuples $< \mathrm{id}_{serie}, \mathrm{id}_{observation} >$ so that can be easily shuffled and yielded to the model (without side-effects) using an arbitrary-large batch size.

First Task To accomplish the first task, data is given to the algorithm so that input matrix contains all values of all features and target variables

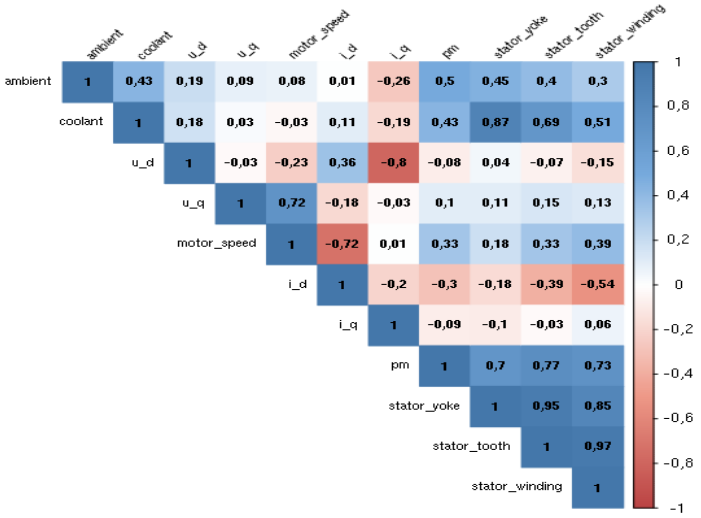

Figure 2: Correlation Plot of the considered variables

throughout the temporal window (starting from an arbitrary moment in the past t_{-max} (the maximum lag) to the previous observation t_{-1}), while target vector contains following values of only target variables at time t_0.

Second Task For the second task instead, dataloaders yield a input matrix containing only features values for a window time starting in a moment t_{-max} to t_0; in addition a vector of values of target variables at time t_0 is given as ground truth: it's a real time prediction of the target and in this case there is a unique model which predicts all four targets.

3.2 Models

All models are implemented as Sequence to Value (Seq2Val) so that the training process can be paralleled and shuffled, with improvements in both speed and quality. In addition, a Sequence to Value model can be easily rewritten as Sequence to Sequence (Seq2Seq), providing real-time information to the driver.

RNN In order to accomplish both tasks with a lightweight model easily usable by a car, a simple Recurrent Neural Network is implemented: this model is provided with one hidden layer that depends on both current data

and previous observations, and uses it to estimate the target variables. After the hidden layer, data flows through two independent feed-forward neural networks (with two layers both) to estimate values for `pm` and `stator`s variables: this approach is justified by the higher correlation between `stator` temperatures and lower correlation with `pm`.

Which layer (the previous input, the estimated output or the embedding) the RNN yields to the next step and how to use this information (where it is linked) are parameters settable in the model construction and optimized through Auto-ML; in addition, *learning rate*, number of neurons per hidden layer, and length of the input sequence are optimized in the same way: to do so, the `Custom_RNN.forward()` method is implemented recursively to make it as adaptable as possible.

LSTM and GRU Two more type of recursive neural architecture are tested, the first one uses a LSTM layer while the other uses a GRU layer instead. For the first task two independent models are created: one to predict only the `pm` variable and one to predict the 3 `stator` variables; time window size is fixed to 60 lags: it does not need to be optimized since the LSTM and GRU should automatically "understand" how many lags are important.

The model is pretty simple (as required by the task): it has one recurrent layer (either LSTM or GRU) followed by two fully connected layers, of which the second one can be omitted if the number of neurons is 0; lastly there is the output layer which has one neuron while predicting `pm` and three neurons while predicting `stator`. During the second task all the four targets are predicted together.

The hyperparameter optimizer needs to optimize the number of hidden units, the number of neurons for each fully connected layer, the learning rate and the batch size of the data. The objective score is the MSE and it is calculated, after one epoch of training, on the validation set.

CNN A CNN model is also tested, even though the number of parameters increases compared to the other approach described before.

For the first task, the model has 2 Convolutional layers, each one followed by a Max Pooling layer; the last convolution-pooling block is flattenized and followed by 2 Fully Connected layer and then the output layer. The second Convolutional and Fully Connected layers can be omitted if the number of

filters or neurons are 0. For the second task the architecture is simplified: poolings are removed.

