"Soft sensor" measuring for ball mills based on LSTM neural networks

Abstract. In the present work, considering the strong uncertainty in ball mills, a soft sensor is made to determine the amount of material processed in a ball mill, feed level, the paper proposes a novel soft sensor based on Deep learning technique (a neuronal network, RNN-LSTM type), to improve the accuracy and reliability of fill level measurement. The data for modeling the system are collected by vibration and sound sensors attached on the ball mill shell, with these two signals combined are avoid distortions produced by sound and/or vibrations in the environment.
First the power spectral densities of the vibration signals are extracted, as characteristics, it is realized by Power Welch method. Second, all the characteristics are process by one neural network, because the system to model is nonlinear, a neural network is used, the activation function gives the nonlinearity to the model system. The neural network is set on regression mode for gives the measure data. The deep learning techniques and specially the RNN networks demonstrated are good in sequence tasks, as sequence predictor, (commonly used for spoken language recognition).

Keywords: Ball mill, Ball mill fil level, Soft sensor, LSTM regression.

1. Introduction

A ball mill is a slightly inclined rotating horizontal cylinder, partially filled with steel balls, which grind the material by friction and impact with the balls that go round to the degree of fineness necessary [1]. In order to achieve the desired particle size, the milling under industrial conditions is usually performed in "grinding circuits" [1]. In the "grinding circuit", ball mills have the lowest energy efficiency and also is difficult to measure the amount of material processed [2]. The amount of material processed is directly related to energy expenditure, so an efficient fill measurement of material contributes to the following advantages, can perform the automation of the plant (ball mill), energy savings of up to 10%, stability of the system, program the use of the mill in a flexible way, improve the stability of the system, prevent system overflows and jams [3].

The problem in ball mill is the difficulty to have a precise measurement for fill level, because material processed is measured indirectly. For solve this problem already exist some methods and measurements, including differential pressure sensors method [4]. The power consumption in the motor method [5], the ultrasonic method [6], the acoustic sensors method [7], [8], the vibration sensor method [9], [10], [11], but each of these methods have problems with measurement. The differential pressure method has problems with measurement delays and is heavily influence by the environment [12]. The power consumption in the motor method is widely used in the industry since it is not affected by the environment, but has a poor resolution and also is affected by the wear coals in the electric motor [12]. The ultrasonic method has problems with the processed material and steel balls, because when the fineness of the material changes the ultrasonic sensor signal also changes [12]. The acoustic sensors method is able to measure the fill level but is also affected for the environment noise [12]. The vibration sensor method, is commonly used in many equipment but also has the environment problem, but is not heavy affected as others methods [13] [12].

In this paper, the percent work is going to be combined two kinds of signals, the sound and vibration signal, it is to avoid the environment problems and improve the measurement process. In the last years some works about ball mill fill level measurements have being published [3], [13], [8], [9] [11], etc. All these works afford the problem with nonlinear methods like Back Propagation Neural Network (BPNN) [8], Cloud model [3], etc. These researches have made great progress for measurement of fill level in ball mills, but did not take the variable time into account. The time variable is important, because the fill level in ball mills is a sequence. For improve the accuracy, this paper introduces a LSTM neural network, it works with data sequence to make regressions, proving to be very efficient in other areas gives grates results in the speech generation, or predictions of written text [14].

The structure of the paper is as follows, in the second section is explain the concepts of LSTM networks, then the third section shows the acquisition of data and formation of the database, in the fourth section takes place the regression based on the LSTM neural network, in the fifth section are the results, making comparisons with other methods using the same dataset, and finally in the sixth section are the conclusions.

2. Concepts of LSTM networks

Neural networks, especially Deep learning type, are generally used for character recognition, speech to text, translations, artificial vision, image generation, etc. The RNN (LSTM) type neuronal networks are widely used for language modeling and speech recognition [15], reprehending the state of the art in these areas. The Recurrent Neural Network (RNN) is a deep neural network (DNN) that is adapted to sequence data, and as a result the RNN is also extremely expressive. RNNs maintain a vector of activations for each time step, which makes the RNN extremely deep. Their depth, in turn, makes them difficult to train due to the exploding and the vanishing gradient problems [15]. There have been a number of attempts to address the difficulty of training RNNs. Vanishing gradients were successfully addressed by Hochreiter & Schmidhuber in 1997, they developed the Long Short-Term Memory (LSTM) architecture which is resistant to the vanishing gradient problem.

