Final project

School of Systems and Enterprises

CS 583 – Deep Learning

Prepared by:

Raif Bucar – CWID 10401402

Stevens Institute of Technology

May 9th, 2019

The problem

The problem chosen for the final project is the task of classifying the rare phenomenon of Partial Discharge (PD) of an electric grid's transformer.

PD is a phenomenon potentially destructive because if left unchecked, it may cause significant damage to the transformer. According to Wikipedia (accessed 05-01-2019): "*In electrical engineering, partial discharge (PD) is a localized dielectric breakdown (DB) of a small portion of a solid or fluid electrical insulation (EI) system under high voltage (HV) stress, which does not bridge the space between two conductors. While a corona discharge (CD) is usually revealed by a relatively steady glow or brush discharge (BD) in air, partial discharges within solid insulation system are not visible*."

A deeper look at the competition's discussion board reveals some more information on the problem, for example VSB-TU ENET data scientist Tomas Vantuch posted a few good insights at the discussion board:

- The dataset contains a significant amount of noise caused from the environment.
- PD may be detected by using pattern recognition.
- The dataset is unbalanced, due to the destructive nature of the PD phenomenon.

The dataset

According to the competition's website: "Each signal contains 800,000 measurements of a power line's voltage, taken over 20 milliseconds. As the underlying electric grid operates at 50 Hz, this means each signal covers a single complete grid cycle. The grid itself operates on a 3-phase power scheme, and all three phases are measured simultaneously."

The dataset consists of a training and test set, containing 8712 and 20337, respectively, time series of 800,000 voltage measurements. The dataset has data on all three phases, and the relation between each phase's PD occurrence is assumed to be independent (each phase could be experiencing PD independently of the others).

The dataset is in fact unbalanced, the number of positively labeled examples is 525, which represents roughly 6% of the dataset.

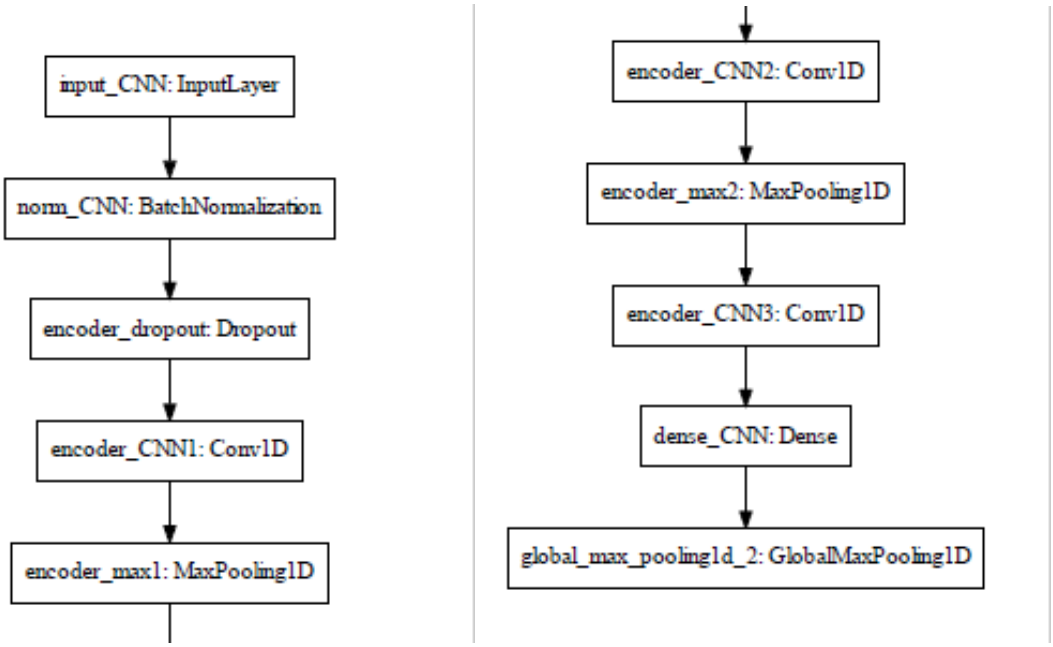
Some examples of the time series can be seen below:

The voltage values range from +45 V to -45 V, an important information in case a self-supervised method such as an auto-encoder is used.

