Authors: Julian Kuypers, Manuel Demetriades, Luis Riveros Dias, Marius Löwe

Deep Learning Report

Epileptic Seizure Recognition

Motivation

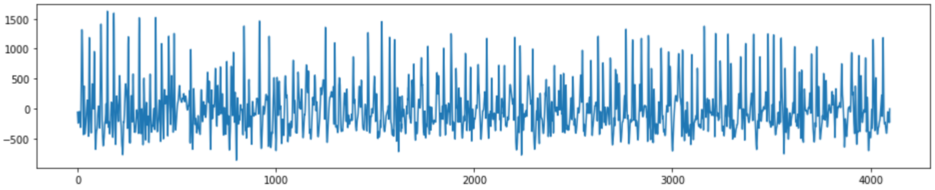
Epilepsy is the second most common brain disorder after migraine; automatic detection of epileptic seizures can considerably improve the patients' quality of life. Current Electroencephalogram (EEG)-based seizure detection systems encounter many challenges in real-life situations; EEG data are prone to numerous noise types that negatively affect the detection accuracy of epileptic seizures. To address this challenge, we propose a deep learning-based approach that learns the discriminative EEG features of epileptic seizures and to distinguish between the different types of patient recordings. More specifically, we aim to tackle this issue by using a Long Short-Term Memory network, and explore the capabilities of this model.

Data Description

The data, published on Bonn University's Epileptology department website, presents Electroencephalogram (EEG) recording of 500 individuals. For each individual, brain activity was recorded for a duration of 23.5 seconds; these recordings are represented by 4096 evenly-spaced, consecutive data points (i.e every 0.0057 seconds). Each row of the dataset, representing an individual's recording, also has a column with the classification of the recording. The five labeled datasets (A, B, C, D, E) are presented below along with their corresponding target classes:

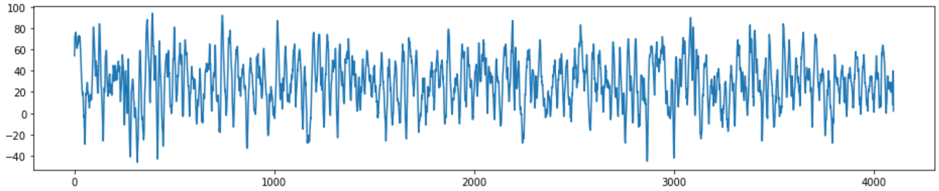
Set A - Class 4: EEG recording of a non-epileptic awake patient with eyes open.

Figure 1 - Class 4 EEG example



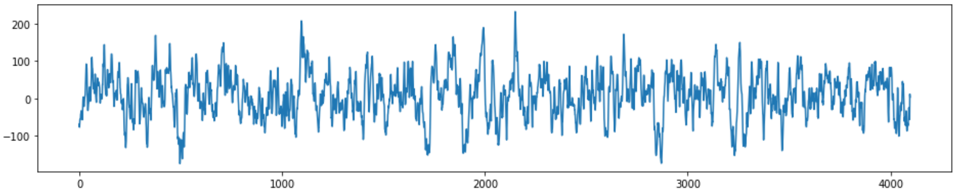
Set B - Class 3: EEG recording of a non-epileptic awake patient with eyes closed.

Figure 2 - Class 3 EEG example



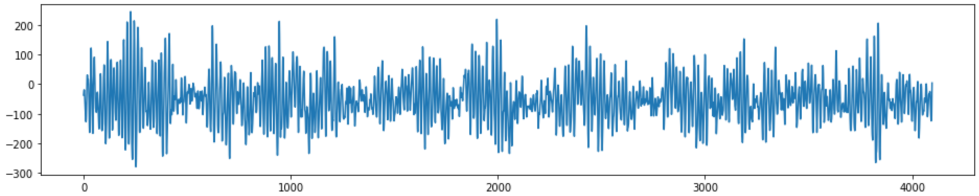
Set C - Class 2: EEG recording of an epileptic patient during seizure free period using electrodes implanted in the brain epileptogenic zone.

