BIDIRECTIONAL SPECTRAL-BASED TRANSFORMER FOR REMAINING USEFUL LIFE PREDICTION

FIELD OF THE INVENTION

[0001] The present invention relates to a bidirectional spectral-based transformer for remaining useful life prediction, in particular a method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model.

BACKGROUND OF THE INVENTION

[0002] Prognostic and Health Management (PHM) is crucial to preventing the physics of failure and monitoring the degradation of sophisticated engineered systems. Remaining Useful Life (RUL) is one of the most important indicators for PHM because it allows for proactive monitoring of the health condition of systems, enabling timely predictive maintenance along with failure prevention. By using RUL prediction, appropriate maintenance can be scheduled for the system, without expending extra resources and time.

[0003] There are two main approaches to prediction for PHM. A physical-based model is often used when there is a good understanding of a less obfuscated system. However, since most interactions in the system are not well understood, the practicality of the physical-based model is greatly reduced. Another popular approach to prediction is data-driven. The increasing availability of sensors, along with advances in machine learning algorithms, has enabled the data-driven approach to achieve more accurate performance. The model can learn to estimate the health state of the system from its past data.

[0004] Specifically, deep learning has been advancing rapidly and has superseded traditional machine learning methods in almost every field. An advantage of using a large neural network is that task-specific feature engineering is not required. Therefore, less handcrafted feature engineering is involved, and deep models can automatically generalize to different datasets without adjusting the network architecture. Furthermore, deep neural networks have much larger learning capacity, so they can capture very complicated mappings from input to output. There were many traditional machine learning methods applied to the prognostic dataset, such as Support Vector Regressor (SVR), Relevance Vector Regressor (RVR), and Multi-layer perceptron (MLP). They all treat RUL estimation as a supervised regression problem.

[0005] Most models failed to perform well on complex datasets as these models could not effectively extract the features of long and complex time series data. To predict the RUL for a complex system, the prediction model requires longer and longer prediction lengths. This implies the model should be capable of handling more past information. In order to meet the need for prediction in the long run, traditional time series prediction models, including ARIMA, are not competent enough as their concrete models need manual selection to account for various factors.

[0006] In recent years, a new type of deep learning architecture has emerged and dominated various fields, including computer vision and natural language processing. The Transformer takes sequential data as input, and it has replaced LSTM as a popular choice for modelling sequential data. The excellent performance of the Transformer in handling time series can be attributed to its ability to preserve dimensionality. Both the sequence dimension and the feature dimension can be preserved and propagated through the layers.

[0007] China Patent No. 113642414A discloses a rolling bearing residual service life prediction method based on Transformer model. The method for predicting the remaining service life comprising the steps of: denoising the original vibration signal of the rolling bearing by using Discrete Wavelet Transform (DWT); extracting time domain statistical characteristics from the denoised vibration signals of the rolling bearing to represent the degradation state of the rolling bearing; and training and using a deep neural network based on a Transformer model to complete the prediction of the residual service life of the rolling bearing. However, the positional encoding is added to the input embedding directly, without learnable attention in the Transformer model. The multi-head attention processed the positional encoding and input embedding together which could create noise. There may also be a feature collapse in the deeper layer of the model. When features are passed to deeper layers, some information may be lost and subtle patterns may not be discoverable. Therefore, there is a need to have an improved method of predicting the remaining service life by solving the problem of mixing noise, enhancing feature diversity and avoiding feature collapse in the deeper layer of the model.

[0008] United States Patent No. 20160069775A1 discloses a method for processing data obtained from a condition monitoring system, which comprises the steps of obtaining dynamic signal data in the form of a time waveform and/or a Fast Fourier Transform (FFT) from at least one sensor; extracting at least two parameters from the time waveform and/or FFT and transmitting or displaying the at least two parameters instead of the dynamic signal time waveform data and/or FFT. The invention step of extracting at least two parameters is carried using Discrete Wavelet Transform (DWT), Continuous Wavelet Transform (CWT) or another time domain analysis. However, there may be a mixing noise and feature collapse in the deeper layer of the model. When features are passed to deeper layers, some information may be lost and subtle patterns may not be discoverable. Therefore, there is a need to have an improved method of predicting the remaining service life by solving the problem of mixing noise, enhancing feature diversity and avoiding feature collapse in the deeper layer of the model.

[0009] China Patent No. 114297918A discloses an aero-engine residual life prediction method based on full attention deep network (Transformer) and dynamic ensemble learning. The method for predicting the remaining life of aero-engines, including the steps: calculating the remaining life RUL value of the training set data and the test set data; performing dimensionality reduction on the data subset FD001, and constructing a simulation data set; clustering the class centers and corresponding samples of the simulation data set; building the Transformer network module as the base learner; determining the weight of the base learner and the weighted ensemble output; and making predictions on test data. However, the positional encoding is added to the input embedding directly, without learnable attention in the Transformer model. The multi-head attention processed the positional encoding and input embedding together which could create noise. There may also be a feature collapse in the deeper layer of the model. When features are passed to deeper layers, some information may be lost and subtle patterns may not be discoverable. Therefore, there is a need to have an improved method of predicting the remaining service life by solving the problem of mixing noise, enhancing feature diversity and avoiding feature collapse in the deeper layer of the model.

SUMMARY OF THE INVENTION

[0010] It is an objective of the present invention to provide a method for predicting the remaining useful life of an engine based on a spectral-based Transformer model, which is lightweight, preserves more signal information and can reduce memory footprint and computational complexity.