The time series were reduced by taking the average of the series every 800 measuresments, reducing each time series to 1000 inputs. This was done for the sole purpose of saving processing time, given the limited processing power available.

First attempt – CNN classifier

The first attempt at this problem was to use a CNN classifier, since pattern recognition is one of CNN's strongest suits. The architecture is as seen below:



Total params: 8,106

Trainable params: 8,104

Non-trainable params: 2

Three aspects about this architecture are important to note:

1- Batch normalization has been applied because the data ranges from values close to zero and values up to +- 45V.
2- Dropout has been applied before the layer with most parameters because the number of trainable parameters surpasses the training sample size.
3- L1 regularization has been applied to all layers to further prevent overfitting.
4- Given the unbalanced dataset, a sample weight method has been applied, increasing the loss for misclassified PD=1 samples by 10 times.
5- The loss used is the categorical cross-entropy, since this is a classification problem.

These same concepts will be applied to all other proposed architectures as well. The summary of optimized hyperparameters are:

- Batch normalization: Aplied.
- Dropout: 20% masked inputs.
- L1 regularization parameter: 0.0001.
- Activation functions: tanh.
- Last activation function: softmax.
- Learning rate: 1E-4.
- Sample weight 10:1.

Which yielded the following results:

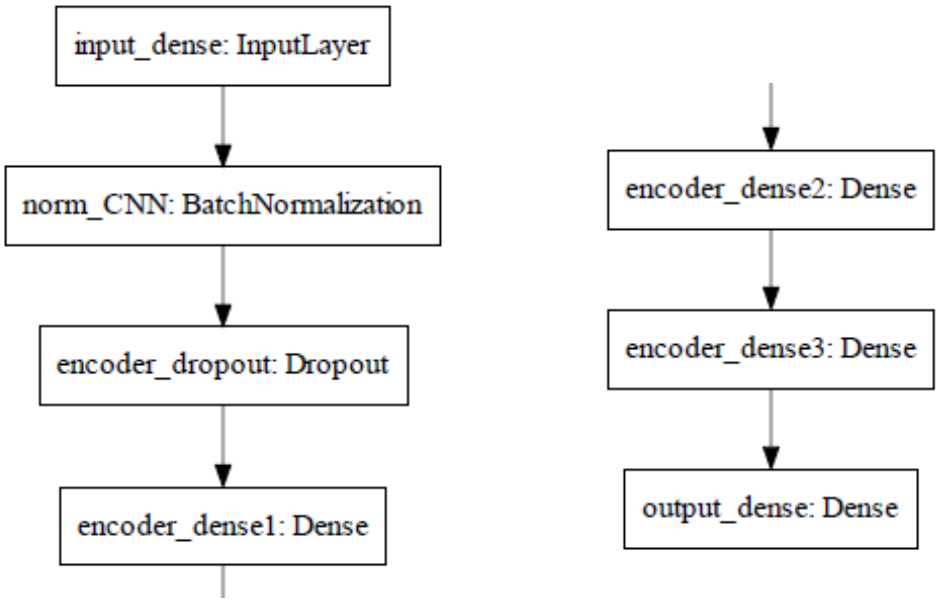
loss: 50.4426 - categorical_accuracy: 0.6958 - val_loss: 11.4552 - val_categorical_accuracy: 0.7889

And the confusion matrix looks as follows:

True / Predicted	PD=0	PD=1
PD=0	2720	5467
PD=1	198	327

Second attempt – Dense classifier

The second attempt was to use a densely connected neural network classifier, the architecture used was as follows:



Total params: 109,682

Trainable params: 107,682

Non-trainable params: 2,000

The following are the optimized hyperparameters:

- Batch normalization: Aplied.
- Dropout: 20% masked inputs.
- L1 regularization parameter: 0.001.
- Activation functions: tanh.
- Last activation function: softmax.
- Learning rate: 1E-4.
- Sample weight 15:1.