LSTM neural networks have the special ability to remember important data over time designed exclusively to solve the long-term memory dependency problem [15], so they have very good results in sequencing tasks. It is for this reason that this method was chosen to perform the modeling of soft sensor for the ball mill. All LSTM networks are in the form of a chain and their common structure is as show the follow picture [15].

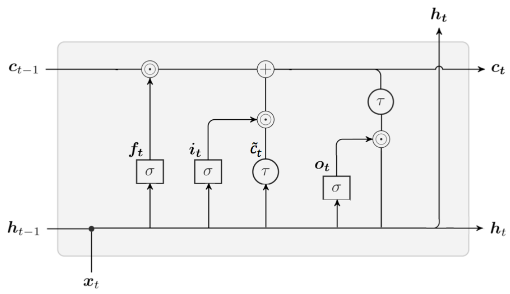

figure 1 classic structure for LSTM networks

The following diagram explains the intersections of each line, taking into account that each line loads with a vector, from the output node to the next input node.

figure 2 Wiring diagram in the figure 12

The most important part in the LSTM neural network structure is the "state", the state is represented by the horizontal line running through the whole diagram (figure 1), LSTM networks has the ability to adhere or suppress information to the state of the “Cell”, through gates. Gates are composed of σ (sigmoid) functions and a multiplication operation at the output of each neural network. The first step in an LSTM type network is passes the information through the state line, this decision is made through the σ (sigmoid) function, this stage is called "forget gate layer". The following equation represents the forget gate.

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f)$$

Equation 1 forget gate layer

This gate takes h_{t-1} and x_t in the input and to the output delivery 0 and 1, for each state, in the weights of the neuron C_{t-1}.

The next step is to save the new information in the weights of the neuron, this step is divided in two, the first stage is a σ (sigmoid) function called "input gate layer" which decides the values that will go to the next stage. The second stage is a τ (tanh) function, it creates the values that will be selected by the σ sigmoid function, $\tilde{C}_t$ [16].

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c$$

Equation 2

The end result of this stage is:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t$$

Equation 3

Finally, the output of the neuronal state the output of the neural network is based on state filtering, the filtering is performed with a sigmoid function, which again according to training data, decides which values pass and which values do not, After the state values pass through the τ (tanh) function, this function throws values between -1 and 1 [16].

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$
$$h_t = O_t \times \tanh(C_t)$$

Equation 4

The LSTM networks has different configuration, it depends on how many time steps or h_t we are going to take, in the sequence to sequence task with can set n inputs and m outputs, one input m outputs, n inputs and one output, in the present task we have the configuration of n inputs and one output, corresponding to n=6 and the output is h_7.

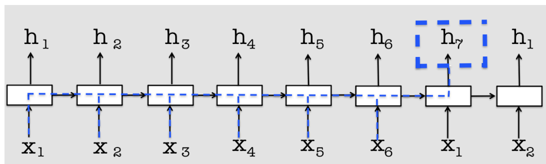

figure 3 configuration of the structure in the percent task, n=6 inputs an one output h_7

3. Data base

3.1 data acquisition

The "soft sensor" are commonly used in control systems, for the measurement of material processed in ball mills [17]. Many of these "soft sensors" use sound or vibration signals, the signals come from a ball mill shells [3]. According to the literature, the procedures for processing data, in the "soft sensor" are: transformation, cleaning and reduction of data [18]. This helps to have more efficient models for recognitions or regressions. The audio and vibration signals are collected by one accelerometer and one microphone coupled to the ball mill shell, both signals has the sampling frequency of $F_s = 51200\ Hz$, corresponding a Nyquist frequency of $FN = 25{,}6\ KHz$. Subsequently, the sampling of the signal is a time interval between 60 and 70 seconds of each fill level, the length of the data is approximately $51.2\text{KHz} \times 65\text{s} \cong 3{,}3 \times 10^6$ this procedure is similar to the procedure performed in [13].