[0011] It is also an objective of the present invention to provide an enhanced of feature extraction ability of the network and an effective method for predicting the remaining useful life of an engine based on a bidirectional Transformer architecture.

[0012] It is also a further objective of the present invention to provide a method for predicting the remaining useful life of an engine, which can prevent mixing noise into the model, feature collapse in the deeper layer and enhance feature diversity.

[0013] Accordingly, these objectives may be achieved by following the teachings of the present invention. The present invention relates to a method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model comprising: obtaining sensor data from at least one sensor; filtering sensor data; normalizing the sensor data; inputting the normalized data into a bidirectional spectral-based Transformer model by adopting Discrete Cosine Transform (DCT); training the bidirectional spectral-based Transformer model based on a pre-set training data set; and predicting the remaining useful life of the engine based on the trained bidirectional spectral-based Transformer model.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] The features of the invention will be more readily understood and appreciated from the following detailed description when read in conjunction with the accompanying drawings of the preferred embodiment of the present invention, in which:

[0015] FIG. 1 illustrates a diagram of the bidirectional spectral-based Transformer for remaining useful life prediction;

[0016] FIG. 2 illustrates a diagram of the bidirectional architecture Transformer in FIG.1;

[0017] FIG. 3a illustrates a diagram of the spectral-based Transformer in FIG.1;

[0018] FIG. 3b illustrates a diagram of the spectral-based Transformer method flow in FIG. 3a;

[0019] FIG. 4 illustrates a diagram of the spectral-based attention with multi-head shortcut in FIG. 3a and FIG. 3b.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[0020] For the purposes of promoting and understanding of the principles of the invention, reference will now be made to the embodiments illustrated in the drawings and described in the following written specification. It is understood that the present invention includes any alterations and modifications to the illustrated embodiments and includes further applications of the principles of the invention as would normally occur to one skilled in the art to which the invention pertains.

[0021] The present invention teaches a method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model **200** comprising: obtaining sensor data from at least one sensor; filtering sensor data; normalizing the sensor data; inputting the normalized data into a bidirectional spectral-based Transformer model **200** by adopting Discrete Cosine Transform (DCT); training the bidirectional spectral-based Transformer model **200** based on a pre-set training data set; and predicting a remaining useful life **108** of the engine based on the trained bidirectional spectral-based Transformer model **200**.

[0022] In a preferred embodiment of the present invention, the bidirectional spectral-based Transformer model **200** is configured as a bidirectional neural network comprises a forward Transformer **202** and a backward Transformer **204**. The forward Transformer **202** and the backward Transformer **204** each comprise an encoder and a decoder.

[0023] In a preferred embodiment of the present invention, the inputting of the normalized data into a bidirectional spectral-based Transformer model **200** in the encoder, further comprises the steps of: inputting the normalized data to a linear projection layer for generating contextual embedding in the encoder; passing through a multi-head spectral transform attention of the contextual embedding with a first multi-head shortcut **330** in the encoder to form a feature data; diversifying the feature data by the first multi-head shortcut **330**; passing the extracted feature data to an addition and normalization layer followed by a feed forward network; and providing a first encoder output to a next encoder along with an initial temporal attention.

[0024] In a preferred embodiment of the present invention, the method further comprises the steps of: encoding positional and contextual information in the encoder by a United Positional and Contextual Attention (UPCA) **320**; creating a positional embedding by a temporal-attention model; passing the positional embedding to a separate multi-head spectral transform attention with a second multi-head shortcut **332**; fusing the output of the temporal attention with the output of the multi-head spectral transform by concatenation on the temporal dimension; and wherein the fusion of the output is a linear layer in the embedding dimension; and the output dimensionally is the same as the original contextual embedding.

[0025] In a preferred embodiment of the present invention, the inputting of the normalized data into a bidirectional spectral-based Transformer model **200** in the decoder, further comprises the steps of: passing embedding from the contextual embedding of the encoder as input in the decoder; passing through a first multi-head self-attention of the contextual embedding with a third multi-head shortcut **334** in the decoder to form a feature data; diversifying the feature data by the third multi-head shortcut **334**; passing the extracted feature data to an addition and normalization layer; passing through a second multi-head self-attention with a fourth multi-head shortcut **336** in the decoder; diversifying the feature data by the fourth multi-head shortcut **336**; and passing the extracted feature data to an addition and normalization layer and followed by a feed forward network; and flattening the output data and sending to a linear layer for obtaining a remaining useful life **108** for every time step in a sliding window.

[0026] In a preferred embodiment of the present invention, the passing of the data through the multi-head shortcuts **330, 332, 334, 336** further comprises the step of inputting the positional embedding and contextual embedding into a multi-head spectral transform, a residual connection **400** and a linear layer without any activation function. The method further comprises of combining and adding all the outputs before passing to an addition and normalization layer.

EXAMPLE

[0027] A diagram of the bidirectional spectral-based Transformer for remaining useful life prediction is illustrated in **FIG.1** wherein the flow comprises sensor data collection **100**, data feature engineering **102**, data pre-processing **104**, bidirectional spectral-based Transformer **200** and RUL information **108**. The sensor data collection **100** involves collecting any sensor reading from at least one sensor and arranging it in a format to be processed by the RUL prediction model. The Commercial Modular Aero Propulsion System Simulation (C-MAPSS) dataset is used in the present invention, which is a widely used prognostics dataset. The dataset contains the sensor data from a run-to-failure simulation. During the simulation, a series of sensor data was recorded and each engine with a different initial condition was degraded until a threshold was reached. Therefore, this dataset is used as a common RUL prediction benchmark. The objective of RUL prediction is to predict the number of cycles before a system runs into failure based on the sensor data.

