# UNSUPERVISED ANOMALY DETECTION USING VARIATIONAL AUTOENCODERS - ASSESSING ROAD CONDITIONS FOR LIRA PROJECT

*Jonas Søbro Christophersen (s153232) & Lau Johansson (s164512)*

## ABSTRACT

In this paper we construct a variational autoencoder with attention mechanism using a Recurrent Neural Network (RNN) with long short-term memory cells (LSTM) in the encoder and decoder networks. We propose to use the model for unsupervised classification of anomalies in sensor data collected from cars related to the LiRA[1] project [1]. The model is trained on simulated sensor-data. It is shown that when passing anomalous road defects to the model it fails to reconstruct them, whereas normal series are properly reconstructed. This lack of reconstruction of anomalies is the basis for a classification model of road defects in further work. Principal component analysis of the latent representations shows that the model might be subject to the bypassing phenomenon. This is alleviated to some extent by increasing the regularization of the self-attention network, furthering the emphasis on the latent representation in reconstructing the input series. The model is easily extendable to additional road-sensors once real data becomes available, and the model seems promising in assessing road conditions in relation to the LiRA project.

## 1. INTRODUCTION

This project serves as preliminary research for the LiRA project conducted in collaboration with Danish Technical University. By analyzing sensor data collected by Green Mobility cars, road conditions can be assessed and road wears can be identified preemptively, enabling roads to be maintained more efficiently [1]. The use of machine learning related to anomaly detection is challenged by the amount of available anomalies in training data and the labelling of these. The LiRA project faces these challenges too. A generic and unsupervised framework for detection of anomalies in time series data is proposed in this paper, inspired by works of J. Pereia and M. Silveira [2]. We propose in this paper how variational autoencoders with self-attention mechanisms and recurrent neural networks utilizing LSTM-cells as encoders and decoders are able to detect anomalies in time-series.

Utilizing the insights of this paper project stakeholders can plan how to approach the real data once available. The model is built using PyTorch and can be applied to detect road defect anomalies without any predefined labels on training or test data. The data used for this project is simulated data which resembles the real data which is yet to be gathered.

## 2. MODEL ARCHITECTURE

Motivations behind the choices of the model architecture are outlined below. Code is available at [13].

### 2.1 RNN and LSTM

As the data collected from the Green Mobility cars are time series they should be treated as such. An effective way of capturing time dependencies in neural networks is achieved by using recurrent neural networks (RNN). The RNN introduces memory by sequentially processing time-steps, carrying information through the time series by inclusion of a hidden state vector $\mathbf{h}_t$. The hidden state at timestep $t$ depends on the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$. RNNs can suffer from the vanishing gradient problem, in which the hidden states attribute a higher importance to more recent timesteps - thus causing vanishing importance to values at much previous timesteps. Long-Short Term Memory network (LSTM) is a variation of RNN which overcomes the problem by adding four interacting terms to the RNN: A cell state and three gates. In the LSTM network information is processed through input, output and forget gates which enables the network to attribute weight to all time-steps if they are relevant for prediction. Thus, the network is less prone to forgetting long term dependencies. The hidden states are used for predicting the next timesteps and the long-term memory is held in a cell state [5][6]. Conventional LSTM networks are not capable of utilizing information from future time-steps, however with the introduction of Bi-LSTM proposed in [8], LSTM networks are constructed to iterate both frontwards and backwards through the input sequence when processing the input, enabling inclusion of information from future time steps to the predictions. In this project we choose to use Bi-LSTM as it is not deemed necessary to forecast sequence-progressions (when no future information is available) but rather analyze entire sequences and identify if these are anomalous.