Figure 3 - Class 2 EEG example



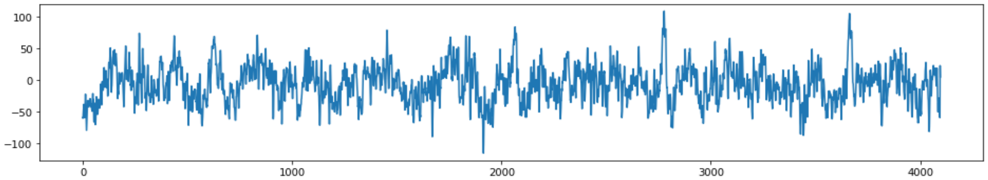
Set D - Class 1: EEG recording of an epileptic patient during seizure free period from the hippocampal formation of the opposite hemisphere of the brain from C.

Figure 4 - Class 1 EEG example



Set E - Class 0: EEG recording of a patient experiencing an active epileptic stroke.

Figure 5 - Class 0 EEG example



Our objective for this project is to correctly predict these five target classes using an LSTM Neural Network. In order to highlight the capabilities of an LSTM layer, three classification problems will be tackled where the LSTM's performance will be compared to that of a regular deep neural network:

1. A two-class problem that classifies whether a patient is having a seizure or not at the time of recording.
2. A three-class problem that requires the classifier to distinguish between a patient that is having a seizure, a patient that is between seizures (*inter-ictal*), and a healthy patient.
3. A five-class problem that aims to classify all five classes, meaning it should be able to distinguish between a patient that is having a seizure, a patient that is between seizures, and a healthy patient. Additionally, it will be able to determine in which part of the brain the recording is made (Sets C & D) and whether the patient's eyes are open or closed (Sets A & B).

Evaluation Measure

To measure the degree of success of our classifiers we have to look at the three problems separately. Especially in medical cases correct classification is of high importance (compared to marketing problems for instance), due to the direct impact on people.

For the two-class classification problem we have a very unbalanced dataset at hand. A high accuracy could still mean, that we classify those with an active seizure incorrectly, without representing this impact adequate when optimizing for accuracy. Therefore, the correct measure would be precision or recall. As it is an ethical question though, whether it is more important to classify someone having a seizure correctly, or someone not having one incorrectly, we opted to include the confusion matrix that takes both into consideration. But we display the accuracy in the other graphs to keep the results comparable. For the three-class classification problem the data is almost balanced in the sense that we have 300 patients with, and 200 without epilepsy. We consider this an edge case, where there are equal arguments for and against accuracy, but opted for easier interpretability for accuracy. For the five-class classification problem, we have a perfectly balanced dataset, therefore, accuracy is in our perspective the best measure. To reduce the impact of randomness, all results are averaged over five seeds.

Data Preparation

As the data is considered clean, and the entire dataset has to be used for the question at hand, no modifications were applied to the data. As the dataset is balanced with 100 cases for each category over- or under-sampling techniques were not applied. Enriching the dataset to reduce overfitting in this case is from our perspective not possible, as the patterns in the EEG are of high importance to spot epileptic behavior. Trying to replicate these patterns from the existing samples would be prone to errors as the differences between the samples from the five categories are very hard to spot.

Two-Class Classification

The most basic classifier for this dataset would be able to distinguish between an individual experiencing a seizure and one who is not. For this model, the target classes have been reduced to patients that are currently experiencing a seizure (label 1: Set E), and patients that are not experiencing a seizure at the time of recording (label 0: Sets A, B, C, D). Our model (*LSTM*) is tested against a regular deep neural network model (*NN*) so as to compare the performance of both models on this simplified task. Figures 6 and 7 show the validation accuracies and losses of our LSTM model and the NN model (averaged over 5 seeds). We notice that although the LSTM outperforms the NN model, the results are fairly close. We can conclude from this that on a simplified binary problem, the LSTM doesn't deliver significantly more performance than a regular neural network.

Figure 6 - training and validation accuracy

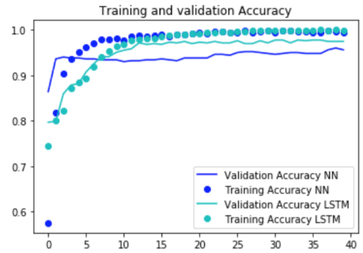


Figure 8 - training and validation loss

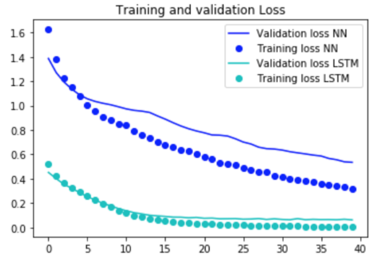
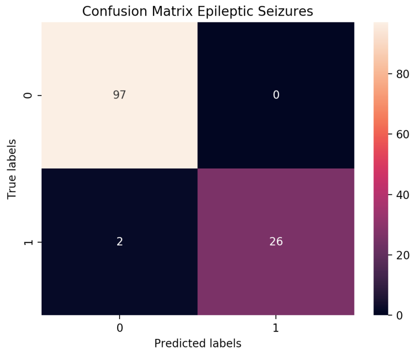