The next step is to cut the beginning and end of each recording, this is done to erase the initialization and finalization of the signals (sounds and vibrations), only taking the stable part. Ball mill machines use electric motors for rotation, these motors are tri-phase motors and its control is performed by frequency, the motor frequency value is recorded at the same time as the sound and vibration signals, in The following figure shows these three signals.

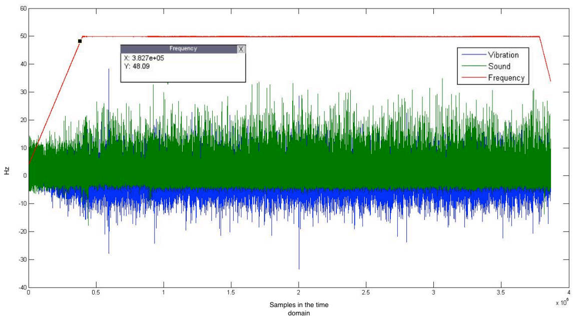

figure 4
This image shows, the frequency signal control in Hz, the vibration and audio signals, in the image the breakpoints in the audio and vibration signals are displayed. (signals of audio and vibration are not real-scale, only for visualization purposes)

For find the cutting points, (deleting the initialization and ending parts in an automatic way) were searched all the position where values are above 48 Hz (the frequency control has a variation between 0 and 50Hz), in a vector $position = [s_n, s_{n+1}, s_{n+2}, \ \ldots, s_{n+p}]$ where $s_n, s_{n+1}, \ldots, s_{n+p}$ represents positions and $n \approx 3\times10^5$ and $n + p \approx 3.28\times10^6 \rightarrow p \approx 3\times10^6$, then the first and last positions number is taken from the *position* vector, s_n *and* s_{n+p}, then are taken all the points between thus tow positions in the sound and vibration signal, this results in signal with length between 3×10^6 and 3.28×10^6. The total of the samples made in laboratory are 140 audio and 140 vibration signals and taking into account that each sample is part of a steady state in the grinding (without transition between level filling, or initialization-stopping states), a sweep of each signal is performed, with a window size of $L = 3\times10^6/22 = 136\,363$ samples (3×10^6 is the smallest sample size), without overlap, at each step a sub-sample is extracted and 22 steps are taken in each sample (components not enter the sweep are discarded), with what is formed 140 * 22 = 3080 samples in total.

3.2 Sound transformation

To extract the main characteristics of the audio and vibration signals, the transformation to the frequency domain is done with Welch method [19], because it has little computational cost and had a good performance in previous ball mill soft sensors, [3], [13], is choose. Welch method starts from a signal x (j) where j = [0, 1, ..., N-1], which has a spectral density $P(f)|f| \leq \frac{1}{2}$, segments Size L, overlapped with D samples between each segment [19]. The next step is to choose a window, in this case were choose a Hamming = W window type, with a "L" length. Then is calculated the convolution, and with this result the finite Fourier transform is calculated and take the real part and absolute values of the vector.

$$A_k(n) = \frac{1}{L}\sum_{j=0}^{L-1} X_k(j) \cdot W(j) e^{\frac{-2kij}{L}} \qquad i = \sqrt{-1}$$

Equation 5

$$I_k(f_n) = \frac{L}{U}\,|A_k(n)|^2$$

Equation 6

Where

$$f_n = \frac{n}{L} \qquad n = 0, 1, 2, \ldots, \frac{L}{2}$$

Equation 7

$$U = \frac{1}{L}\sum_{j=0}^{L-1} W^2(j)$$

Equation 8

With this we have the spectral power of each segment in the signal, to conclude we must calculate the mean of the vectors, as expressed in the following formula.

$$\hat{P}(f_n) = \frac{1}{k}\sum_{k=1}^{k} I_k(f_n)$$

Equation 9

The result is a vector of L/2 dimension, which contains all the frequencies present in the signal [19]. Welch method was performed on all the sound and vibrations samples, because hade choose a Haming window of L = Fs / 2, where Fs = 51200 Hz were resulting L = 25600 and a vector dimension of L / 2 = 12800. Following is showing the transformation of one sample.