Which yielded the results:

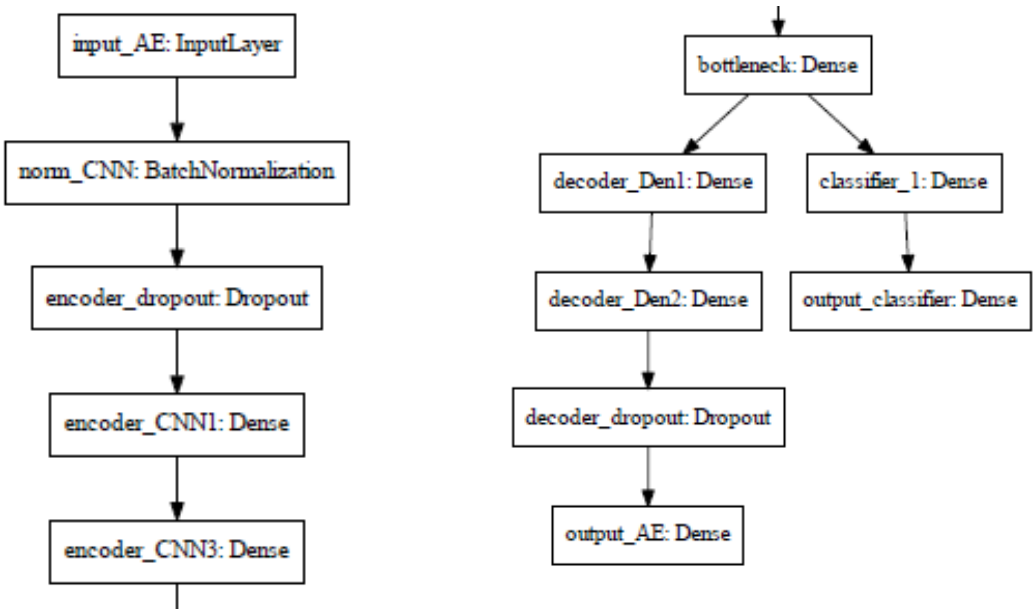
- 1s - loss: 2.4356 - categorical_accuracy: 0.6294 - val_loss: 1.7828 - val_categorical_accuracy: 0.5634

And the confusion matrix looks as follows:

True / Predicted	PD=0	PD=1
PD=0	2761	5426
PD=1	194	331

Third attempt – Dense autoencoder classifier

The third architecture was a densely connected autoencoder, which uses the bottleneck layers for multi-task optimization for a classifier, the structure is as follows:



Total params: 1,226,982

Trainable params: 1,224,982

Non-trainable params: 2,000

The following are the optimized hyperparameters:

- Batch normalization: Aplied.
- Dropout: 50% masked inputs.
- L1 regularization parameter: 0.0001.
- Activation functions: tanh.
- Last activation function: softmax.
- Learning rate: 1E-3.
- Sample weight 20:1.

Which yielded the results:

- 1s - loss: 2.4356 - categorical_accuracy: 0.6294 - val_loss: 1.7828 - val_categorical_accuracy: 0.5634

And the confusion matrix looks as follows:

True / Predicted	PD=0	PD=1
PD=0	2710	5477
PD=1	151	374

Comparison and chosen concept

Upon comparison of all the proposed architectures, it's clear that the autoencoder concept has a better performance:

- For a rather similar true negative rate (TNR) of roughly 33%, the last concept yielded the best true positive rate result (TPR) of 71%, this is still relatively better than the all negative (TNR=100%, TPR=0%) or all positive (TNR=0%, TPR=100%) naïve estimate.
- Given the nature of the problem at hand, trading off negative labeling accuracy for positive labeling accuracy is adequate, and thus chosen for the final submission.