[0028] The C-MAPSS dataset consists of four sub-datasets (FD001-FD004), each with different operational conditions and fault conditions, as shown in Table 1.

Table 1: A summary of the sub-datasets in C-MAPSS

Sub-dataset	FD001	FD002	FD003	FD004
Training set	100	260	100	249
Testing set	100	259	100	248
Operational conditions	1	6	1	6
Fault conditions	1	1	2	2
Number of training samples	20631	53759	24720	61249
Number of testing samples	100	259	100	248

[0029] The operational conditions include single or different operational conditions, and the fault modes include fan degradation or high-pressure compressor (HPC) degradation. RUL prediction is essentially a time series regression:

$$X = (x_t | t = 1, \dots, RUL_{max}) \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(1)$$

wherein $x_t \in R^n$,

RUL_{max} is the maximum number of cycles of the engine

$$Y = (y_t | t = 1, \ldots, RUL_{max})\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

wherein $y_t \in R$

[0028] The aim in the present example is to use *X* to predict *Y*. During training, a fixed-length sliding window of size *w* is applied to the time series. The input is now represented as:

$$X = (X_t^w | t = w, \ldots, RUL_{max})\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

wherein $X_t^w = [x_{t-w+1}, \ldots, x_t] \in R^{n \times w}$

$$Y = (Y_t^w | t = w, \ldots, RUL_{max})\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

wherein $Y \in R^{1 \times w}$

[0029] The window size is 40 for FD001 and FD003 whereas 60 for FD002 and FD004. This is because the FD002 and FD004 are more complicated, and hence require a longer sequence to train the model.

[0030] The data features are further filtered and selected according to the Kolmogorov-Smirnov test and Kendall's tau test in the tsfresh library. The sensors that are not statistically significant are dropped from the dataset. No additional feature engineering is used to generate more features for training. The filtered sensor data will be further normalized to make the training more stable. A standard way to normalize the sensor data is to subtract it from the mean and then divide it by the variance, as follows:

$$z = (x-u)/s \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

wherein *u* and *s* denote the mean and standard deviation of the sensor data.

[0030] The input of the normalized data will be further input to the bidirectional spectral-based Transformer model **200** as illustrated in **FIG.1** wherein the input is a 2-D tensor of data from multiple sensors over a sliding window of length w. The DCT is used to transform a signal into the frequency domain. The attention mechanism is much faster than self-attention when running on a Graphics Processing Unit (GPU) with the use of said DCT. Rather than using complex exponential kernels in the Fourier transform, DCT Attention (DCTA) uses cosine kernels. Therefore, DCT has the properties of high energy compaction and having a real spectrum, which can preserve more signal information and less computational complexity. In the present example, DCTA is adopted instead of the Fourier transform for self-attention in the encoder. The output y is real. A 1D DCT, denoted as D_emb, is first applied to the embedding dimension, and then another 1D DCT, denoted as D_seq, is applied to the sequence dimension, as follows:

$$y = D_{seq}D_{emb}(x) \quad \text{......................................(6)}$$

wherein x and y denote the input embedding and the DCTA output, respectively.

[0031] The bidirectional spectral-based Transformer model **200** is trained based on a pre-set training data set to obtain a trained bidirectional spectral-based Transformer model **200**. The remaining useful life of the engine is predicted based on the trained bidirectional spectral-based Transformer model **200**. The last timestep in the sliding window of the output of the model is the current RUL prediction for the particular system.

[0032] The bidirectional spectral-based Transformer model **200** in the present invention is configured as a bidirectional neural network. **FIG.2** illustrates a diagram of the bidirectional architecture Transformer **200** wherein a reverse of the input sequence provides another perspective on the model in the bidirectional neural network. The bidirectional architecture Transformer **200** consists of a forward Transformer **202** configured in the normal sequence and a backward Transformer **204** configured in the reverse sequence. They can be viewed as an ensemble of two models. The outputs of the two separate Transformers are concatenated and passed to a fully connected layer with ELU activation and batch normalization. Then, the result is passed to a linear layer to generate an output of the same length as the sliding window and estimate the RUL for each time step. The whole sequence in the bidirectional Transformer **200** can be processed without needing to be split into one input per timestep. Said bidirectional Transformer **200** provides excellent feature extraction for some complex tasks which have a long sequence. The weight from the past values is learnt to predict the future values with said bidirectional learning.

[0033] **FIG. 3a** and **FIG. 3b** illustrates a diagram of the spectral-based Transformer **200**, which comprises an encoder and a decoder, and details of the spectral-based Transformer method flow, respectively. There are two layers in the encoder and decoder. The output of the Transformer **200** is a vector, where each value represents a RUL for every time step in the sliding window. The Mean Square Error loss is used to compare the predicted RUL and the ground-truth RUL. The inputting of the normalized data into a bidirectional spectral-based Transformer model **200** in the encoder, comprises the steps of: inputting the data to a linear projection layer for generating contextual embedding in the encoder; passing through a multi-head spectral transform attention of the contextual embedding with a first multi-head shortcut **330** in the encoder; diversifying the feature by the first multi-head shortcut **330**; passing the extracted feature to an addition and normalization layer and followed by a feed forward network; and providing the first encoder output to the next encoder along with the initial temporal attention.