Figure 7 - confusion matrix two class classification



The confusion matrix shows us that our LSTM model performs well, and that even though our dataset for the two-class classification is unbalanced (400 instances of non-seizure patients (*0*) and only 100 instances of seizure patients (*1*)), accuracy is an appropriate measure since it there are no false positives and only 7% false negatives.

Three-Class Classification

In the three class-problem it becomes clearer that sharing time related information between neurons significantly increases performance and give the LSTM an advantage. Not only does the LSTM perform better, it has much better generalization ability than the regular neural network. We can conclude that the LSTM is appropriate to use for more complex time-series classifications. Figure 9 shows us the confusion matrix of our LSTM; we notice that our model can correctly distinguish between non-seizure patients (*label 0*), inter-ictal patients (*label 1*), and seizure patients (*label 2*).

Figure 9 - Confusion matrix three class classification

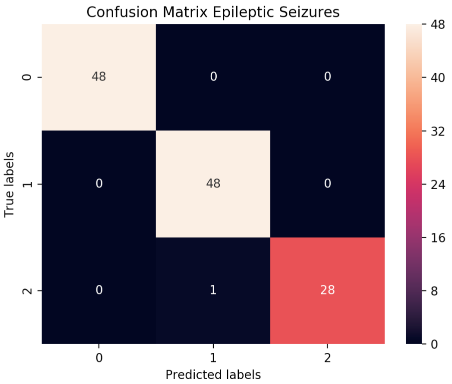


Figure 10 - training and validation fitness

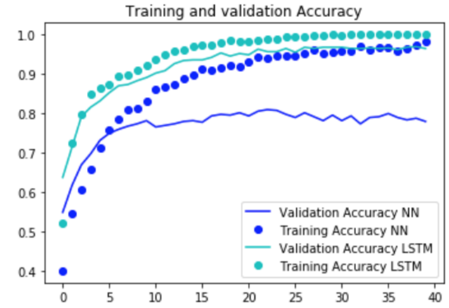
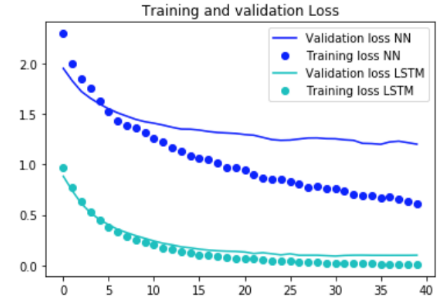


Figure 11 - training and validation loss



Five Class- Classification

The LSTM performs better than the normal neural network on the five-class problem; this was expected given the results of the three-class problem. However, we noticed that while the LSTM still performs well, there is a significant drop in generalization ability (i.e the LSTM overfits). This holds even more for the generalization ability of the NN. In this section we will further explore the LSTM through a series of optimization techniques, architectural changes, and feature engineering techniques will be implemented and tested in order to further increase its performance and reduce overfitting.

Figure 12 - training and validation fitness

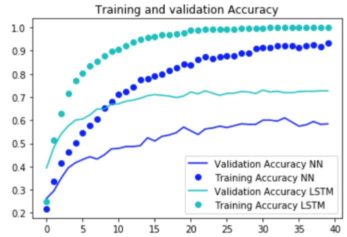
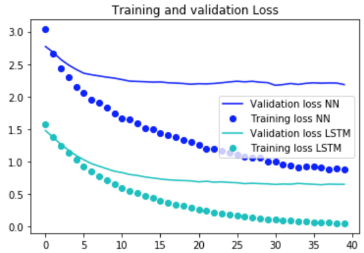
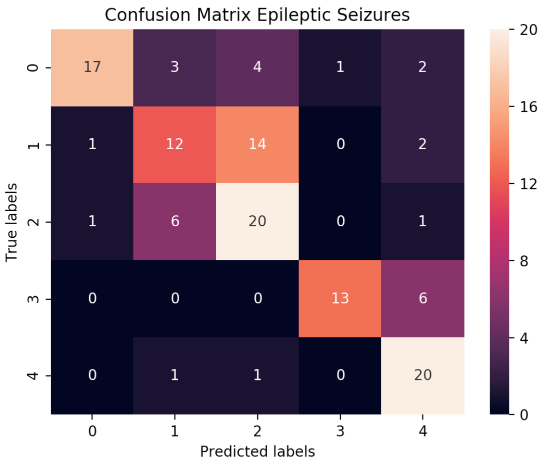


