ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Anomaly detections on the oil system of a turbofan engine by a neural autoencoder

Jean Coussirou, Thomas Vanaret and Jérôme Lacaille

DataLab, Safran Aircraft Engines
Rond-Point René Ravaud, 77550 Moissy-Cramayel, France

Abstract. The turbofan engine uses oil to lubricate and cool its components. This extremely sensitive system can cause in-flight engine shutdowns in the event of a failure. This article presents the implementation of a fully automatic anomaly detection system capable of detecting both known phenomena and exceptional cases using weak signals.

1 Introduction

Several algorithms developed by the PHM (Prognostics and Health Monitoring) team monitor the LEAP oil system. Most of these algorithms are based on a business approach by building indicators specific to each phenomenon to be monitored. They use snapshots type data (snapshots of measurements taken during several flight phases). The oil system may however present other types of operating anomalies not covered by existing algorithms, and which will not be detected because they are still unknown to the engineers.
Since recently, we have access to continuous CEOD (Continuous Engine Operating Data) data collected during flights, which offers much richer content. It is probably possible with the help of this new functional type of information to identify original and unusual events.
The proposed application is a generic algorithm (independent of the platform, even if it is only tested for the LEAP engine for the moment) capable of identifying weak signals. The tool is therefore not built for a specific anomaly, but on the contrary, it is able to detect any problem in the oil system before too much deterioration compromises the normal operation of the engine.
Auto-encoders are increasingly used to detect local anomalies on time signals, for example acoustic data [1], or on electric actuators [2] which, unlike the heavier hydromechanical versions, can undergo a seizure phenomenon which must be controlled by analysis. These neural networks are often linked to classification models [3], [4] rather than just anomaly detection because they allow to reduce the dimension of signals and process them as a whole.

2 Collected data

CEOD data is made up of hundreds of measurements and calculations performed by the on-board computers and retrieved on the ground after each flight. To address the oil management system, we are only interested in 10 parameters, including 4 oil data parameters (quantity, temperature, main pressure, and pressure difference around the

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

filter) and 6 context data parameters (fan and core speed, fuel flow, external pressure and temperature, position of the fuel/oil temperature exchange system valve). All the corresponding time series have a frequency of 1Hz. To calibrate and validate the algorithms, we also use maintenance feedback information as well as on-board indicators developed by our PHM team.

3 Methodology

The objective is to detect abnormal behaviors on the 4 oil parameters without any prior knowledge in a given flight. This can occur on any time scale, at any time location and be potentially multivariate. To tackle this problem, the developed algorithm uses two fully convolutional auto-encoders. Indeed, as oil system incidents are scarce events, the available database of recorded flights is mainly composed of signals with no anomalies. To take advantage of that, the auto-encoders are trained to encode and decode the nominal multivariate temporal patterns encountered during a common flight. Thus, when the models are fed with such an input, a good reconstruction is expected and on the other hand, a high reconstruction error would reveal the existence of a problem.

The length of the time series motivates the use of convolutional layers: the mean duration of a flight is about 8000 seconds, which is beyond the memory capability of recurrent neural networks even based on GRU or LSTM cells, and convolutional layers are far easier to train on such long sequences. However, in order to detect an anomaly occurring on a certain time span, the top layer of a convolutional block (i.e. stack of multiple convolutional layers) encoding or decoding an input signal, must have a receptive field with respect to the input, of more or less the same temporal order of magnitude. This is especially true for mild and moderate abnormalities on signal dynamics developing slowly over time in which all parameters stay in their nominal range. Therefore, the use of two auto-encoders rather than one is preferable: to achieve a receptive field of thousands of seconds would require stacking dozens of layers resulting in a network with a great depth and number of parameters, difficult to train with a probable poor performance in the end. Skip connections commonly employed in this context are unusable as the objective is to reconstruct the input: the model would exploit these trivial pathways without learning anytime.