[0034] The method of inputting normalized data into the bidirectional spectral-based Transformer **200** in **FIG.3a** and **FIG. 3b** further comprises the steps of: encoding positional and contextual information in the encoder by UPCA **320**. The UPCA **320** solves the problem of mixing noise, sensor data and positional data using an attention mechanism. Temporal attention is the output of learnable attention specifically designed for processing ordered sequences. The method further comprises of: creating a positional embedding by a temporal-attention model; passing the positional embedding to a separate multi-head spectral transform attention with a second multi-head shortcut **332**; fusing the output of the temporal attention with the output of the multi-head spectral transform by concatenation on the temporal dimension, as below:

$$Fusor = Concat\begin{pmatrix} Contextual\ embedding, \\ Temporal\ Attention \end{pmatrix} W_F, \ldots\ldots\ldots\ldots (7)$$

[0035] wherein the fusion model is a linear layer in the embedding dimension and the output dimensionally is the same as the original contextual embedding. Hence, a two-dimensional relationship map is produced by the model. Unlike the traditional Transformer, the positional embedding is not directly added to the contextual embedding.

In doing so, the order of the sequence can be preserved without mixing up the contextual embeddings and positional encoding. The dimensionality of the input and output from temporal attention remains the same. Most importantly, the input and output of temporal attention are two-dimensional, which means that the time series has not been flattened, and more information and sequence order can be preserved during propagation.

[0036] In addition, the concatenation operation does not remove any information from the embedding or the sequence dimension. The learnable fusion model selects the important embedding in the contextual attention and temporal attention for every time step, based on weighted summation. The learnable attention helps the fusion model investigate the relationship between each time step without explicit coding and assigning a weight to each time step for every embedding dimension.

[0037] As further shown in **FIG.3a** and **FIG. 3b**, the inputting of the normalized data into a bidirectional spectral-based Transformer model **200** in the decoder, further comprises the steps of: passing the data from the contextual embedding of the encoder as input in the decoder; passing through a first multi-head self-attention of the contextual embedding with a third multi-head shortcut **334** in the decoder; diversifying the feature by the third multi-head shortcut **334**; passing of the extracted feature to an addition and normalization layer; passing through a second multi-head self-attention with a fourth multi-head shortcut **336** in the decoder; diversifying the feature by the fourth multi-head shortcut **336**; and passing of the extracted feature to an addition and normalization layer followed by a feed forward network; and flattening the output and sending to a linear layer for obtaining a remaining useful life for every time step in a sliding window.

[0038] **FIG. 4** illustrates an example diagram of the spectral-based attention with the first multi-head shortcut **330** in **FIG.3a** and **FIG. 3b.** The spectral-based attention is applicable to the first **330**, second **332**, third **334** and fourth **336** multi-head shortcuts in the present invention. The first multi-head shortcut **330** uses a simple linear projection layer, which comprises the steps of inputting the positional embedding and contextual embedding into a multi-head spectral transform, a residual connection **400** and a linear layer without any activation function. The second step of inputting via residual connection **400** is different from the residual connection in the original Transformer, because the shortcut passes the input to the fusion model, not directly to the normalization layer. The features are projected into a different feature space in the third step of inputting via linear layer without any activation function. This can produce a distinct feature in parallel to the original one. All the outputs will be combined and added before passing to an addition and normalization layer. Therefore, the first multi-head shortcut **330** could enhance the feature diversity and avoid feature collapse in the deeper layer, in addition to the standard residual connection. Said multi-head shortcut is introduced by said three pathways, as illustrated in **FIG.4** and below:

$$MultiHead\ Shortcuts\ =\ T(x) +\ x\ +\ \sum_{i=1}^{d} W_i(X_i),\ (10)\ ....(8)$$

wherein T( ) is the Multi-Head Spectral Transform,
x is the input, and
d is the dimension of the input embedding.

[0039] The method for predicting the remaining useful life of an engine based on the spectral attention mechanism in the present invention outperforms state-of-the-art deep models for RUL prediction by at least 30% in terms of RMSE and 50% in terms of score on hard sub-datasets as illustrated in Table 2. Accordingly, the proposed BST (DCT) achieves the best result in the hardest sub-dataset FD004. It indicates that BST (DCT) has the capacity to handle more complicated trend and pattern.

Table 2: Comparison table

Model	FD001	FD002	FD003	FD004
MLP [9]	37.56	80.03	37.39	77.37
SVR [9]	20.96	42.00	21.05	45.35
RVR [9]	23.80	31.30	22.37	34.34
CNN [9]	18.45	30.29	19.82	29.16
DW-RNN [7]	22.52	25.90	18.75	24.44
MTL-RNN [7]	21.47	25.78	17.98	22.82
LSTMBS[8]	14.89	26.86	15.11	27.11
LSTM [6]	16.14	24.49	16.18	28.17
DCNN [39]	12.61	22.36	12.64	23.31
Semi supervised [40]	12.56	22.73	12.10	22.66
EN [41]	13.58	19.59	19.16	22.15
Transfer learning [42]	19.65	29.43	22.40	29.95
Proposed				
BST (DCTA)	**12.18**	**12.81**	12.32	**13.95**
BST (FCTA)	12.33	12.82	12.71	14.02
BST (FTA-RI)	12.55	12.84	**12.09**	14.07
% change (DCTA)	-3%	-35%	+2%	-37%
% change (FCTA)	-2%	-35%	5%	-37%
% change (FTA-RI)	0%	-34%	0%	-36%

