EDRVASS

Fase 7

Team: TMEC
Sergio Sánchez Vallés
Álvaro Campillos Delgado

TABLE OF CONTENTS

01

Introduction

Team and methodology of work

Team members

Sergio Sánchez Vallés

Software Engineering (URJC)

Data science, software development, aerospace, DevOps.

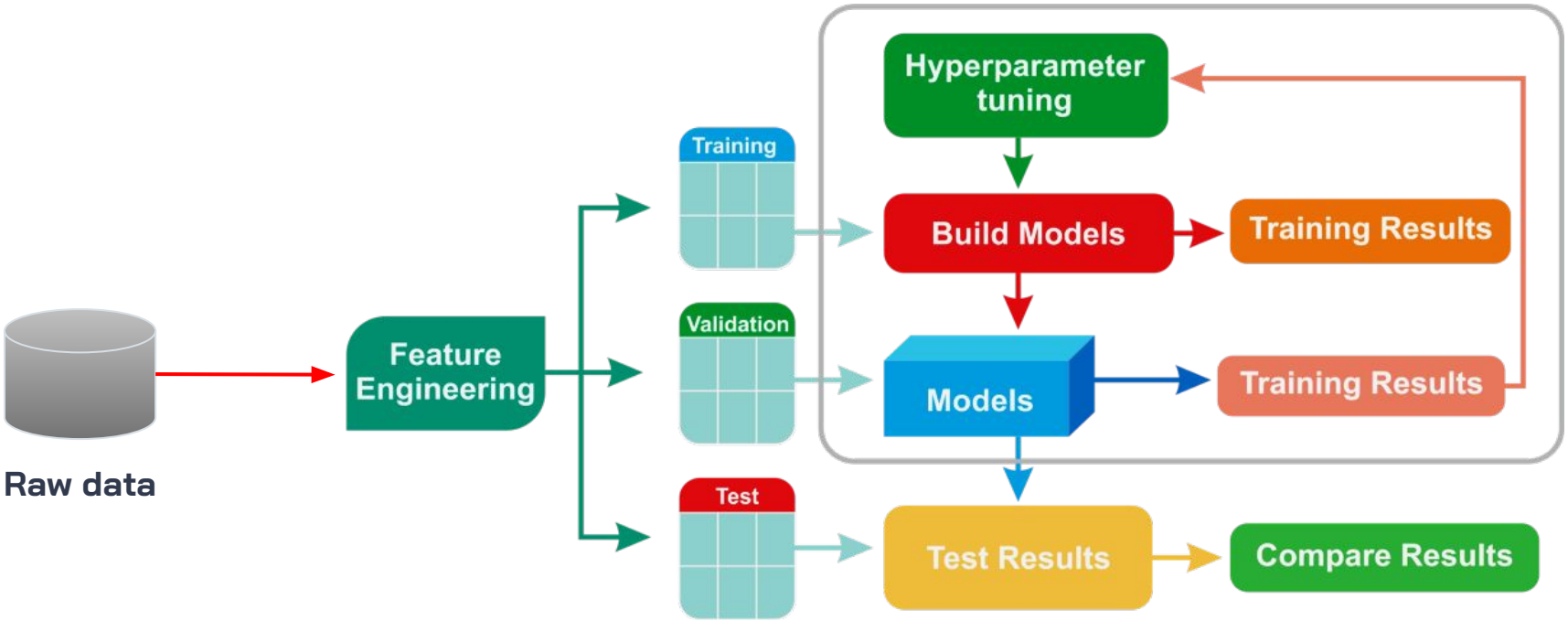
Álvaro Campillos Delgado

Computer Science (Universidad de Sevilla)