The first auto-encoder, which will be called high frequency (HF), is based on filters exploiting the initial data at 1 Hz, and the second, low frequency (LF), will use a sub-sampling at 0.01 Hz. The subsampling consists in dividing the 1Hz data in contiguous non-overlapping windows of length 100 time steps and taking the respective mean for each 10 parameters; if the last window is incomplete, the corresponding data is discarded. This subsampling procedure allows preserving the shape of the signals. Hence, in the rest of the paper the term flight refers to an entire multivariate time series of dimension 10 with a time resolution of 1Hz or 0.01Hz (obtained using the subsampling method described) from engine start to engine shut down.

The low-frequency part is intended to detect large-scale unusual phenomena taking place over a long period, while the first auto-encoder will focus more on fast or almost instantaneous events. Figure 1 shows the detailed architecture of the models, the main difference being the use of stacked dilated convolutions to increase the receptive field of the blocks on the LF model. For each model, the output layer is a 1D convolution with 10 filters, a kernel size of 3 with padding and no activation.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

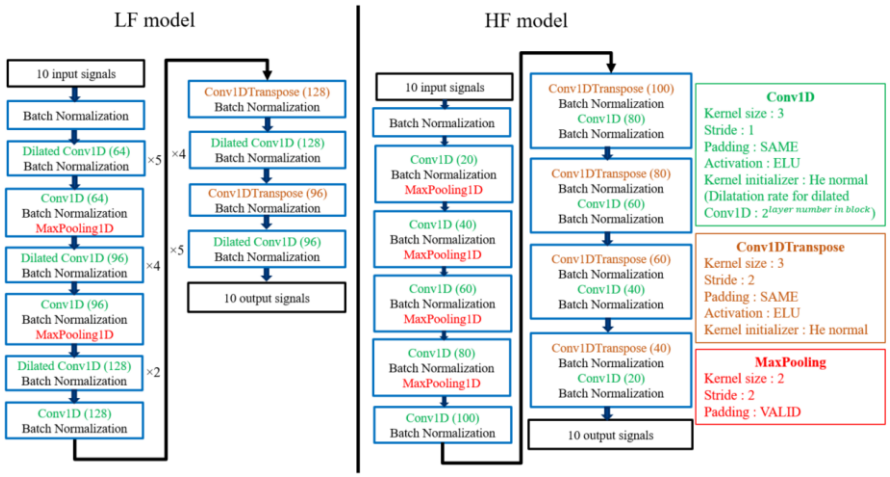

Figure 1. Architecture of the auto-encoders (LF model on the left and HF model on the right). The multiplication symbol and the parenthesis refer respectively to the number of dilated convolution layers stacked in the block and the number of filters

The models are fully convolutional which makes it possible to handle easily flights with any duration: the HF and LF models require input multiple of 16 and 4 respectively to output sequences of the same length. The input time series are padded with zeros to reach these requirements, after the subsampling step for the LF model, and the corresponding time steps are then ignored in their respective output.

4 Training procedure

The available database is divided in three datasets. The train dataset composed of 19567 flights spread over 51 engines, is for the batch gradient descent algorithm. The validation data composed of 4909 flights spread over 25 engines, different from those belonging to the training set, is used for two purposes: initially to monitor any overfitting during the training phase and to compute the hyper-parameters of the anomaly score described in Section 5. The flights corresponding to these two data sets are selected taking care to extract the abnormal flights already identified, knowing that there is still a small proportion with anomalies, and with a large geographic and mission diversity. The test set is introduced in Section 6 and used to assess the models performance. The datasets for the LF model are the same as the HF model ones except for the subsampling pre-processing. Every sequences at training and inference time are standardized based on the 2 metrics, global mean and standard deviation of each 10 parameters, computed beforehand from all flights of their respective training dataset.
The auto-encoders are trained to reconstitute the input identically, using a reconstruction error based on least squares.
In the case of the HF model, for the training phase only, to avoid too much padding caused by the different durations, each flight is divided in windows of 2000 seconds (which is a multiple of 16) with an overlap of 50%. If the length of the last window is lower, it is padded with zeros. Thereby, the HF model training instances consist in