### 2.2 VAE

Autoencoder neural networks are a type of networks which aim to reconstruct input data from a lower dimensionality representation. By lowering the dimension of input data using

1 Abbreviation for Live Road Assessment

an encoder network, a latent space representation (denoted **z**) of the data is obtained, which then can be reconstructed using a decoder network. Traditionally, autoencoders were used for dimensionality reduction or feature learning but have since gained a lot of recognition for their uses in unsupervised anomaly detection [3]. Variational autoencoders store the latent space representation as a probability distribution (which is sampled) rather than a fixed value using the reparameterization trick [11]. By storing the latent space representation as a distribution rather than a fixed value, the network can be used for generative processes, and also serves as a regularization of the latent space. This regularization is caused by the inclusion of the Kullback-Leibler (KL) divergence between the prior ($p_\theta(\boldsymbol{z})$) and posterior ($\widetilde{q_\phi}(\boldsymbol{z}|\boldsymbol{x}^{(n)})$) of the latent space representation in the loss function of the network [4]. The chosen prior distribution is a standard normal i.e. Normal(**0**,**I**). The latent space is generated using two linear layers, with tanh activation for the first moment of the latent normal distribution, and softplus activation for the second moment of the latent normal distribution. The input for these two linear layers is the last hidden states in both directions from the encoder LSTM network. To avoid negative values for the second moment, the log-variance is used which is then transformed prior to sampling.

Thus when passing many normal (well-known) time series of road conditions the network learns to reconstruct these well, however when the network is presented to anomalous samples, the network is not capable of identifying a proper latent representation of the sample - as these anomalies were never part of network training, and ultimately achieves a high reconstruction error for these anomalies. This error can then be used to identify whether the sample was normal or anomalous. The chosen error-measure for these reconstructions are Mean Squared Error (MSE), but the network may benefit from changing these to reconstruction probabilities as seen in [2] which is planned to be investigated in further work.

### 2.3 Attention layer

As presented in [7], self-attention mechanisms encourage the network to focus its attention on whichever relevant parts of the input-sequence are useful for current predictions. The mechanism is essentially imitations of how us humans perceive information through sight, in which typically instead of scanning entire scenes, we choose to focus on specific portions according to current needs [9].

We include a self-attention mechanism to improve performance of reconstructions. Self-attention has been gaining popularity in NLP seq2seq models, but has also shown useful for fields outside of NLP [2]. Since the latent representation is reliant on only the last state of the hidden states in the LSTM network, the representation can't sufficiently store all relevant information in the sequence, and thus inclusion of selective attention to relevant parts of the sequence enhances reconstruction performance. The attention layer receives as input encoded hidden states for all timesteps and computes deterministic context vectors, (denoted $\mathbf{c}_{det\_t}$) for all timesteps. The computation of these are identical to those presented in [2]. These context vectors are connected to a variational layer with a similar approach as for the latent representation, and thus context vectors (denoted $\mathbf{c}_t$) are obtained by sampling from the first and second moment nodes connecting to the deterministic context vectors. This is done to circumvent the *bypassing phenomenon* presented in [10], in which the latent space representation is not learning meaningful representations, as the required information for the decoder is primarily carried in the attention layer. By removing the direct connection between the attention layer and the decoder and regularizing the context vector distribution using the KL divergence between the posterior ($\widetilde{q_\phi}^a(\boldsymbol{c_t}|\boldsymbol{x}^{(n)})$) and prior distribution ($p(\boldsymbol{c_t})$), the network is compelled to learn meaningful latent space representations. Similarly to the variational layer for the latent representation, the prior for the context vectors is a standard normal and the KL divergence between the prior and posterior distribution for the context vector is included in the loss function.

### 2.4 Data

The model is built using simulated data which represent three different real-world road defects: cracks, patches and potholes. The model uses simulated time series sequences with a length of 150 timesteps. Each simulated sequence resembles a few seconds of data gathered while driving over one of the three possible defects. A sequence contains two features, with the first feature corresponding to sensor data from an accelerometer and the second feature capturing the severity of the road defect. As proposed in [2] the data is corrupted by adding gaussian noise $\mathbf{n} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\sigma}_{2n}\mathbf{I})$ to the input data **x.** Additionally the data is standardized. This is done to ensure that the network puts equal emphasis on both features in reconstruction, since the scale and variance of the two features are different.