Data science, programming, math

Technology stack

Project workflow



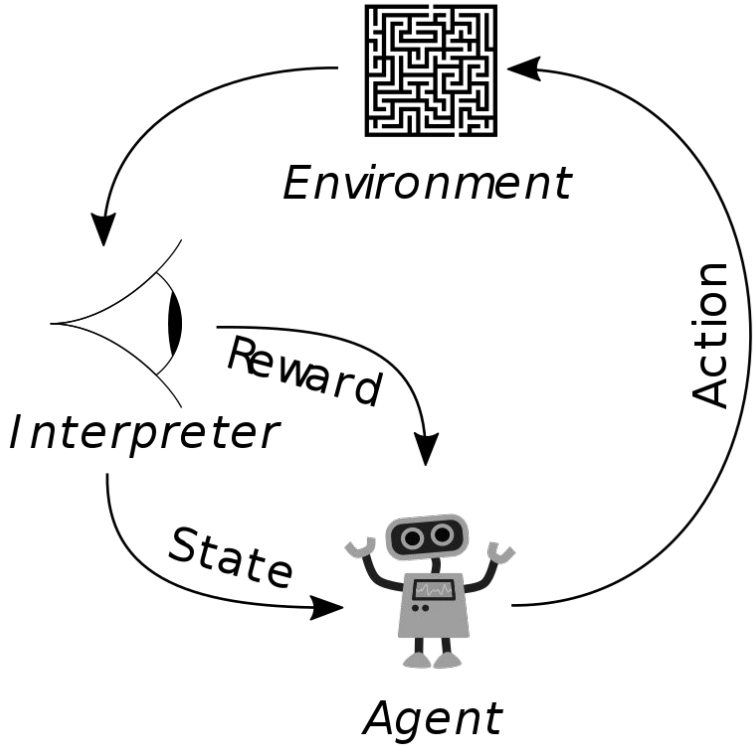
02

Initial approaches

RL and RNN

Reinforcement learning

- **State:**
  - x_pos, y_pos
  - x_vel, y_vel
  - angle
  - angle_vel
  - leg1, leg2
- **Actions:**
  - main_booster
  - lat_booster
- **Environment:**
  - gravity
  - wind_power
  - turbulence_power



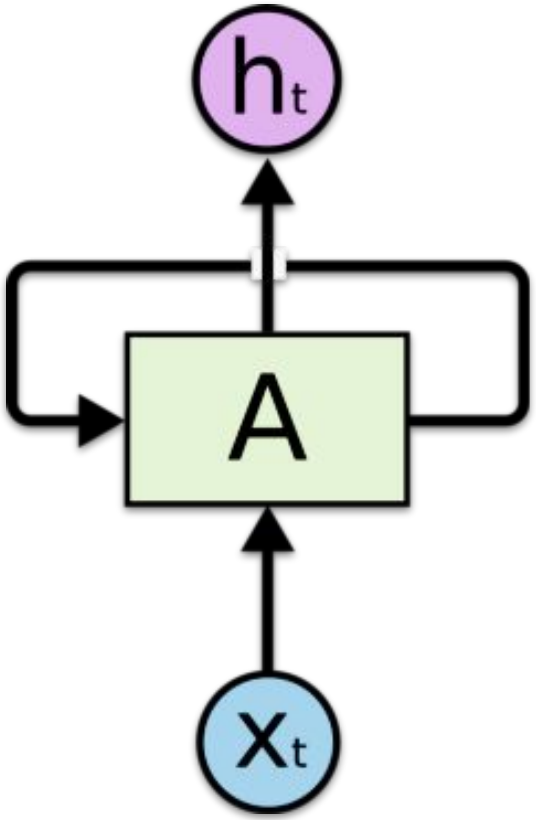
Recurrent neural network

Main **advantages** for the problem:

- Learns complex patterns in time series data (nonlinearity)
- Learns temporal dependence present in data

Main **disadvantages**:

- Vanishing/exploding gradient
- Weak memory
- Computationally expensive

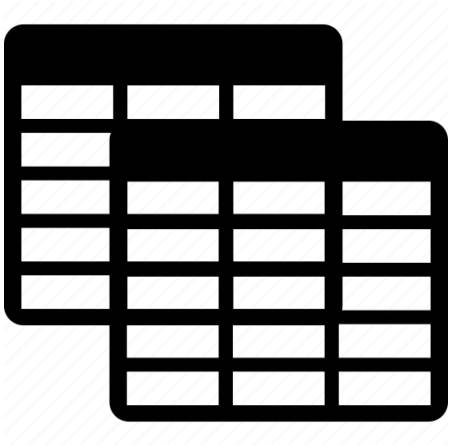


03

Exploratory data analysis

Statistics, correlation, missing values and visualization

Problem’s dataset



Experiments’ summaries

Experiments’ series

Experiments’ summary

filename	total_timesteps	gravity	wind_power	turbulence_power	efficiency
experiment_1.csv	185	-5.375066	14.337393	1.902789	415.69
experiment_2.csv	428	-3.958084	16.376438	1.697932	104.37
experiment_4.csv	193	-3.640152	1.614948	0.192214	411.14
experiment_5.csv	166	0.000000	0.000000	0.242178	414.93
experiment_6.csv	172	-5.289123	15.771525	1.453806	373.85

experiments_summary_train.xlsx

Experiments' series

x_pos	y_pos	x_vel	y_vel	angle	ang_vel	leg_1	leg_2	main_booster	lat_booster
-0.001548	1.403302	-1.571687e-01	-3.385856e-01	0.002068	4.082241e-02	0.0	0.0	-0.327187	0.481895
-0.003124	1.395385	-1.598507e-01	-3.518934e-01	0.004395	4.654207e-02	0.0	0.0	-0.341337	0.935360
-0.004636	1.387138	-1.516633e-01	-3.665214e-01	0.004803	8.163137e-03	0.0	0.0	0.265217	0.893026
-0.006214	1.379139	-1.566717e-01	-3.554864e-01	0.003728	-2.151341e-02	0.0	0.0	-0.214743	0.577881
-0.007766	1.370820	-1.531046e-01	-3.697289e-01	0.001665	-4.125303e-02	0.0	0.0	-0.151192	0.531269