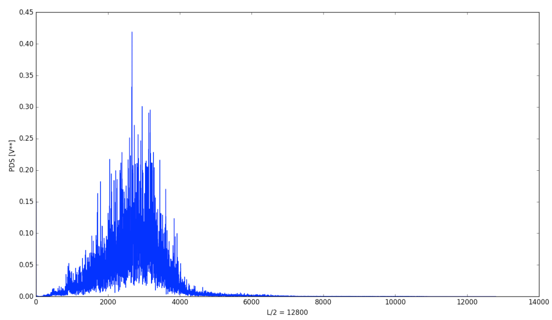

figure 5 This image shows the first sample of sound signal in the frequency domain.

Once all the samples were transformed to the frequency domain, was formed a matrix with all of them, this matrix has a dimension of 3080 number of samples, per 12800 number of frequencies (it is for sound and vibration signal), following is showing part of these matrices in picture format, where the brightness colors represent larger values and dark colors represent lower values.

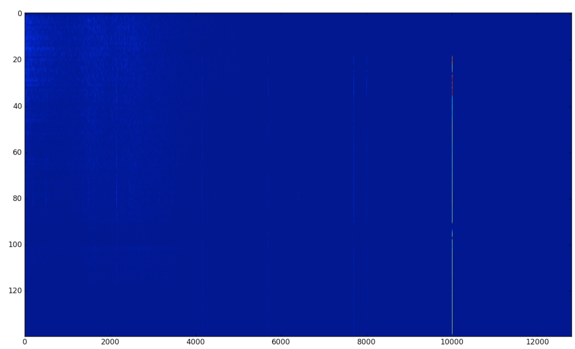

figure 6 This image shows the sound matrix with 140 samples, one at each fill level, from level 0 to level 139 with 12800 components.

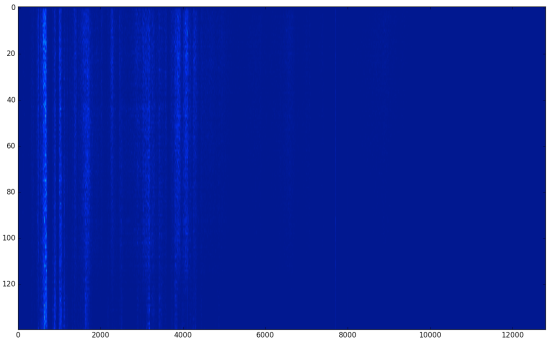

figure 7 This image shows the vibration matrix with 140 samples, one at each fill level, from level 0 to level 139 with 12800 components.

3.3 data labels

The neural network needs data entry and associated labels, where each sound and vibration signal has associated one label, it label is the filing level in the ball mill. For generate the dataset, the ball mill was filing with a range of 0L to 139L, with a deferens of 1L in every step, these data represent the labels. Up to this point we have 140 audio and vibration samples, with 140 labels, the next step is dived the signals. From the 140 audio and vibration samples, each sample were divided into 22 subsamples of the same size already explained in the 3.1 segment, each of these subsamples will keep the corresponding label of the original sample it is from. It results in a database of 3080 samples with 3080 labels (140 x 22), the following figures shows better the labels database.

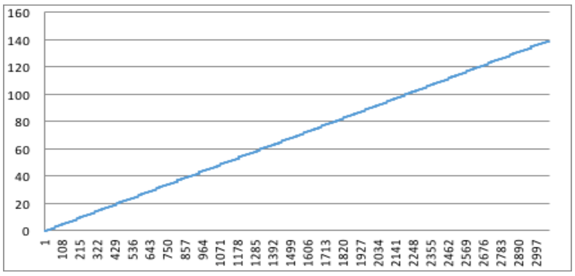

figure 8
This image shows all signs of file level, which representin g the tags in the linear regression.