### 2.5 Loss function

The training objective (for each sequence) is to minimize the loss function seen below:

$$L = ||\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{x}}^{(n)}||^2 + \lambda_{KL}[D_{KL}(\widetilde{q_\phi}(\boldsymbol{z}|\boldsymbol{x}^{(n)})||p_\theta(\boldsymbol{z})) + \eta \sum_{t=1}^{T} D_{KL}\left(\widetilde{q_\phi}^a(\boldsymbol{c}_t|\boldsymbol{x}^{(n)})||p(\boldsymbol{c_t})\right)]$$

The reconstruction term is the mean squared error between the corrupted input data and the reconstructed time series. The second loss term ensures minimization of the KL divergence between posterior and the prior latent space. The third term ensures to minimize the KL divergence between posterior and the prior attention space. Optimization aims to find approximate posterior parameters $\phi$ that minimize the divergences. An annealing scheme on $\lambda_{KL}$ is applied, which

increases the parameter for each epoch. The parameter is initially set to 0 and thereby focusing on the reconstruction term in early stages of training. This is done to avoid the KL vanishing problem [12]. The hyperparameter η balances the latent representation KL term and the attention KL term.

### 2.6 Model overview

A quick overview of the network architecture can be seen below.

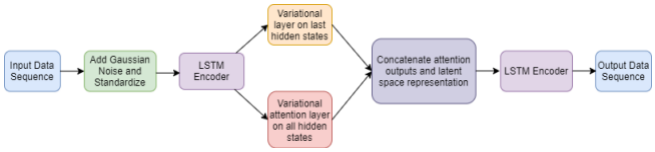


Figure 1: Network architecture

As the network is inspired by the one presented in [2], for further details investigate Fig. 1 in [2] - with the exception that the model proposed in this paper does not return reconstruction probabilities, but rather the reconstruction values. This addition is to be investigated in further work.

## 3. BUILDING THE MODEL ON SIMULATED DATA

### 3.1 Tuned hyperparameters

To achieve the best performing model the number of layers in the encoder and decoder LSTM are set to one. The chosen LSTM hidden unit size for both the encoder and decoder is 128. The batch size in training and testing is 5. The dimension of the attention space is 2 and the latent representation space dimension is 20. The chosen hyperparameters for the $\lambda_{KL}$ annealing scheme are 0 for the initial value and update-rate of 0.1 and maximum value 1. The balancing parameter η is 0.01. In the testing phase L (the number of samples from each **z**) is set to 20. The chosen optimizer is AdamW with learning rate 0.01.

### 3.2 Tested models

To improve the generalization of the model dropout was implemented in the LSTM encoder and decoder - however this did not increase performance and was thus left out. Additionally, several LSTM layer sizes were tested, but as no increase in performance was gained the number of layers were set to 1. Further, we tried a model with no KL divergence on the context vectors, although this caused the model to reconstruct anomalies extremely well, indicating presence of the bypassing phenomenon. A reconstruction as such is illustrated in the appendix [14].

## 4. RESULTS

To test the performance of the model both normal time series and anomalies have been passed to the model. Anomalies are generated manually by adding different kinds of noise to a normal series. Plotting the 20 samples of the reconstructed outputs (orange) against the original input (blue) provides a visual and intuitive insight into whether the model can reconstruct the input.

### 4.1 Model performances

The model was trained using 5000 training series and 1000 test series over 10 epochs. In Figure 2 the reconstructions (orange) and input time series (blue) are plotted for a normal pothole time series - the reconstructions are visualized as 20 series, since the sampling was done 20 times from the latent space representation. The MSE loss of the normal data on the plot is 6.56. When inspecting Figure 2 both the acceleration and the severity have been nicely reconstructed.

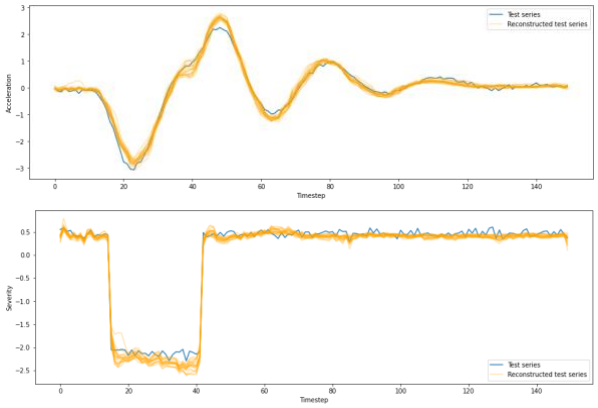


Figure 2: Reconstruction of normal pothole data

Figure 3 visualizes two different anomalies for the acceleration feature, which shows how difficult the model finds it to reconstruct them.