experiment_1.csv

04

Data processing

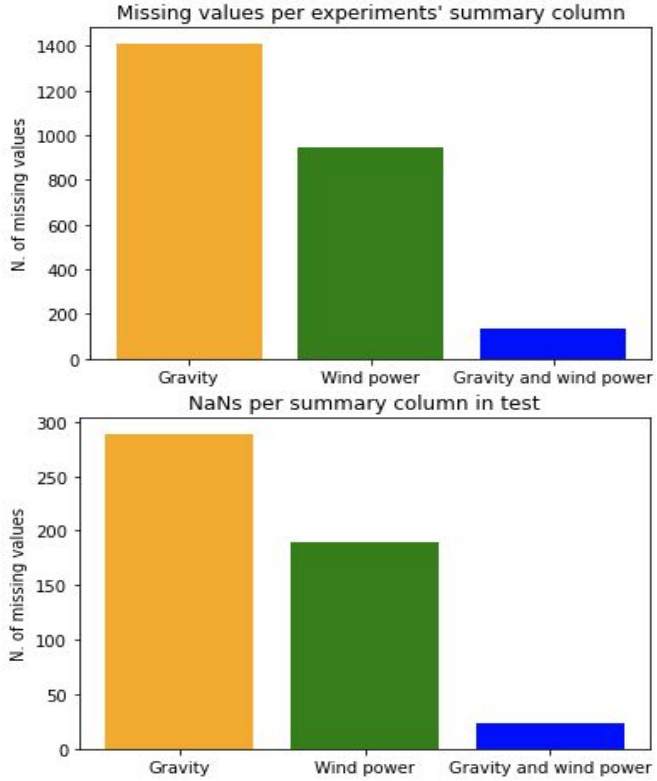
Missing values imputation, padding, scaling and reshaping

Data imputation

Missing values in experiments' summary

Columns gravity and wind_power can have missing values registered as 0.

These columns can have missing values simultaneously.

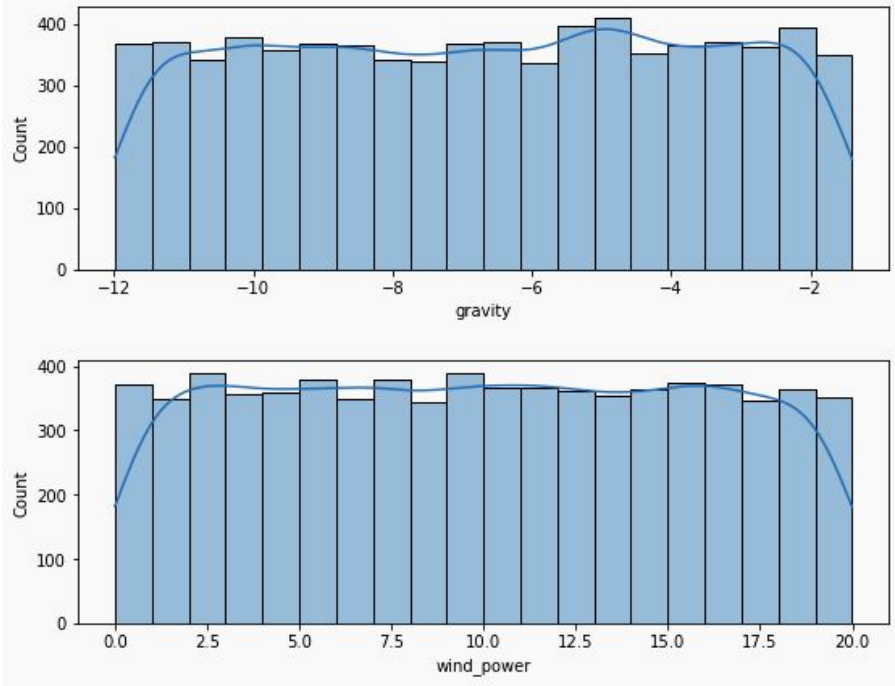


Experiments' summary after removing missing values

Experiments' summary

Imputation methods:

- K-nearest neighbors
- Median
- Average
- Bayesian Ridge



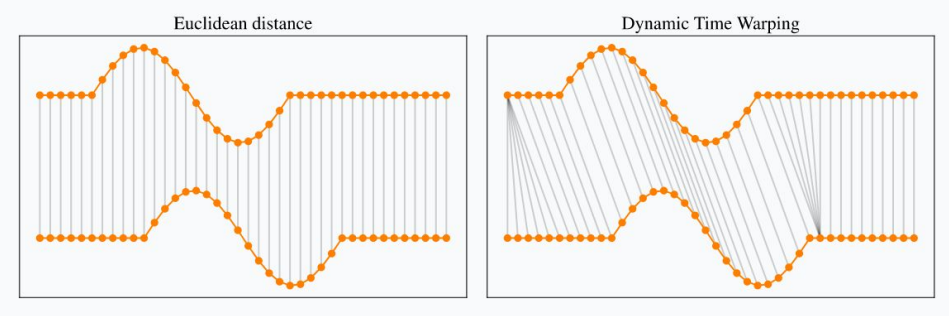
Experiments' summary after removing missing values

Imputation approaches

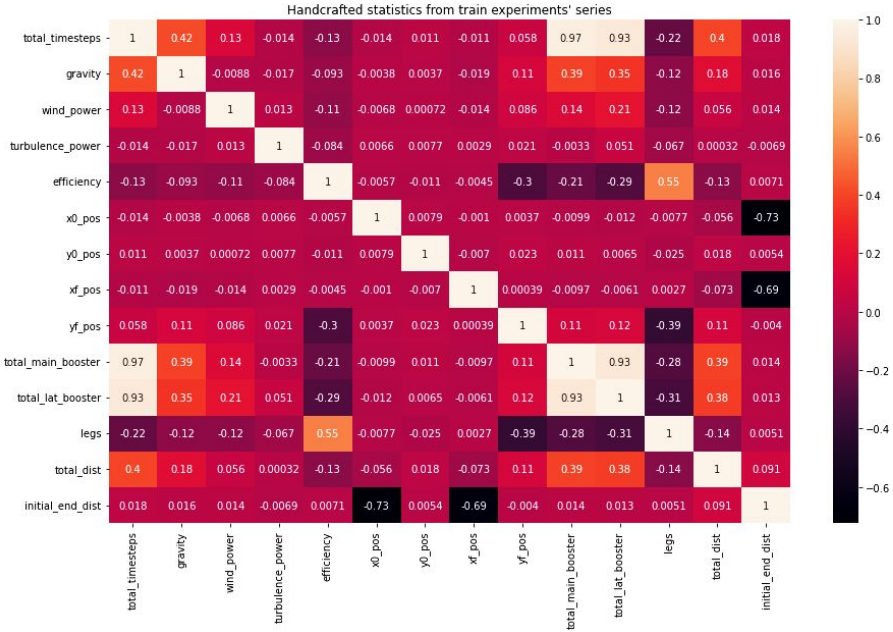
Imputation using summaries and series

Find similarity between experiments based on their time series to impute missing values. Two approaches:

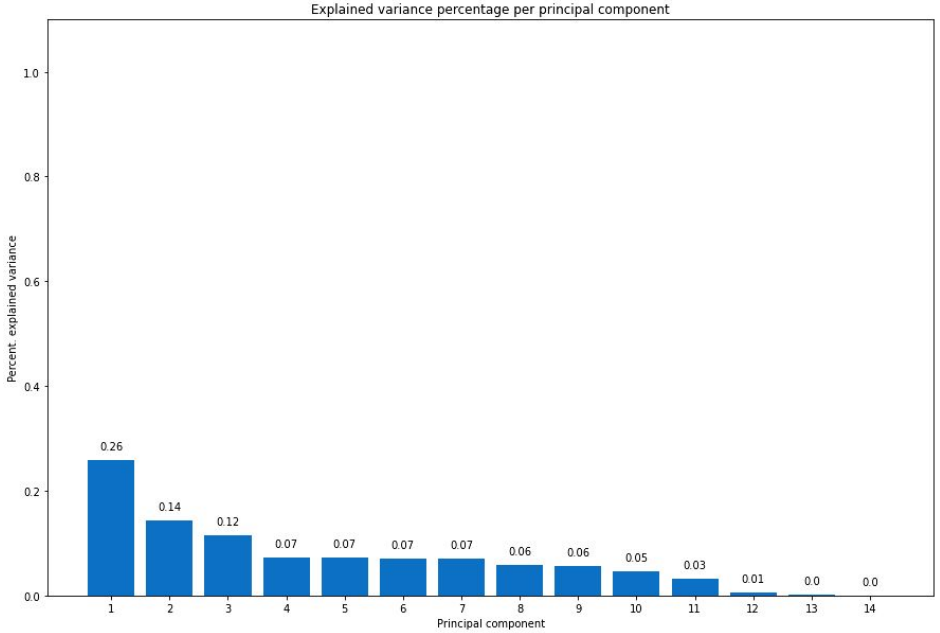
- Similarity between series **using each time serie entirely** (requires finding an appropriate **distance measure** for time series)
- Compute **statistics** from them and consider the **distance between these** (less complexity, computation already done for EDA, requires finding appropriate statistics that summarize the series, allows hand-crafted statistics)