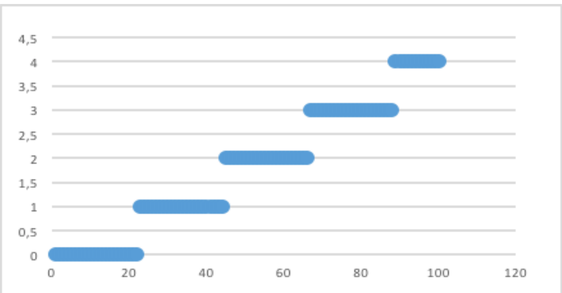

figure 9
This image shows 100 samples or labels, only for visualiza tion purposes.

As we can see the graph is in the form of a ladder, because weights were measured in a discrete form, ranging from 0 to 139, with a step of 1.

3.4 Database reduction

The reduction of the database is made to alleviate the computational load in the process of regression or classification [18], the reduction of the database also offers improvements, eliminating the noise or zeros from the characteristics [18]. In the present work, to perform the reduction of the database, we choose three methods, PCA, ranking and a heuristic method. The first method, based on PCA (Principal Components Analysis) is a method used for dimensional reduction, maintaining 99% of the variance the PCA method maintains a dimension of 550 characteristics (remember that the database has 12800 characteristics).The second method is ranking, which gives a ranking of the most important components, this ranking is based on linear regression, as the PCS method was taken 550 components.
The last method is a heuristic method, which consists of eliminating the characteristics that are mostly zero. The heuristic method was performed as follows.

We first calculate the maximum of each characteristic (12800 columns in the matrices are taken as characteristics), this results in a vectors of 12800 components, then were searched the vector maximum value, then the positions with values is less than 5% of the maximum value were set to 0 and the other values were set to 10.

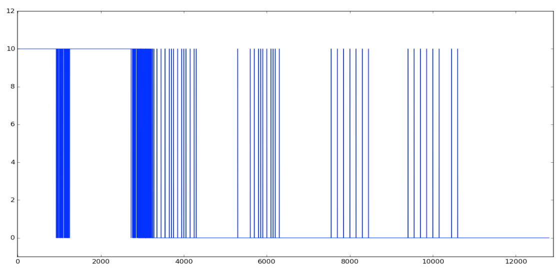

Figure 10 The image is showing the columns included and columns not included, following the criterion, greater than 5% of the maximum value.

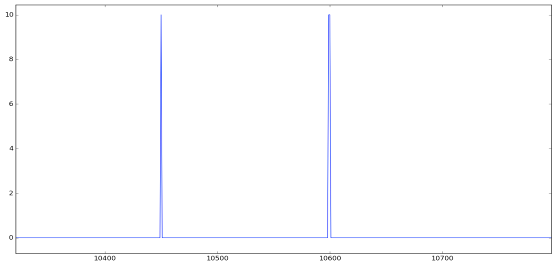

Figure 11 This image is an zoom of Figure 7, to have a better view of the latest components of the matrix.

For simplicity it was decided to cut all the characteristics that are after the 11000 position, since after position 10600 all the components have a spectral power less than 5%, this process results in matrices of 3080 x 11000. The last step in the heuristic method is the pooling, since the matrices stile has a large computational cost, the technique of “mean pooling” was used, with a window of 20 components without overlapping. This results in matrices of 3080 x 550 (sound a vibration correspondingly).

The pooling method consists in dividing each vector v(i) where i = [0, 1, ..., N-1] into non-overlapping windows of size L, then obtain the mean of each window and regroup each result in a new vector. The mean pooling function is as follows:

$$v_j = mean_{N\,xN}(v_i{}^{n\,x\,n} u(n\,x\,n))$$

Equation 10

Where $u(n\,x\,n)$ it is a window or also called kernel [10] which divides an array, in this case the matrix is $u(1\,x\,L)$ representing a vector. The mean function, for each interval, is shown below.