Figure 13 - training and validation loss



We notice from the confusion matrix in figure 14 that the model is mainly having trouble distinguishing between classes *1* & *2*, these correspond to the ‘*recording of an epileptic patient during seizure free period using electrodes implanted in the brain epileptogenic zone*’ and the ‘*recording of an epileptic patient during seizure free period from the hippocampal formation of the opposite hemisphere of the brain from C*’. In other words,

Figure 14 - confusion matrix five class classification



while our model can correctly classify inter-ictal patients, it is having difficulties distinguishing between the different parts of the brain that are being recorded.

It seems our model faces two main issues; overfitting and the inability to distinguish between the two types of inter-ictal recordings. In the following sections of this report, we will further explore the architecture and parameters of our model in order to try and improve it.

Methodology

In the previous part of the report we used a working configuration of the NN and LSTM model to compare them on each respective problem. We did not finetune the respective model for this part of the report. In the following part we outline elements of the LSTM that we further addressed when finetuning the model. The LSTM consisted of a first LSTM layer with 100 neurons, followed by a 10% dropout, followed by a Time Distributed Dense layer with 50 neurons and a Global average pooling layer. The last layer has the number of neurons fitting to the problem. Further details are in the code.

Epochs

Figure 16 - training and validation fitness

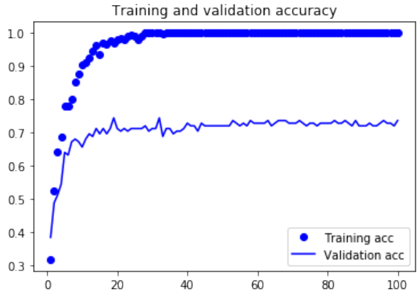
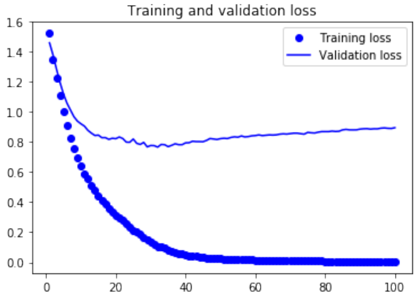


Figure 15- training and validation loss



Running our basic LSTM implementation over 100 epochs highlights that the appropriate number of epochs used for our system is at around 40 epochs. In Figure 15, we see our validation accuracy stabilizing at around 40 epochs whilst on the right-hand side we observe that beyond 40 epochs, an increasing dispersion between training loss and validation loss arises, suggesting we would be increasingly overfitting.

Pooling

Global Pooling layers in convolutional networks play an extremely important role as they reduce the dimension of data from the prior layer through the combination of its neuron clusters into a single neuron in the following layer.

Figure 17 - training and validation fitness

Global Max Pooling

Global Max pooling on the other hand extracts the maximum value from a cluster of neurons in the previous layer. Perfect test accuracy is achieved at around 20 epochs and we see our validation accuracy stabilizing at around 55% at 33 epochs and beyond. We observe a final training/validation loss spread of roughly 1.1.

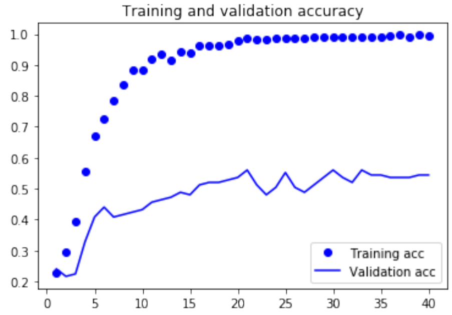


Figure 18 - training and validation fitness

Global Average Pooling

Applying the Global Average pooling layer extracts the average value from a cluster of neurons in the previous layer to then in turn create the appropriate output. Our validation accuracy terminates at just above 70% whereas we observe convergence to a perfect test accuracy at around 25 epochs. A final training vs validation loss spread of around 0.8 still provides some evidence of overfitting.