Handcrafted stats analysis from series



Correlation plot for static experiments' summary variables



PCA explained variance for static experiments' summary variables

Comparative using time series statistics

	Gravity MAE	Wind power MAE
K-nearest neighbors	1.6530	4.6752
Median	2.8344	5.0531
Average	2.6918	4.9913
Bayesian Ridge	2.2202	4.7954

MAE without using time series statistics for imputation (train)

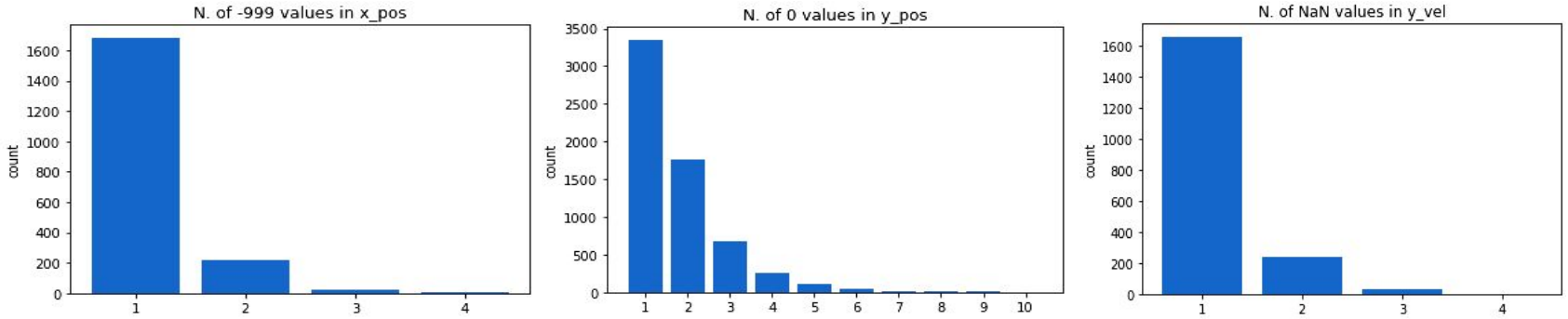
	Gravity MAE	Wind power MAE
K-nearest neighbors	1.4806	4.2995
Median	2.7153	4.9973
Average	2.7673	5.1631
Bayesian Ridge	1.2339	3.6660

MAE using time series statistics for imputation (train)

Missing values in experiments' series

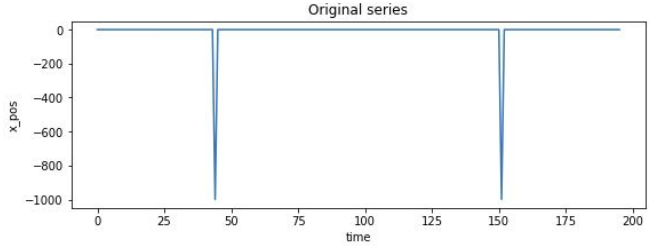
	NaNs	-999	0
x_pos	✗	✓	✗
y_pos	✗	✗	✓
y_vel	✓	✗	✗

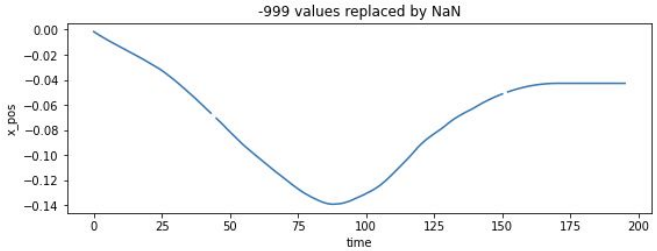
Type of missing value for x_pos, y_pos and y_vel columns in time series



Missing values distribution in x_pos, y_pos and y_vel columns of train time series