The convolution happens on the vectors that contain all the variables for each specific time in $[t_{-max}, t_{-1}]$. Basically this model is a lighter version of a FC model (it has less number of weights) and it extracts the most relevant features to pass to the FC layers. The temporal window is the same used by the LSTM model.

In this architecture the hyperparameters to optimize are the stride and kernel of the convolutions, the number of neurons for each FC layer, the learning rate and the batch size.

4 Results and Evaluation

4.1 First Task

In the Figure 3 results of hyper-parameter optimization are shown: for both models one predicting the `pm` variable and the other predicting the three `stator` variables. RNN is excluded because of it's poor performance (loss is two order of magnitude bigger) when compared to the other three models.

Results of the model on Training, Validation and Test are reported on the Table 2, in this case all the output variables are concatenated together and a general loss is calculated.

	MSE			Weighted MSE		
Models	Train	Validation	Test	Train	Validation	Test
RNN	1.82e-3	2.17e-3	4.46e-3	4.21e-3	4.18e-3	2.55e-3
CNN	2.32e-5	2.28e-5	2.39e-5	1.78e-5	1.80e-5	1.98e-5
GRU	1.29e-5	1.24e-5	1.28e-5	2.42e-5	2.18e-5	2.48e-5
LSTM	0.98e-5	0.96e-5	1.00e-5	2.20e-5	2.00e-5	2.23e-5

Table 1: Results on Training, Validation and Test Set for the first task. Values can change (in a small range) due to stochasticity of algorithms.

In addition, to improve predictions of critical temperatures, a custom weighted version of MSE is used as described in the previous section. Results of hyperparameter optimization with the new loss are shown in the Figure 4.

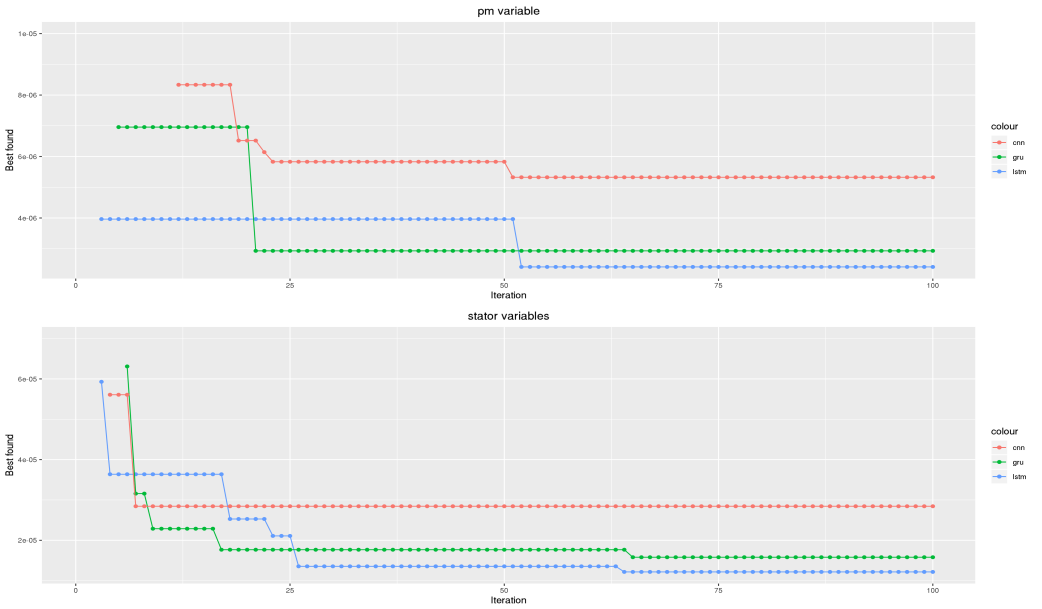

Figure 3: Results of the optimization process on the pm and the three stator variables respectively using the MSE for the first task (RNN is excluded due to its poor performance). Values outside the specific range are excluded.