$$v_1(t) = \frac{1}{L}\sum_{i=0}^{L\text{-}1} V(i) \qquad i = 0, 1, \dots, L\text{-}1$$
$$v_2(t) = \frac{1}{L}\sum_{i=0}^{L\text{-}1} V(i+L) \qquad i = 0, 1, \dots, L\text{-}1$$
$$v_3(t) = \frac{1}{L}\sum_{i=0}^{L\text{-}1} V(i+2L) \qquad i = 0, 1, \dots, L\text{-}1$$
$$v_k(t) = \frac{1}{L}\sum_{i=0}^{L\text{-}1} V(i+(k\text{-}1)L) \qquad i = 0, 1, \dots, L\text{-}1$$

Equation 11

Where we have k segments $v_1(t),\ v_2(t), v_3(t), \dots, v_k(t)$, these contain all samples $(K\text{-}1)L + L = N$, where N is the number of components of each vector.
The resulting vector has a k = N / L length. In the present work, a L = 20 is selected, resulting in a vector of N / L = 11000/20 = 550, this is performed on the sound and vibration signal, obtaining two vectors, each of 550 components.

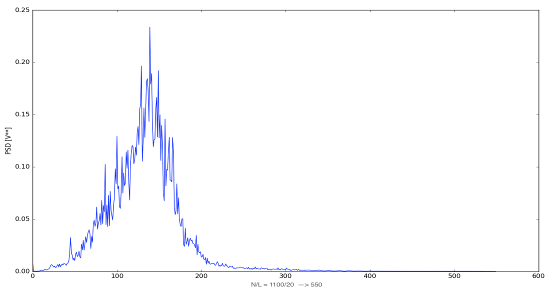

figure 12 This image shows a sample of the audio signal, after the process of pooling mean in the heuristic method, this signal is the same as is showed in the picture 4.

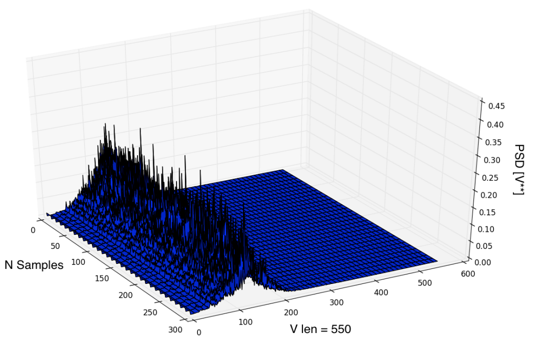

figure 13
This image shows the first 300 signals coming from the au dio signal, after the pooling in the heuristic method.

3.5 The train-test data separation

In this section is done the separation of the dataset to the train and test dataset, to achieve this objective were performing a simple algorithm. The samples obtained are 3080, after the reduction procedure, each sample is a vector of 550 components and all the samples forms a matrix of 3080 x 550, for sound and vibration signal, then these tow matrices are appending, resulting in a matrix of 3080 x 1100. To form the train and test datasets this matrix is dived in two parts, is dived by the sample dimension (3080), the even samples are part of the train set (1540 x 1100) and the odd samples are part of the test set (1540 x 1100), in this way we have a continuous group of samples for the training and testing of the neural network.

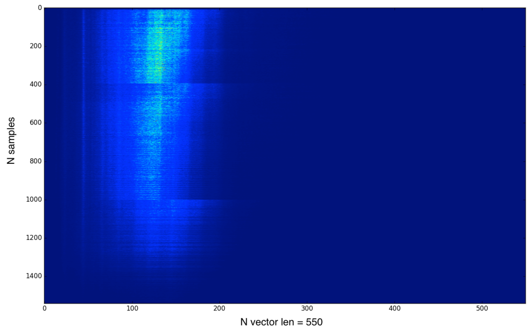

figure 14 This image shows the training array, which contains 1540 samples and each sample has 5 50 components.

4. Model regression

4.1 The training

The core of the model consists of an LSTM cell that processes one sample at a time and computes probabilities of the possible continuations of the fill level. The memory state of the network is initialized with a vector of zeros and gets updated after reading each simple and for computational reasons, the data is processed in mini-batches of size 250. The mini-batch is formed with a window of 6 size samples, a sweep of the training set (1540 samples) is performed, the sweep step is equal to one (overlapped sweeping). In order to make the learning process tractable, it is a common practice to truncate the gradients for backpropagation to a fixed number in this case the number is 2240 or until the loss reach 0.05 RMSE value, and the learning rate was set to 0.006.