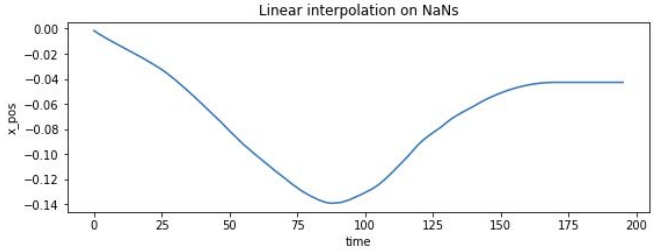
x_pos column

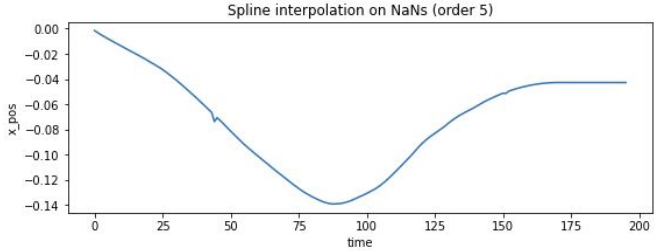


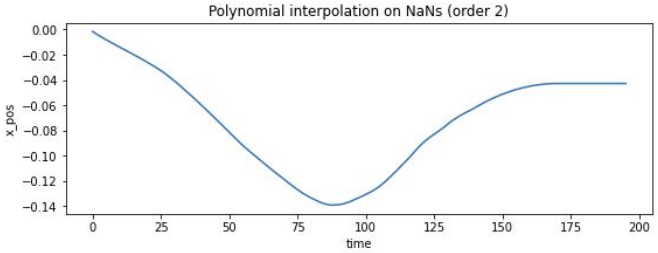


experiment_2274.csv

Interpolation

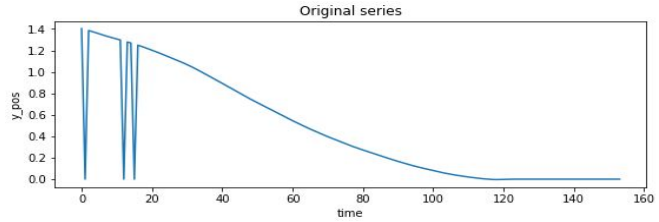


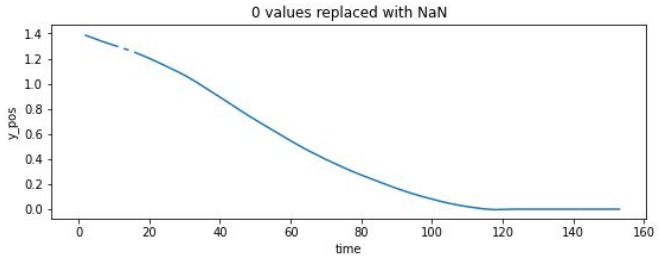




Missing values imputation with different interpolation methods

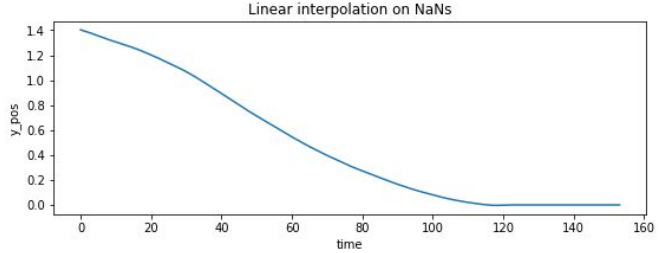
y_pos column

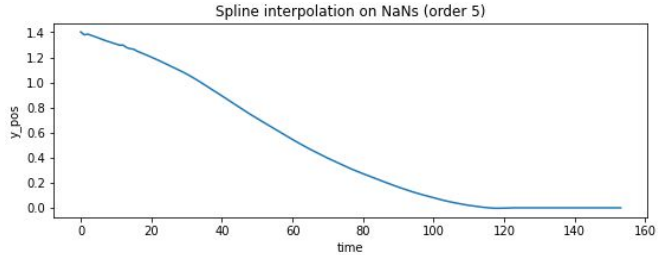


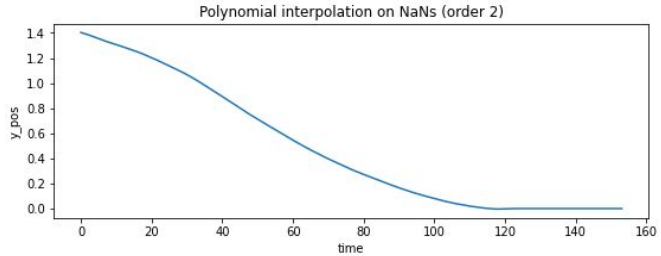


experiment_7503.csv

Interpolation

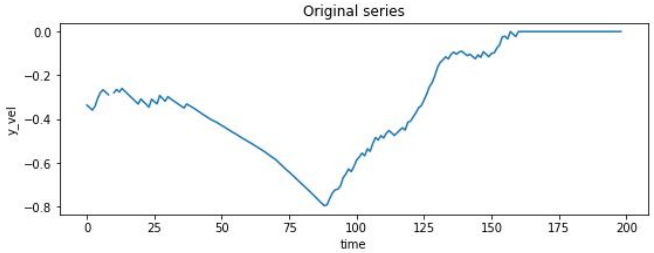






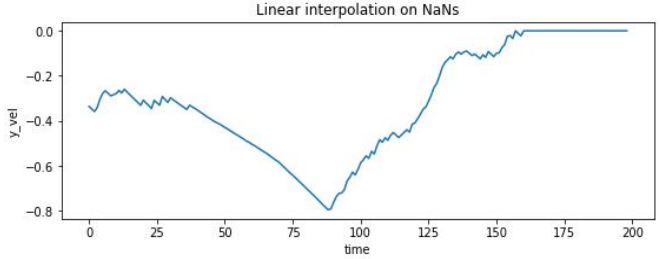
Missing values imputation with different interpolation methods