Model	FD001	FD002	FD003	FD004
MLP [9]	17972	7802800	17409	5616600
SVR [9]	1382	589900	1598	371140
RVR [9]	1503	17423	1432	26509

CNN [9]	1287	13570	1596	7886
LSTMBS[8]	481	7982	493	5200
LSTM [6]	338	4450	852	5550
DCNN [39]	274	10412	284	12466
Semi supervised [40]	231	3366	251	2840
EN [41]	228	2650	1727	2901
Proposed				
BST (DCTA)	**212**	**834**	222	**1136**
BST (FCTA)	230	913	243	1296
BST (FTA-RI)	253	1008	**220**	1236
% change (DCTA)	-7%	-69%	-12%	-60%
% change (FCTA)	+1%	-66%	-3%	-54%
% change (FTA-RI)	+11%	-62%	-12%	-56%

[0040] Said spectral-based can effectively bridge the performance gap between the easy and hard sub-datasets in C-MAPSS, where the performance is more stable. This can facilitate the models' generalization on different datasets. In particular, the use of Spectral-based Attention (SBA) based on DCT can make the Transformer have a simpler design and the output of the transform is real, rather than complex. The lightweight attention mechanism, based on SBA, can perform better than the traditional self-attention mechanisms on small datasets, which is very useful because it is difficult and costly to collect sufficient amounts of training data from complicated systems.

[0041] The method for predicting the remaining useful life of an engine based on the bidirectional Transformer **200** in the present invention provides an efficient prediction of RUL. The UPCA **320** and multi-head shortcuts **330, 332, 334, 336** allow the model to better understand the input time sequence and, thus, achieve better performance than absolute positional encoding for RUL prediction. A potential future direction of this research is the use of Transformer and lightweight self-attention for RUL prediction or other time-series regression tasks.

[0042] The present invention explained above is not limited to the aforementioned embodiment and drawings, and it will be obvious to those having an ordinary skill in the art of the prevent invention that various replacements, deformations, and changes may be made without departing from the scope of the invention.

CLAIMS

WHAT IS CLAIMED:

1. A method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model (200) comprising:

obtaining sensor data from at least one sensor;

filtering the sensor data;

normalizing the sensor data;

inputting the normalized data into a bidirectional spectral-based Transformer model (200) by adopting Discrete Cosine Transform (DCT);

training the bidirectional spectral-based Transformer model (200) based on a pre-set training data set; and

predicting a remaining useful life (108) of the engine based on the trained bidirectional spectral-based Transformer model (200).

2. The method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model (200) according to claim 1, wherein the bidirectional spectral-based Transformer model (200) is configured as a bidirectional neural network comprises a forward Transformer (202) and a backward Transformer (204).

3. The method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model according to claim 2, wherein the forward Transformer (202) and the backward Transformer (204) each comprise an encoder and a decoder.

4. The method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model (200) according to claim 2, wherein the inputting of the normalized data into the bidirectional spectral-based Transformer model (200) in the encoder, further comprising the steps of:

inputting the normalized data to a linear projection layer for generating contextual embedding in the encoder;

passing through a multi-head spectral transform attention of the contextual embedding with a first multi-head shortcut (330) in the encoder to form a feature data;

diversifying the feature data by the first multi-head shortcut (330);

passing the extracted feature data to an addition and normalization layer and followed by a feed forward network; and

providing a first encoder output to a next encoder along with an initial temporal attention.

5. The method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model (200) according to claim 4, wherein the method further comprising the steps of:

encoding positional and contextual information in the encoder by an Untied Positional and Contextual Attention (UPCA) (320);

creating a positional embedding by a temporal-attention model;

passing the positional embedding to a separate multi-head spectral transform attention with a second multi-head shortcut (332);

fusing the output of the temporal attention with the output of the multi-head spectral transform by concatenation on the temporal dimension; and

wherein the fusion of the output is a linear layer in the embedding dimension and the output dimensionally is the same as the original contextual embedding.

6. The method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model (200) according to claim 4, wherein the inputting of the normalized data into a bidirectional spectral-based Transformer model (200) in the decoder, further comprising the steps of:

passing embedding from the contextual embedding of the encoder as input in the decoder;

passing through a first multi-head self-attention of the contextual embedding with a third multi-head shortcut (334) in the decoder to form a feature data;

diversifying the feature data by the third multi-head shortcut (334);

passing the extracted feature data to an addition and normalization layer;

passing through a second multi-head self-attention with a fourth multi-head shortcut (336) in the decoder;

diversifying the feature data by the fourth multi-head shortcut (336); and

passing the extracted feature data to an addition and normalization layer and followed by a feed forward network; and

flattening the output data and sending to a linear layer for obtaining a remaining useful life (108) for every time step in a sliding window.

7. The method for predicting the remaining useful life of an engine based on a bidirectional spectral-based Transformer model according to claim 4, 5 or 6, wherein the passing of the data through the multi-head shortcuts (330, 332, 334, 336) further comprising the step of inputting the positional embedding and contextual embedding into a multi-head spectral transform, a residual connection (400) and a linear layer without any activation function.