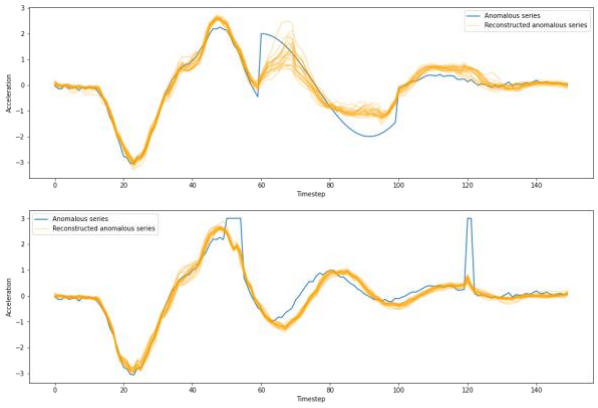


Figure 3: Reconstruction of pothole acceleration. Anomaly occurs between 60th to 100th timesteps (top). Anomaly occurs at the spikes near 50th timestep and 120th timestep (bottom).

The reconstruction errors of the two acceleration anomalies are respectively 82.45 and 12.99. The reconstruction of the anomalies has much higher reconstruction error than reconstruction of the normal data.

Figure 4 visualizes two different anomalies for the severity feature, which shows how difficult the model finds it to reconstruct them.

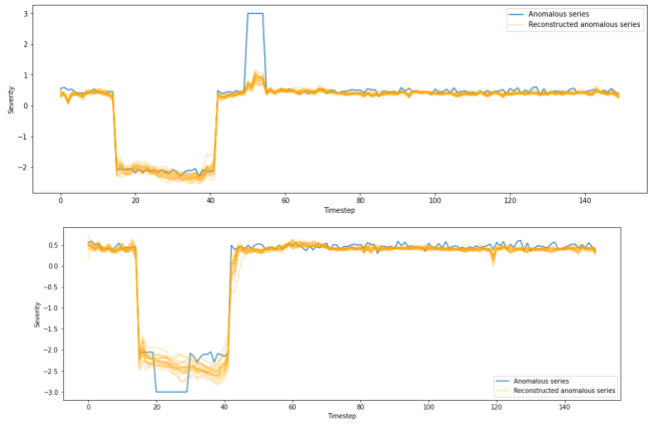


Figure 4: Reconstruction of pothole severity anomalies.

The reconstruction error of the three severity anomalies are 7.55 and 9.40. As in the example with acceleration, the severity anomalies do also cause the reconstruction error to be higher than the normal road defect severity.

### 4.2 Visualization of the latent space

A principal component analysis (PCA) has been applied on the mean latent space representations, $\boldsymbol{\mu}_z$, to investigate the structure of the latent space representations for different values of the latent space dimensions. The latent states are projected down to two dimensions to create a scatter plot of the latent states. Below is a scatter plot of the latent spaces.

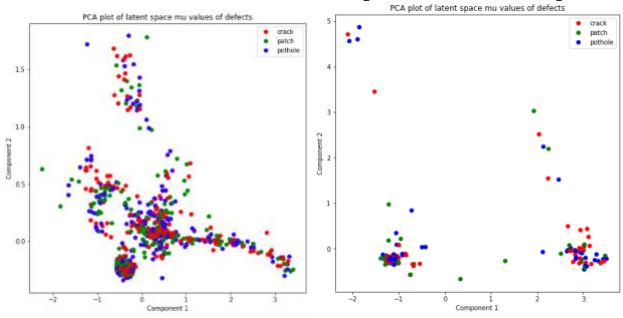


Figure 5: PCA plot of $\boldsymbol{\mu}_z$ (left) and PCA plot of $\boldsymbol{\mu}_z$ with increased η (right).