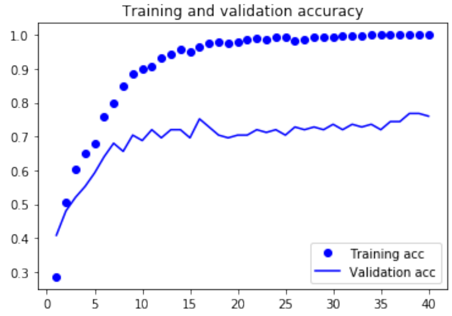
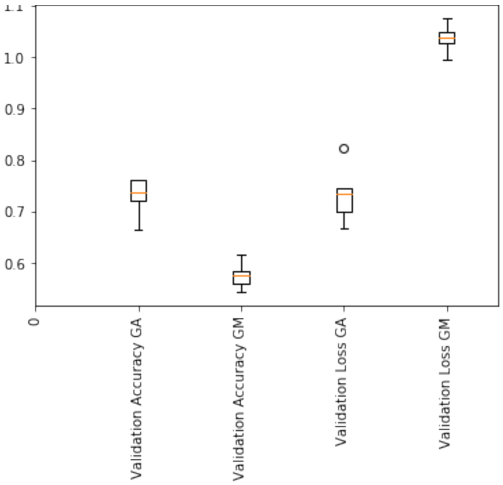


Figure 19 - results comparison pooling

Global Average Pooling vs Global Max Pooling

Seeing the two pooling methods in parallel, highlights a clear winner. The lowest observed Validation accuracy for Global Average Pooling is higher than the highest observed for Global Max Pooling. Furthermore, we see significantly more validation loss for Global Max pooling, pointing out such model would likely struggle to generalize as well as our winner, Global Average Pooling.



Timesteps & Dimension Sizes

Each row of the epileptic seizure dataset represents a 23.5 sec EEG recording of an individual; this data is represented as 4096 evenly-spaced, consecutive points. Although it is possible to create an LSTM neural network with 4096 input variables, our belief is that this data can be aggregated into a series of shorter time dimensions; aggregating the time information according to smaller *timesteps* would reduce computational time and create a simpler model. Our parameter *timesteps* represents the number of times the row data will be divided (i.e the number of input variables our LSTM model will receive). For example, if we choose 128 timesteps, our data will be split into 128 input variables, each containing 32 data points (4096/128 = 32).

Our initial model, the one that was tested against a normal neural network for each classification problem, was divided into 64 timesteps; this means the LSTM model received as inputs time blocks corresponding to 64 consecutive recordings. This division was arbitrary; in reality, our time series can be divided into can be divided by any factor pair of 4096 corresponding to the timesteps and data dimension. The times of the data point in an EEG recording are crucial in identifying the type of patient being recorded. In the interest of exploring the time relation between data points, and improving our model's capacity of correctly classifying the two types of inter-ictal patients, different timesteps and data dimensions were tested.

Figure 20 shows the validation scores for different data dimensions (i.e number of recordings per timestep), averaged over 5 seeds; we can conclude that 256 timesteps (i.e 16 data dimensions per timestep) yield the best validation scores for our data (Figure 22, dim 16). Reducing the data dimension seems to have improved our model's generalization ability while overfitting less. This leads us to believe that shorter timesteps improves our model; this means our LSTM performs better on this dataset when less information is shared on between the neurons because shorter timesteps account for less variation in the recordings.