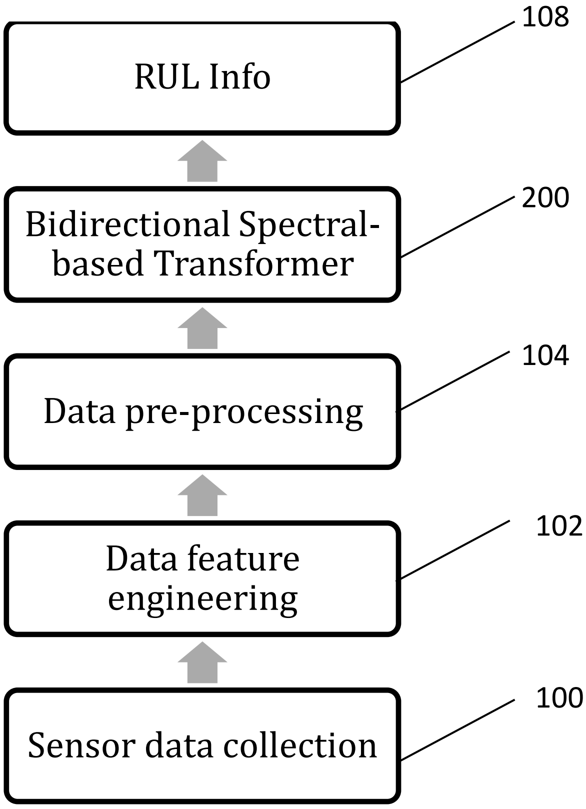
RUL Info
108
Bidirectional Spectral-based Transformer
200
Data pre-processing
104
Data feature engineering
102
Sensor data collection
100

FIG.1

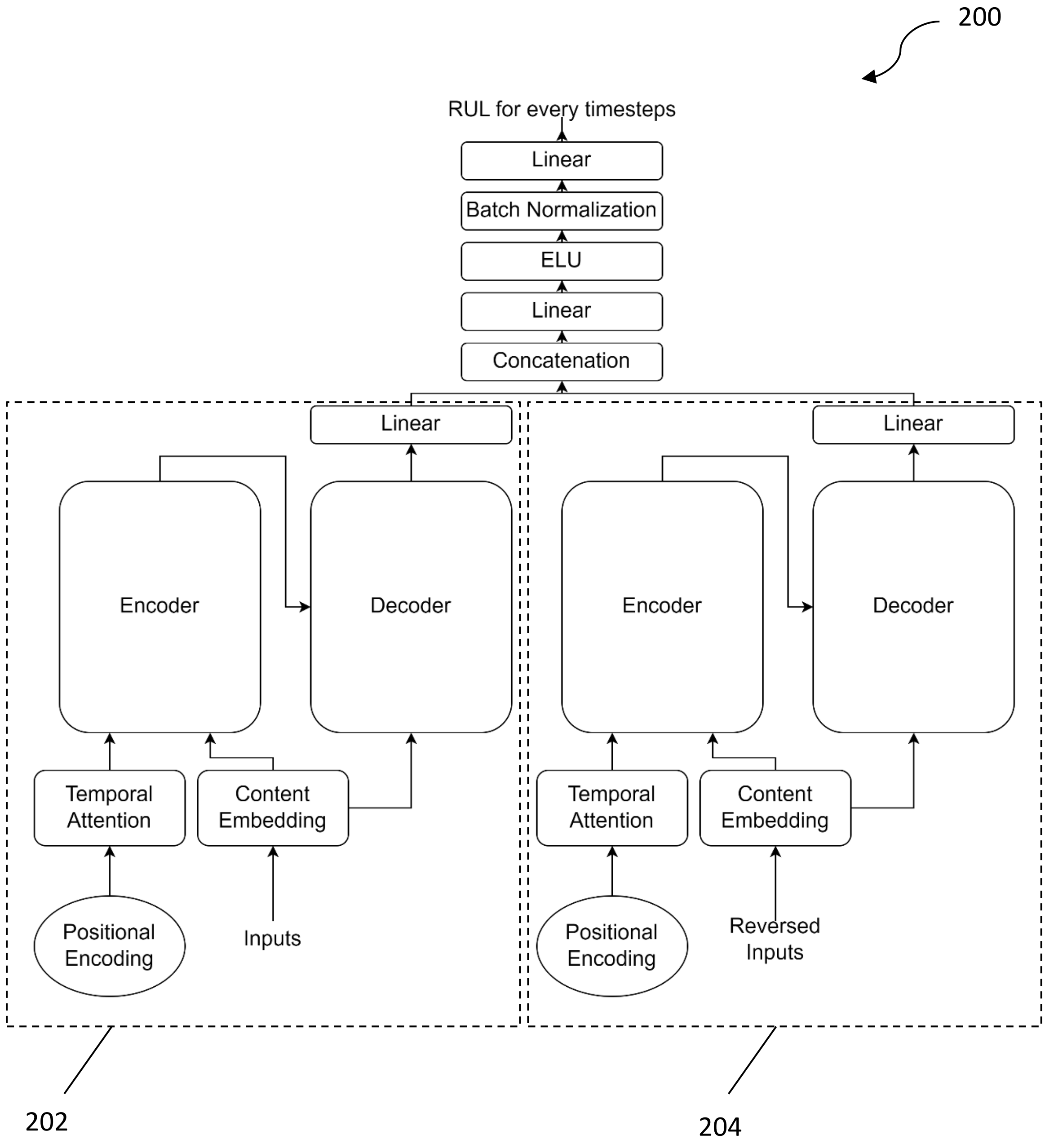
200
RUL for every timesteps
Linear
Batch Normalization
ELU
Linear
Concatenation
Linear
Encoder
Decoder
Temporal Attention
Content Embedding
Positional Encoding
Inputs
Linear
Encoder
Decoder
Temporal Attention
Content Embedding
Positional Encoding
Reversed Inputs
202
204