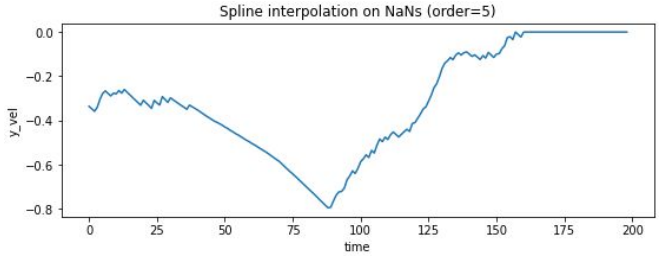
y_vel column

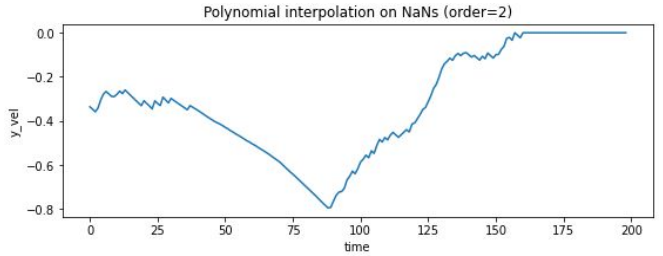


experiment_7956.csv

Interpolation

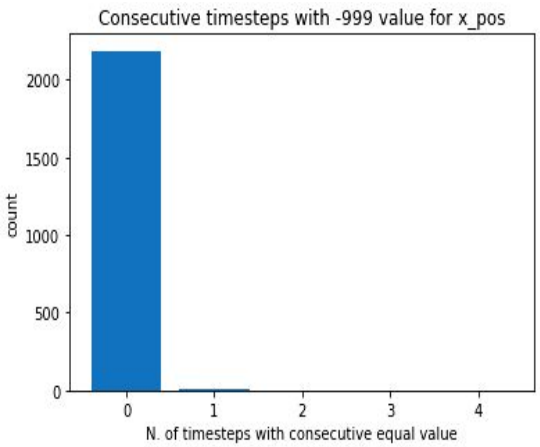


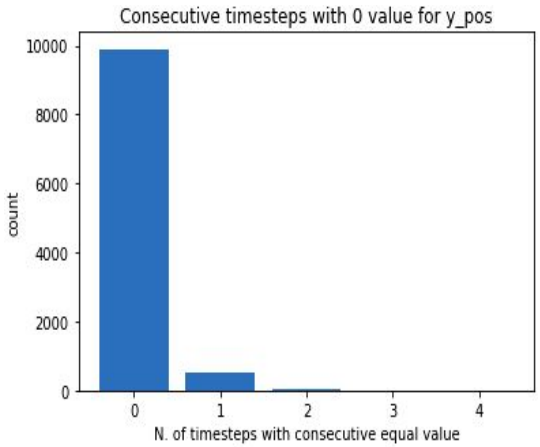


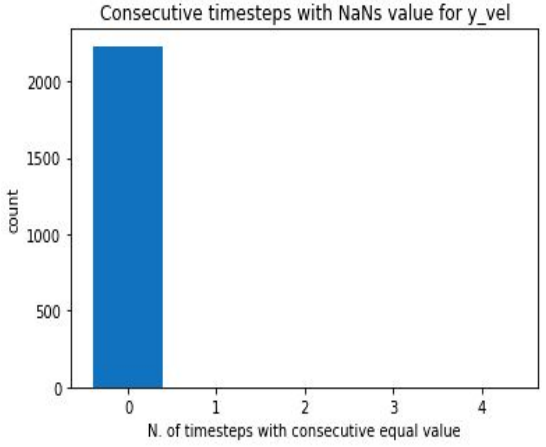


Missing values imputation with different interpolation methods

Missing values distribution in series





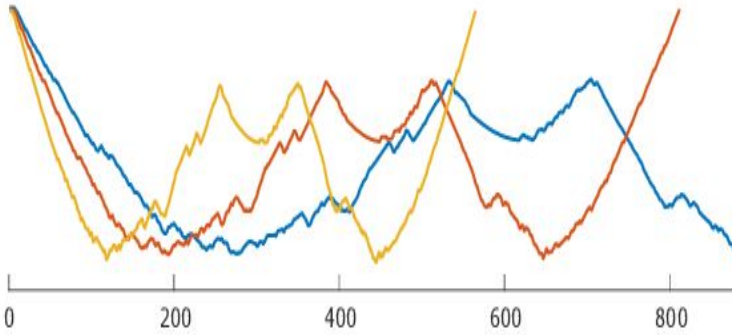


Distribution of equal consecutive missing values for columns x_pos, y_pos and y_vel in train time series data