Figure 21 - results comparison timesteps

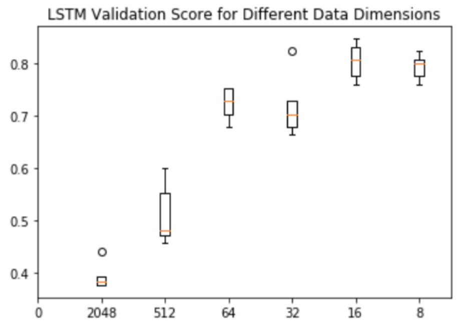


Figure 20 - confusion matrix 128 timesteps

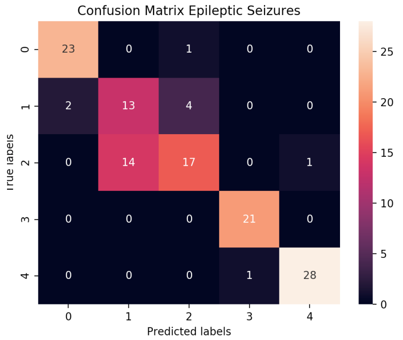


Figure 23 - training and validation fitness

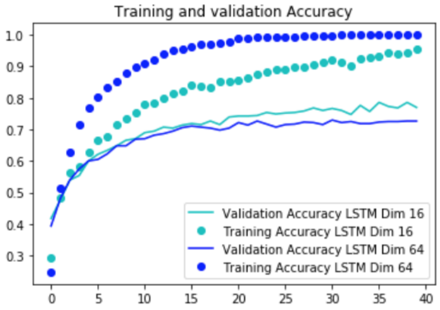


Figure 22 - training and validation loss

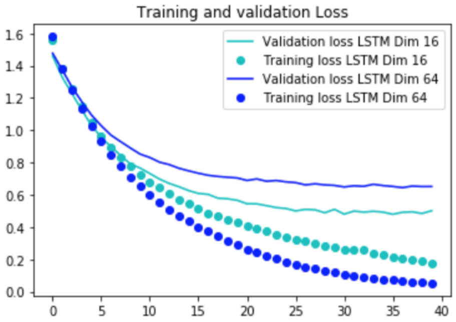


Figure 24 - training and validation fitness

Neuron configurations

To change the configuration of the LSTM model in the dimension of the number of Neurons, two layers can be optimized. Firstly, the number of neurons in the LSTM layer, secondly the number of neurons in the hidden layer. Additionally, further layers can be added to the network, however in theory, this is not necessarily required to solve this problem.

When adding neurons to the LSTM layer the accuracy increases. At 400 LSTM neurons the performance on validation accuracy is up around 5% compared to the model with only 100 LSTM neurons. However, the runtime drastically increases more than four-fold compared to 100 neurons in the first layer. Furthermore, an increase in overfitting is observable as the slope of the validation loss curve on the right in Figure 24 shows after epoch 25.

In other problems several concurrent LSTM layers, especially for text mining applications, proved useful. However, in our case this did not contribute to better results, probably because of the higher complexity of text analysis.

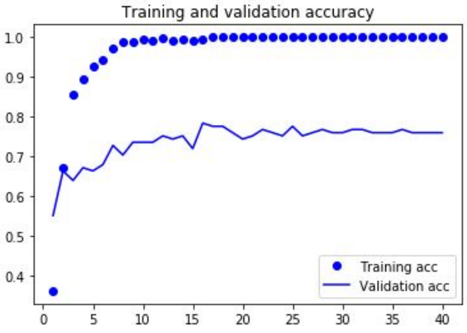
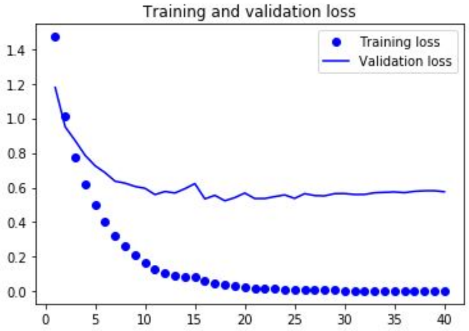


Figure 25 - training and validation loss



When changing the number of neurons in the hidden layer, it seems the optimum number of neurons is at 50. Increasing or decreasing only weakens the performance of the classifier. An increase to over 100 neurons seems to have the same overfitting behavior as described above on the instance of the LSTM layer.

Adding additional hidden layers with different number of neurons did not contribute to better results, but either the results were worse or rather seed dependent, which we interpreted as another sign of overfitting.

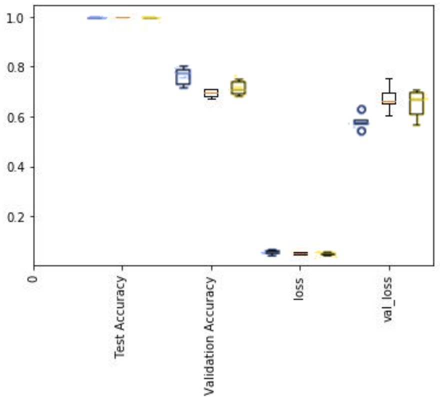
Dropout & Regularization

Introducing Dropout and Regularization to the model are two measures to counter overfitting of the model. The gap between the training and validation data can possibly be reduced by applying these techniques more successfully.