On the best performing model with η equal to 0.01 (left plot), no obvious clusters appear. Most of the data clusters around zero with no certain structure to the different classes in the test data (although there is some structure to the latent spaces but not obviously related to the classes). This could be an indication that the model is still subject to the *bypassing phenomenon* and would need further tuning. It is also possible that the structure of the different classes is not dis-similar enough to distinguish on the latent space representation. In order to test this a model using a larger η is built. This model is both trained and tested on fewer data points as training and testing is time-consuming. When increasing η there is a tendency that the defect-categories cluster together to some extent, although still with a large overlap. In figure 5 (right) the patches (green) generally cluster left in the plot, i.e. with low component 1 values, while cracks (red) and potholes (blue) have a tendency to cluster in the bottom right, i.e. with high component 1 values. The overlap between cracks and potholes could reflect the severity structure which are similar compared to patches.

## 5. FURTHER WORK

The next step for application of the model is to train on real road data once available.

In [2] it is proposed to use the reconstruction probability rather than the reconstruction error in the loss function. Using reconstruction probability instead of the reconstruction error as done in this project could accommodate for the disadvantage of having data-specific detection thresholds. Additionally, the reconstruction probability takes the variability of the reconstruction into account, which is likely higher on anomalous reconstructions. This idea is to be tested moving forward.

The model could be extended by implementing a new model which is able to classify each of the three road defects. This could benefit the LiRA project in the process of labelling the data for further training and testing other models.

## 6. CONCLUSION

In conclusion we find that the proposed model, a variational LSTM-autoencoder with a self-attention mechanism, is able to identify anomalous road defect time-series. By inclusion of a self-attention mechanism we have enabled the network to attend relevant parts of the input time series, further increasing the model's ability to reconstruct normal time series while anomalous time series are poorly reconstructed. By visualizing the latent space representation (with PCA), it was shown that the model may still be subject to the *bypassing phenomenon* however when increasing the η parameter, thus increasing the regularization of the attention-part of the network, it was seen that the latent space representations were more prone to being clustered together according to the reconstructed class. The model is easily extendable to additional road-sensors, and with the inclusion of the real data and re-tuning of the hyperparameters we believe that the model is applicable in assessing road conditions related to the LiRA project.

## 7. REFERENCES


[1] Lira-project.dk. Visited on May 10th 2020: http://lira-project.dk/

[2] J. Pereira and M. Silveira, "Unsupervised Anomaly Detection in Energy Time Series Data using Variational Recurrent Autoencoders with Attention" *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, pp. 1275-1282, 2018.

[3] I. Goodfellow, Y. Bengio, A Courville, Deep Learning MIT Press, 2016.

[4] J. Rocca, Understanding Variational Autoencoders (VAEs). [Blog post]. Retrieved from https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

[5] N. Abel, How do LSTM networks solve the problem of vanishing gradients. [Blog post]. Retrieved from https://medium.com/datadriveninvestor/how-do-lstm-networks-solve-the-problem-of-vanishing-gradients-a6784971a577

[6] C. Olah. Colah's blog. Understanding LSTM networks. [Blog post]. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L Kaiser, I. Polosukhin, "Attention Is All You Need," arXiv:1706.03762, 2017.

[8] A. Graves, J. Schmidhuber "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". 2005.

[9] A. Cloud, Self-Attention Mechanisms in Natural Language Processing. [Blog post]. Retrieved from https://medium.com/@Alibaba_Cloud/self-attention-mechanisms-in-natural-language-processing-9f28315ff905

[10] H. Bahuleyan, L. Mou, O. Vechtomova, P. Poupart, "Variational Attention for Sequence-to-Sequence Models", arXiv:1712.08207, 2017.

[11] P. Sayak, "Reparameterization" trick in Variational Autoencoders [Blog post]. Retrieved from https://towardsdatascience.com/reparameterization-trick-126062cfd3c3

[12] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, S. Bengio, *"Generating Sentences from a Continuous Space"* , *CoRR*, abs/1511.06349, 2015


## 8. APPENDIX

[13] Link to github repository: https://github.com/LauJohansson/AnomalyDetection_VAE_LSTM

[14] Plot of reconstruction without the KL divergence term between the approximate posterior and prior for the context vectors. Notice that the model reconstructs the anomaly well which is an unintended behavior.