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

chunks of flights of 2000 seconds. As far as the LF model is concerned, as the sequences length is one hundred times lower and the purpose of this model is to detect large-scale anomalies no slicing is performed. Entire flights with similar duration are gathered in 4 buckets of amplitude 80 time steps (multiple of 4). The first bucket is filled with sequences whose length varies in the range [0, 80] and [240, 320] for the last one. Zero padding is applied on a flight inside a bucket if necessary to match the bucket upper bound. During a gradient training step, a bucket is randomly chosen and the batch is composed of flights contained in it. For both HF and LF models, ADAM optimizer is used with respective batch size of 64 and 32, along with the 1cycle learning rate policy [5]. The learning rate increases linearly from $1.\,10^{-4}$ to 1.10^{-3} (from 5.10^{-5} to 5.10^{-4} respectively) during the first half of the training then decreases linearly down to the initial value during the second half except in the last 10% epochs in which it drops linearly by 1000 orders of magnitude. The inertia of the gradient follows the inverse dynamic from 0.95 to a minimum of 0.85 and stays constant to 0.95 in the last 10% epochs. 100 epochs were sufficient to obtain a satisfactory reconstruction and no early stopping procedure was necessary.

5 Anomaly scores

The models are applied on the ground once the flight records have been retrieved and stored in a dedicated database. At inference time, each entire flights are given as input to the models to obtain the reconstructions. The observation of notable differences between the initial signal and its reconstruction therefore gives a good indication of an anomaly. This detection is made both on the reconstruction at 1 Hz and that at 0.01 Hz: two flight scores will therefore be calculated. Detection thresholds are defined for each score, and an anomaly will only be announced if at least one of the two scores exceeds its threshold.

Each score S uses only the 4 observations related to the oil system, the 6 other context variables are not taken into account. At each instant of a flight, a Mahalanobis deviation is calculated from the vector formed of the 4 reconstruction residues whose approximate normality is checked. The correlation matrix (different for each model) is calculated beforehand once and for all on the whole validation data, supposedly composed of flights without problems on the oil system, after stacking all the residual vectors in the time dimension. This allows calibrating the nominal reconstruction errors of the models on signals with no issues.

$$\begin{aligned} r_t^j &= \hat{x}_t^j - x_t^j, j \in \{1 \dots 4\} \\ \Sigma &= cov(r) \\ s_t &= r'\Sigma^{-1}r \\ S &= \sum_{t, s_t \geq \theta} s_t \quad \text{with an alert if } S \geq N \end{aligned} \tag{1}$$

To reduce data inconsistency, we will never consider the 3-minute measurements at the start and end of each flight: the corresponding time steps are ignored in the computation of S. During these periods, the combined effect of starting or stopping the engine

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

generates a side effect which degrades the precision of the algorithm. Moreover, if a flight duration is lower than 3000 seconds, the LF model is deactivated as the subsampling procedure suppresses too many information from the signals resulting in high false alarm rate on those sequences.
The detection thresholds θ and N (different for each model) were selected empirically according to the observed results on the validation data in order to calibrate the ratio of false and true positive on that set. They are chosen once and for all and will be used to compute the score of any flight of any engine. θ corresponds to a quantile of the distribution of all s_t values of all flights of the validation set. The time series s_t which can be seen as a normalization and aggregation of the residues of the 4 oil parameters helps to identify precisely the locations of the abnormalities in the entire flight which is an important information to understand the cause and provide a solution if applicable. Moreover, no correlations between the flight scores and durations have been observed: it can be accounted for the fact that only very bad reconstructions, far beyond the usual expected model errors on nominal data, at specific time steps contribute to it. It makes it possible to follow and compare the scores for a given engine to spot potential trends.

6 Experiment and results

6.1 Tests

To evaluate our solution, a beta-test was carried out early 2022 with the latest flights data available (~5000 flights per week). We collected data from 571 LEAP-1A engines in operation on Airbus A320s of which 167 presented anomalies detected by the algorithm. For the most part, these are standard events such as the detection of the beginnings of clogging of a filter or the overfilling of the oil tank and which have no consequence on the operability of the engines.
More precisely, our database contains 61737 flights from those 571 engines of which 871 flights (1.4%) show anomalies for 167 engines (29.2%). Among those 871 detected, 560 where entirely analyzed, 287 were false alarms due to poor data feedback, missing data, or discontinuities during flights. Among the valid cases, 199 are sensor problems (managed by redundancy). 26 cases (concerning 9 engines) showed anomalies that had not been detected by existing PHM algorithm for the LEAP-1A.
At this step, the algorithm shows a false positive rate of around 0.5%, i.e. 25 flights per week and an alert rate of 0.4 %, about 20 flights per week. Further work is planned to reduce the false positive rate, mainly caused by data quality issues.