4.2 The lost function

The loss function has two parts, the first is to calculate the value to be minimized and the second is the minimization function. The value to be minimized is calculated from the mini-batch, using the RMSE function.

$$P_{target\ i} = \sqrt{(y_i\text{-}\tilde{y}_i)^2}$$

Equation 12

The minimization function, commonly is called the loss function, which we want to minimize the average negative log probability of the calculated target (fill level).

$$loss = -\frac{1}{N}\sum_{i=1}^{N} \ln p_{target_i}$$

$$e^{-\frac{1}{N}\sum_{i=1}^{N} \ln p_{target_i}} = e^{loss}$$

Equation 13

In the training process its value is monitored throughout the process. The typical measure reported is average per-state perplexity.

4.3 The test

The test part is done with a group of data never previously "seen" by the neural network, which have the same configuration as the training data, this test set has 1540 samples and each one has 550 components (as the train set), this test set has the same structure as the train set, except for the mini-batches, only 6 samples are send at ones, sweeping the test set as in training stage, it means the mini-batch has a length of 6. In the training stage the backpropagation function and the loss function are suppressed, and just is taken the seventh state of the neural network, it at every time form the regression.

4.4 Tensorflow

All the code and the work was made with Tensorflow, the deep learning platform from Google, all the code for LSTM regression is upload in github https://github.com/RodrigoGantier/ball-mill-LSTM, the dataset is not proportionated.

5. Results

First, a test of the dimensional reduction method is done, PCA, Ranking and heuristic datasets are tested, the dataset with the best result will be used to make the comparisons with other regression methods (This test was performed only with the LSTM neural network regression method).

For each dataset, in training stage, the neural network was trained until yields an RMSE value of 0.01, later in the test stage, with the test samples, the neural network yields an RMSE, this results are shown in the follow table.

Method	RMSE sound	RMSE vibration	RMSE sound and vibration
PCA	4.33	3.71	3.54
Ranking	10.23	3.36	2.63
Heuristic	2.07	1.70	1.447

Table 1 Comparison between PCA, Ranking and Heuristic method, in LSTM regression.

The heuristic method with the signals combination of audio and vibration, yield the best results, for this reason this dataset is used to perform the comparison with other methods of regression. Results from the neural network are compared with other methods, the following regression methods were used. Random forest regression, SVM regression, Nearest neighbor's regression.

Method	RMSE sound and vibration
Random forest regression	10.487
SVM regression	40.73
LSTM network regression	1.447
Nearest neighbors regression	5.255

Table 2 Comparison between Random forest regression, SVM regression, LSTM network regression and Nearest neighbors regression.

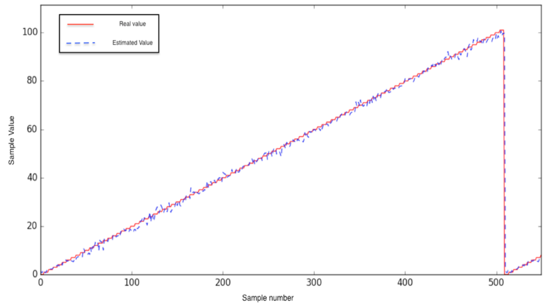

figure 15
Representation of the complete regression, blue line represents the estimated value, red line represents the real value.

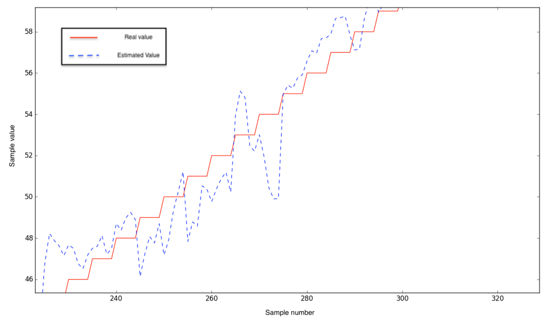

figure 16 Zoom of the figure 18, the initial state.