FIG.2

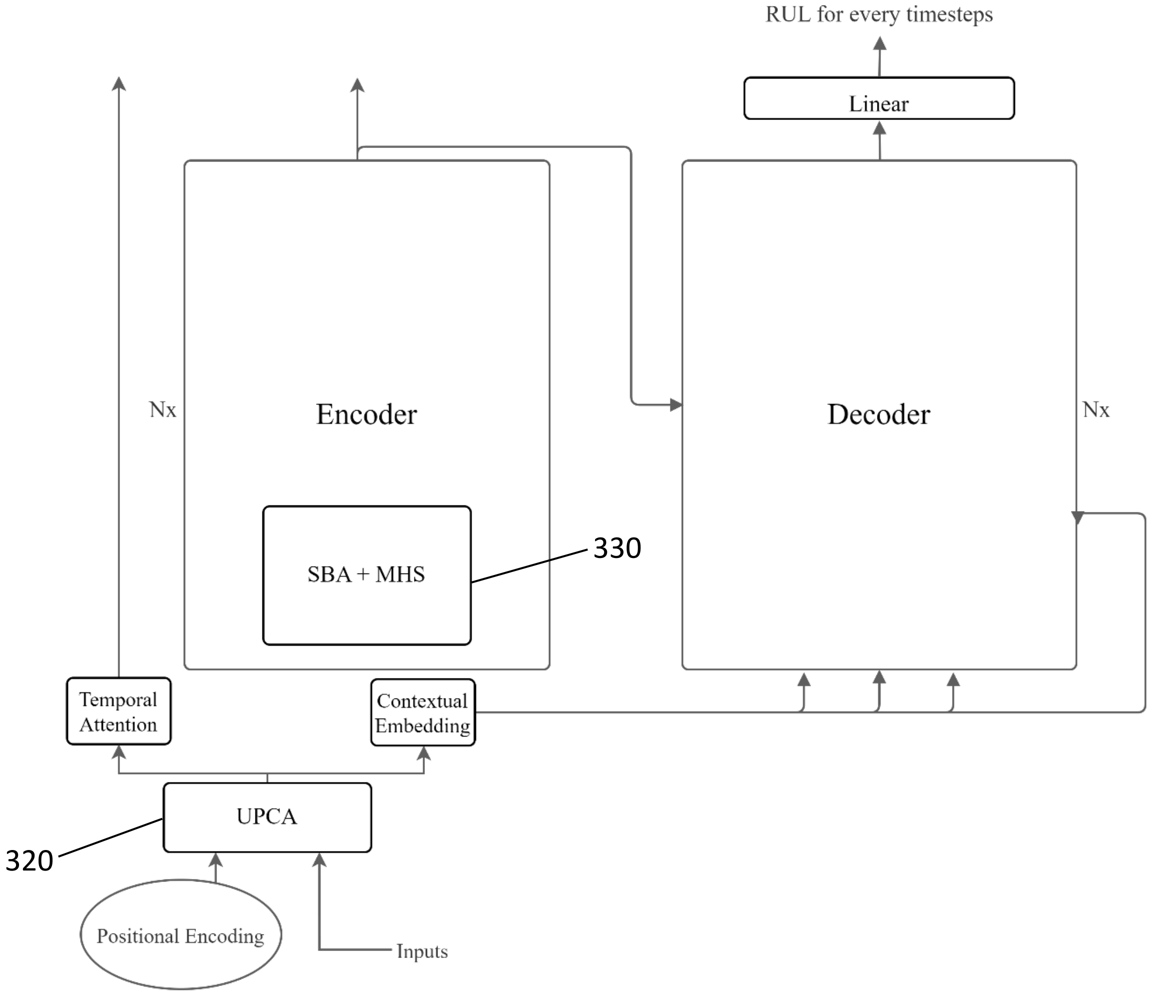
RUL for every timesteps
Linear
Nx
Encoder
SBA + MHS
330
Decoder
Nx
Temporal Attention
Contextual Embedding
UPCA
320
Positional Encoding
Inputs