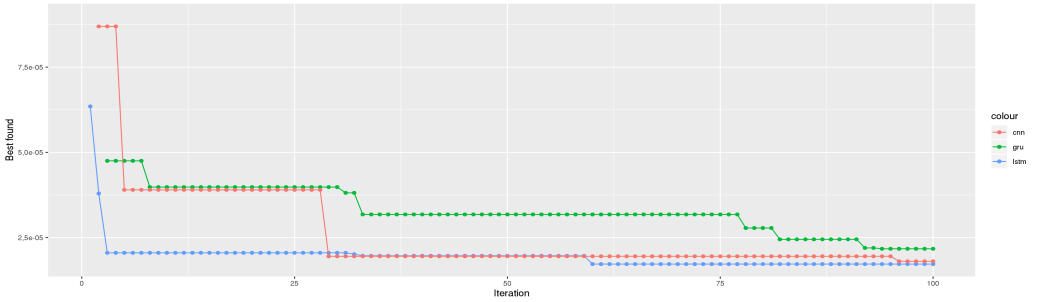

Figure 4: Results of the optimization process using a custom loss for the first task (RNN is excluded due to its poor performance). Values outside the specific range are excluded.

4.2 Second Task

For the secondary task (predicting motor temperature without knowing previous values) results of hyperparameters optimization are shown in the figure 5:

	MSE			Weighted MSE		
Models	Train	Validation	Test	Train	Validation	Test
RNN	0.0118	0.0127	0.0232	0.0131	0.0150	0.0254
CNN	0.0121	0.0122	0.0126	0.0140	0.0142	0.0141
LSTM	0.0143	0.0139	0.0149	NA	NA	NA

Table 2: Results on Training, Validation and Test Set for the second task. Values can change (in a small range) due to stochasticity of algorithms.

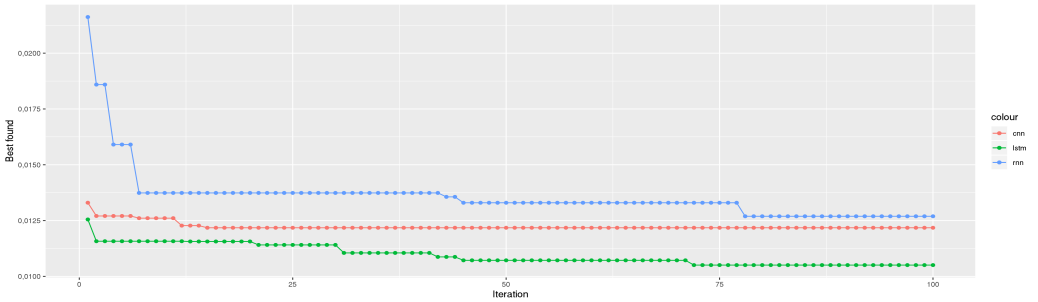

Figure 5: Result of the hyper-parameter optimization for the second task. Values outside the specific range are excluded.

5 Discussion

As shown in Figure 3 and Figure 4, after the optimization process, the LSTM and GRU yield the best results (LSTM is slightly better); CNN provides results comparable with these architecture, while RNN yields poor previsions when compared to the other architectures.

However, comparing results using only the loss value is not satisfying, because the model have to be light enough to be easily used by a car. So both number of parameters and performance are considered and plotted in the Figure

As shown in Figure 6, numbers of parameters is one order of magnitude higher using the better architecture in the first task (CNN) instead of GRU, with little impact on prediction quality. In the second task however, the most-accurate model has the minor number of parameters, so its use is not discutible.

In the second task performances are pretty good for the `stator` variables,

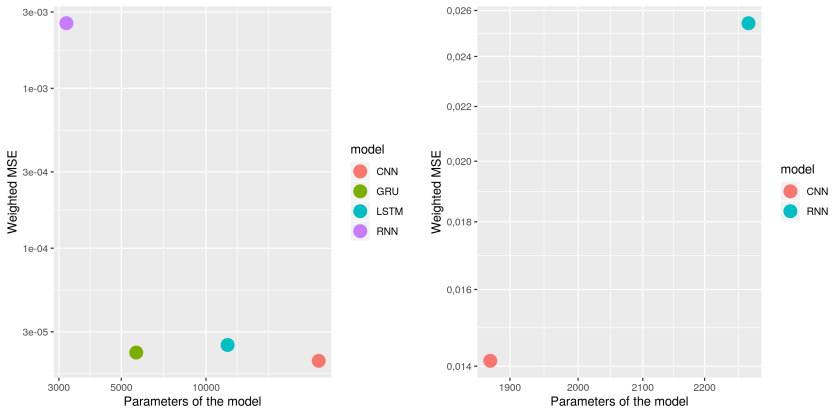

Figure 6: Weighted MSE vs Number of parameters per model on the first (left) and second (rigth) task (on a logarithmic scale).

but not so good for the `pm` variable. The best model in this task is again the LSTM. The RNN's behavior is improved too.
In the prediction of the LSTM model on the test set (Figure 7) a lot of noise can be noted in the prediction of the `pm` variable. The `stator` variable prediction seems to follow the local trend pretty well, despite the fact that there is no direct formula connecting different physical quantities.