Adding dropout after a layer of neurons of any kind will eliminate the weights at random positions of the model differently at every iteration of the model. Therefore, it will introduce a degree of randomness to the training of the weights.

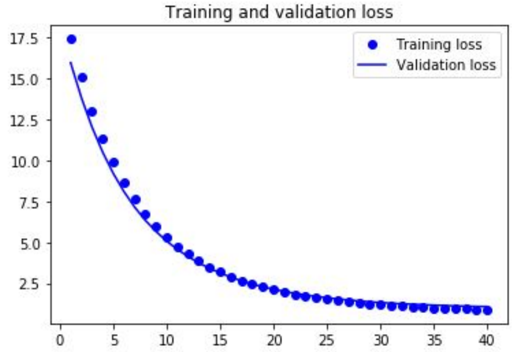
Without any Dropout in the LSTM model introduced (black in figure 26 on the right), the model performs worst in terms of validation accuracy and validation loss compared to the blue boxes, representing one dropout layer after the LSTM and Dense layer each (dropout probability of 0.1 at optimum configuration), and the orange boxes representing the scores with only dropout after the LSTM layer.

Figure 26 - results comparison dropout



For the regularization, three regularizers are commonly used, the L1 (lasso) the L2 (ridge) and L1/L2 (elastic net) regularizers. All these regularizers can be applied to the use of the activation inside the neuron, or to the output of the weights towards the loss function. However, in neither of the configurations, even with very low boundary values, did not improve the performance of the model or reduced overfitting. What can be observed though, is that the model learns a lot slower, and the initial error is much larger (10 fold) than without regularization as it is shown in figure 27 on the right.

Figure 27 - regularized loss development



Conclusion

For the five-class classification, the main objective of this project, we noticed that our model had a tendency to overfit, Furthermore, while it correctly classified the three types of recordings (i.e seizure, inter-ictal, non-seizure), it had issues identifying the part of the brain being recorded for the inter-ictal patients. While this error is not dramatic in the sense that a wrong classification does not affect the health of a patient, a more accurate classifier would help researchers to better understand and treat patients. In order to tackle the issues of overfitting and wrong inter-ictal classification, different parameters and network architectures were explored. Reducing the dimension of the LSTM output through a Global Average Pooling did not improve our model. However, increasing the number of timesteps (i.e reducing the data dimensions) improved the model's generalization ability and slightly improved the model's classification ability; shorter timesteps lead to less recording variation being transferred between neurons in the LSTM layer, which leads to a better assessment of the recording at the previous time step. Finally, adding random neuron dropouts to the model, a known technique for combating overfitting, did not improve this particular model. To conclude, this LSTM model is capable of distinguishing seizure patients, inter-seizure patients, and healthy patients with high accuracy, making it a robust model with strong practical applications.

Best model (*Five-class classification)*

LSTM layer, 100 Neurons input_shape= (265, 16), return_sequences = True, recurrent_regularizer =None))
model.add(Dropout(0.1))

```
model.add(TimeDistributed(Dense(50, kernel_regularizer =None)))
model.add(Dropout(0.1))
model.add(GlobalAveragePooling1D())
model.add(Dense(5, activation='softmax'))
model.compile(loss='sparse_categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

References

Yuan, Y., Xun, G., Jia, K. and Zhang, A. (2018). A multi-context learning approach for EEG epileptic seizure detection. *BMC Systems Biology*, 12(S6).

Hussein R., Palangi H., Ward R., and Wang Z. J.. (2018). Epileptic Seizure Detection: A Deep Learning Approach.

Seed Usage

In order to verify that running our models over a certain number of seeds suffices when calculating our mean valuation accuracy, we implemented a simple student's t-test, testing for the equality of 2 sample means. In other words, we feed two sets of samples to our hypothesis test, investigate the null hypothesis (H0) of mean1 = mean2 and counter-hypothesis (H1) of mean1 =/= mean2. Conducting this analysis identified 10 seeds as being the appropriate number of runs required in the calculation of our mean final validation accuracy.