6.2 Example of anomaly detection

Figure 2 shows the detection of a specific oil system anomaly. Under some degraded conditions, non-magnetic particles can temporarily obstruct the oil circuit and cause sudden variations in the tank level. This is what can be seen in graph D at the bottom right, which shows the oil level measurements, observed on the first flight after detection (graph A at the top left). We can see just to its left (graph C) that the oil temperature is perfectly restored (it would be the same for the pressures). Graph B at the top right shows the evolution of the score calculated from the high definition auto-encoder.

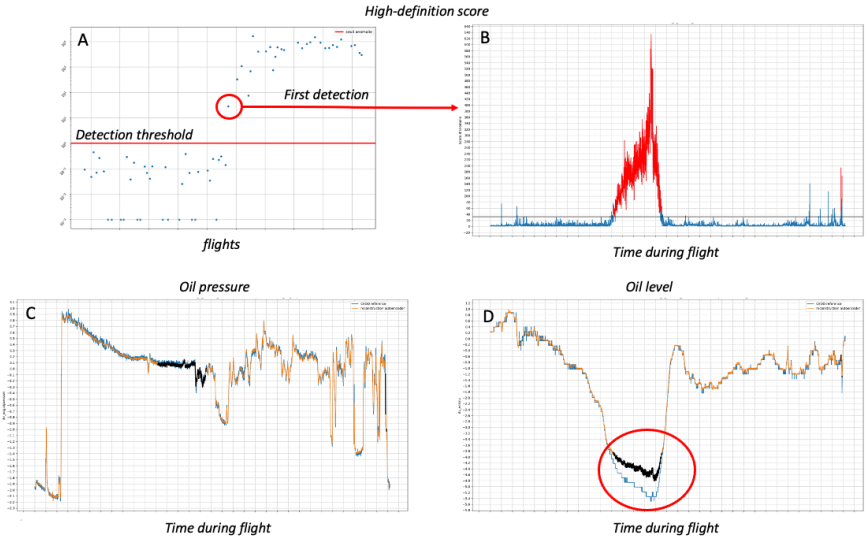

Figure 2: Anomaly detected on an engine that was not part of the calibration and test dataset. A is the high-definition score over successive flights, B is the local high-definition score computed during the first flight after detection, C is the expected (orange) and observed (blue) oil pressure and D shows the expected (orange) and observed (blue) oil level with a localized drop which is the signature of the non-magnetic particles detection.

7 Conclusion

This very promising algorithm is still being studied with a view to integrating it on our operational monitoring system. It has already been transferred to the LEAP-1B engine, which equips the various versions of the Boeing 737. Other engine systems are also being considered for similar applications.

References

[1] T. B. Duman, B. Bayram, and G. İnce, "Acoustic Anomaly Detection Using Convolutional Autoencoders in Industrial Processes," *Advances in Intelligent Systems and Computing*, vol. 950, no. January, pp. 432–442, 2020, doi: 10.1007/978-3-030-20055-8_41.

[2] A. Eid, G. Clerc, and B. Mansouri, "A Novel Deep Soft Clustering for Unsupervised Univariate Times Series," 2021. doi: 10.1109/ICPHM51084.2021.9486468.

[3] F. Forest, M. Lebbah, H. Azzag, and J. Lacaille, "Deep embedded self-organizing maps for joint representation learning and topology-preserving clustering," *Neural Computing and Applications (NCAA)*, no. 33, pp. 17439–17469, 2021, doi: 10.1007/s00521-021-06331-w.

[4] F. Forest, M. Lebbah, H. Azzag, and J. Lacaille, "Deep Embedded SOM: Joint Representation Learning and Self-Organization," 2019. doi: 10.1016/B978-0-323-39396-6.00001-4.

[5] Leslie N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay" 2018, https://doi.org/10.48550/arXiv.1803.09820