FIG. 3a

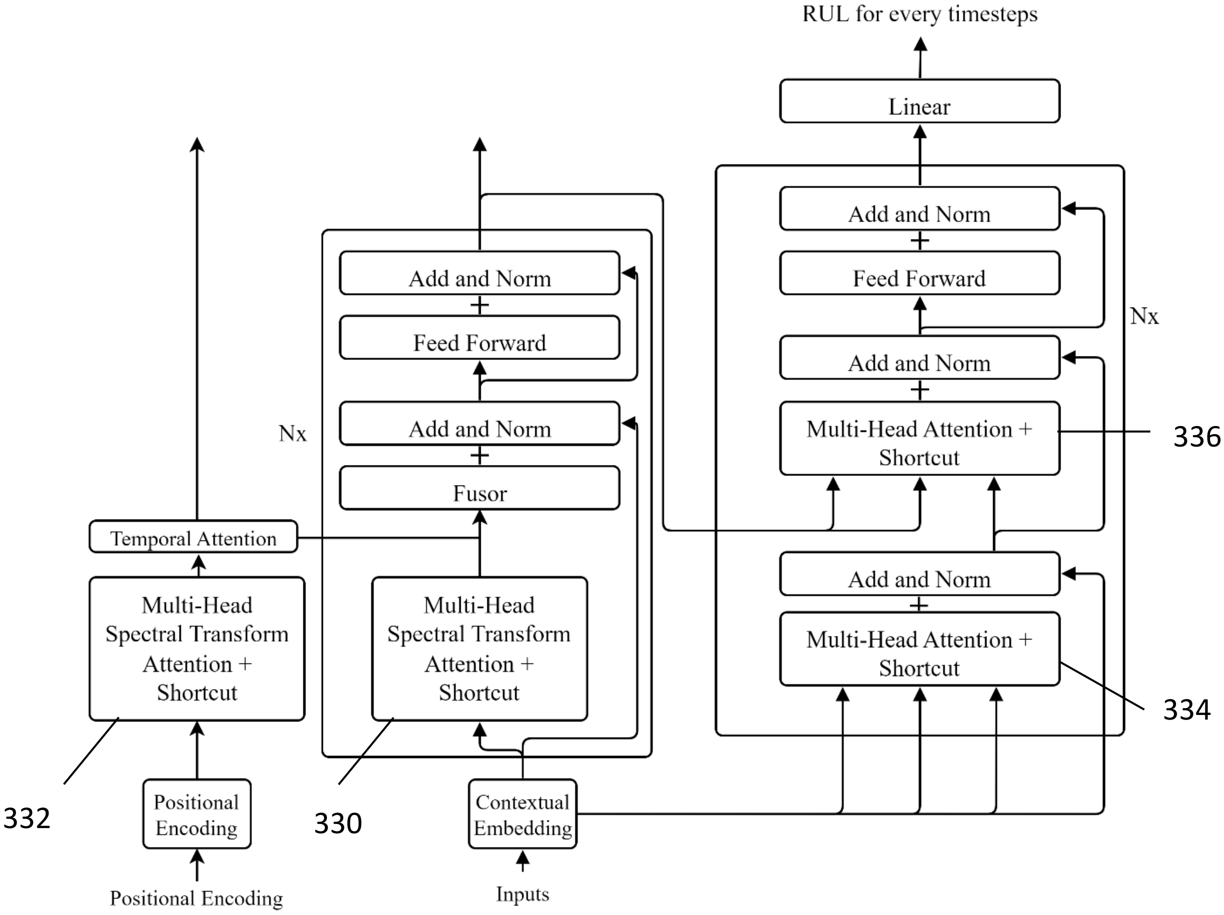
RUL for every timesteps
Linear
Add and Norm
Feed Forward
Add and Norm
Multi-Head Attention + Shortcut
Add and Norm
Multi-Head Attention + Shortcut
Nx
336
334
Add and Norm
Feed Forward
Add and Norm
Fusor
Nx
Temporal Attention
Multi-Head Spectral Transform Attention + Shortcut
Multi-Head Spectral Transform Attention + Shortcut
332
330
Positional Encoding
Contextual Embedding
Positional Encoding
Inputs

FIG. 3b

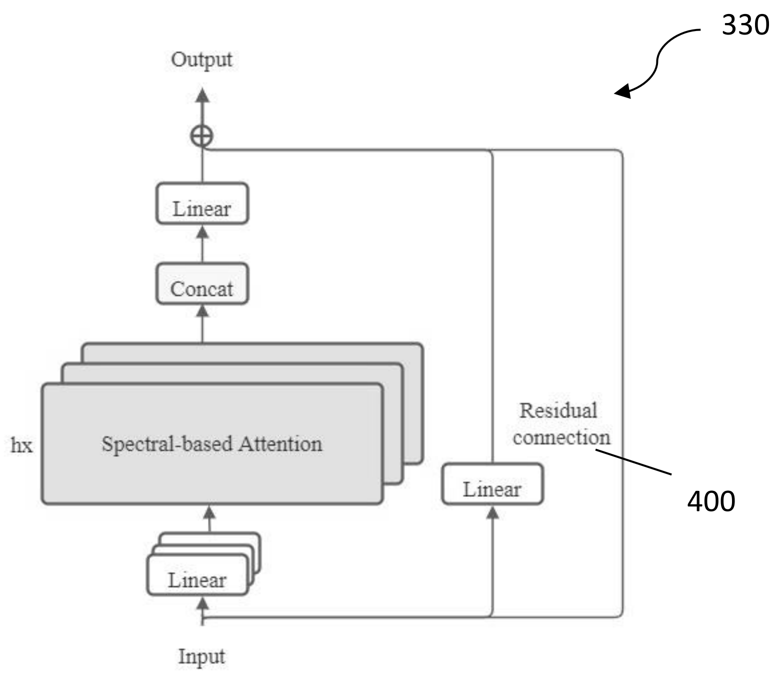
330
Output
Linear
Concat
hx
Spectral-based Attention
Residual
connection
Linear
400
Linear
Input

FIG. 4