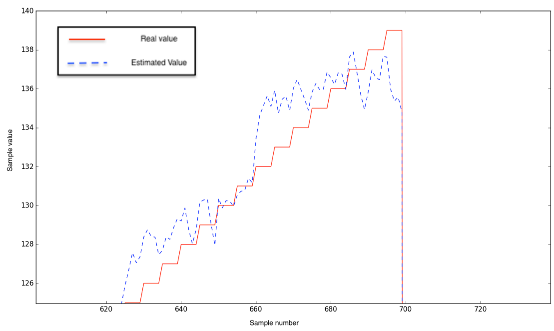

figure 17 Zoom of the figure 18, final state.

6. Conclusions

LSTM neural networks are commonly used in classification and recognition of natural language, since it is composed of sequences, which over time (samples) need to be remembered, this being the context of the sentence or paragraph. The classical systems of neural networks do not have the ability to contextualize (remember important patterns over time), for this reason the neural networks type LSTM have good results with data sequences as in this case, which is a system of continuous measurement through time.

Related Links

[1] V. Monov, B. Sokolov and S. Stoenchev, "Grinding in Ball Mills: Modeling and Process Control.," *CYBERNETICS AND INFORMATION TECHNOLOGIES,* 2012.

[2] T. Ming, B. Jing and F. Yushun, "Hybrid Intelligent Modeling Approach for the Ball Mill Grinding Process".

[3] G. YAN, X. GONG, X. XU and X. HAN, "Conceptual Representation and Measurement Model0f BaU Mm Fill Level Based on Cloud Model," *Chin.Soc.for Elec.Eng,* 2014.

[4] R. Farnham and B.-M. I, "Enhanced control system for non-standard pulverized fuel mills using individual mill capability indices," *Control '96, UKACC International Conference on (Conf. Publ. No. 427),* 1996.

[5] M. G. Melero, J. M. Cano, J. Norniella, F. Pedrayes, M. F. Cabanas, C. H. Rojas, G. Alonso, J. M. Aguado and P. Ardura, "Electric motors monitoring: an alternative to increase the efficiency of ball mills," *International Conference on Renewable Energies and Power Quality (ICREPQ'14),* 2014.

[6] Z. F and F. X.L, "Fuzzy-PID control system for level of ball mill based on level ultrasonic measurement," *Mechanical and Electrical Engineering Magazine,* 2008.

[7] B. ARUP, B. SUMAN and S. JAYA, "Designing of Intelligent Expert Control System Using Petri Net For Grinding Mill Operation," *WSEAS Transactions On Applications,* 2005.

[8] Y. Sha, Y. Chao and G. Y.G., "Analysis of acoustic signal and BP neural network-based recognition of level of coal in ball mill," *Journal of Northeastern University (Natural Science),* 2006.

[9] B. Behera, B. Mishra and C. Murty, "Experimental analysis of charge dynamics in tumbling mills by vibration signature technique," *Minerals Engineering 20,* 2007.

[10] J. Tang, L.-j. Zhao, J.-w. Zhou, H. Yue and T.-y. Chai, "Corrigendum to ‘‘Experimental analysis of wet mill load based on vibration signals of laboratory scale ball mill shell’’ [Miner. Eng. 23 (2010) 720–730]," *Minerals Engineering.*

[11] M. J. B. Z. P. Huang, "Investigation on measuring the fill level of an industrial ball mill based on the vibration characteristics of the mill shell," *Minerals Engineering.*

[12] T. C. J. C. Y.H. Sha, "Measure methods of ball mill's load," *Modern Electric Power ,* 2006.

[13] G. Yan, "Soft Sensor for Ball Mill Fill Level Based on Uncertainty Reasoning of Cloud Model".

[14] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville and Y. Bengio, "A Recurrent Latent Variable Model for Sequential Data," *arXiv:1506.02216v3 [cs.LG],* 2015.

[15] Rafal Jozefowicz, Wojciech Zaremba and Ilya Sutskever, "An Empirical Exploration of Recurrent Network Architectures," *Journal of Machine Learning Research,* 2015.

[16] W. Z. S. Vinyals, "RECURRENT NEURAL NETWORK REGULARIZATION," 2015.