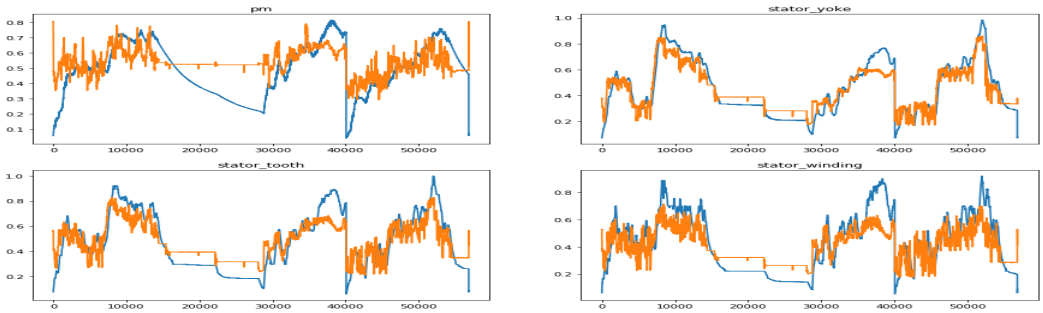

Figure 7: Prediction of the LSTM model in the second task.

The weighted loss has been applied also in this case (with the CNN model) and it improves, as expected, the prediction for higher temperature as shown in Figure 8. Prediction for lower and mid range temperature is more noisy, thus the overall MSE decreases.

Even creating a separate independent model for the `pm` variable does not improve its performance. Obtaining real data from actual cars might be

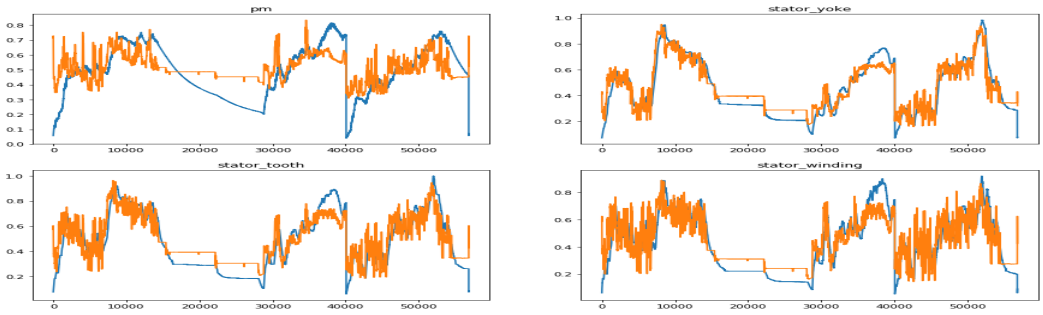

Figure 8: Prediction of the CNN model in the second task with the custom loss.

useful to improve the model further: the AC current and voltage of the 3 phase reference system is supposed to be balanced, but all real systems have some irregularities, so a dq0 transformation is just a simplification of the problem.

6 Conclusions

Recurrent models, since they are used for this kind of tasks, are expected to work well, but new trends in literature, such as [**?**] [**?**], suppose that a CNN should also work well. This Project was a confirmation of this supposition: performance are at least comparable to recurrent models, if not better.
This report, however, cannot be a proof of them working better or worse because NAS was not done for either model, only the hyper parameter were optimized with a simple pseudo NAS where the number of layers can vary in some cases.

Target variables have a normal distribution when collected from different independent profiles, but in each profile the distribution is not normal, thus the optimal predictor for a series might not be linear, and the arima ACF/PACF plot has different ARIMA coefficients and components. So, for each series a new ARIMA needs to be trained, making it unpractical to use; a deep learning model however generalizes well on all series, being a non-linear predictor. If the model needs to be simplified even more, a great option would be an exponential smoothing model for the first task and a generalized linear regression for the second, given the high correlation among some of the variables, or even combining both models in a single one.

To improve the second task's performance, recurrent models can be used in a different way: after the prefixed lag number of event have taken place, the predicted values of past can be used with the model created in the first task to "smoothen" the model's prediction.

References