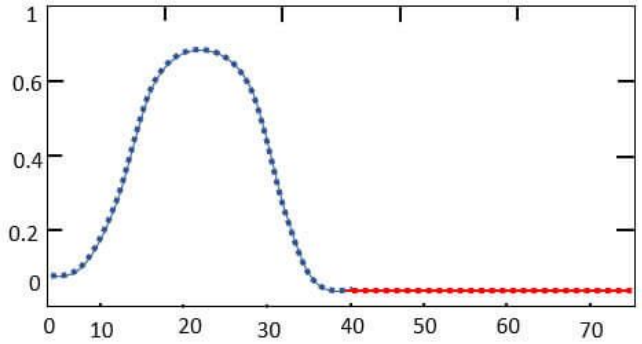
Scaling

Padding

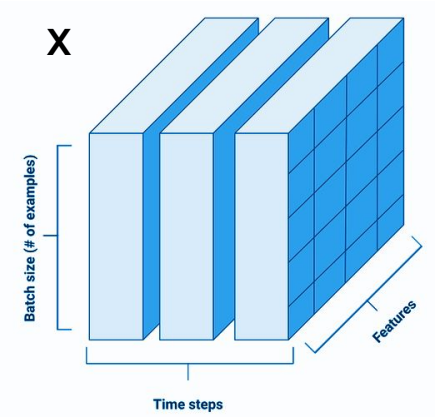
Variable-length time series

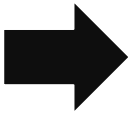


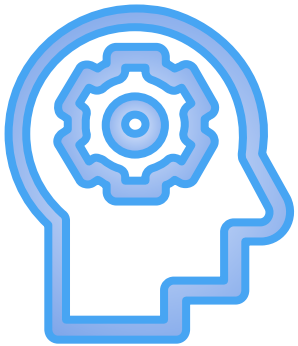
Zero padding

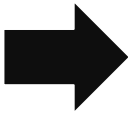


Reshaping









y

05

Models

Missing values imputation, padding, scaling and reshaping

Features

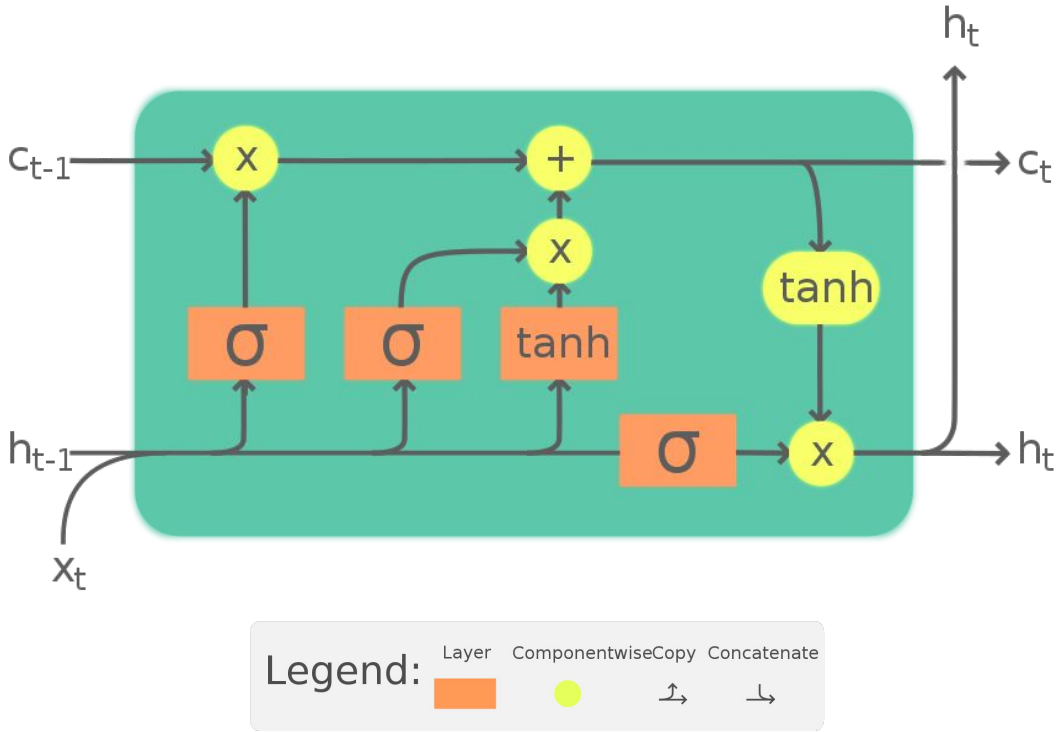
Static and time dependent data

Original time series columns

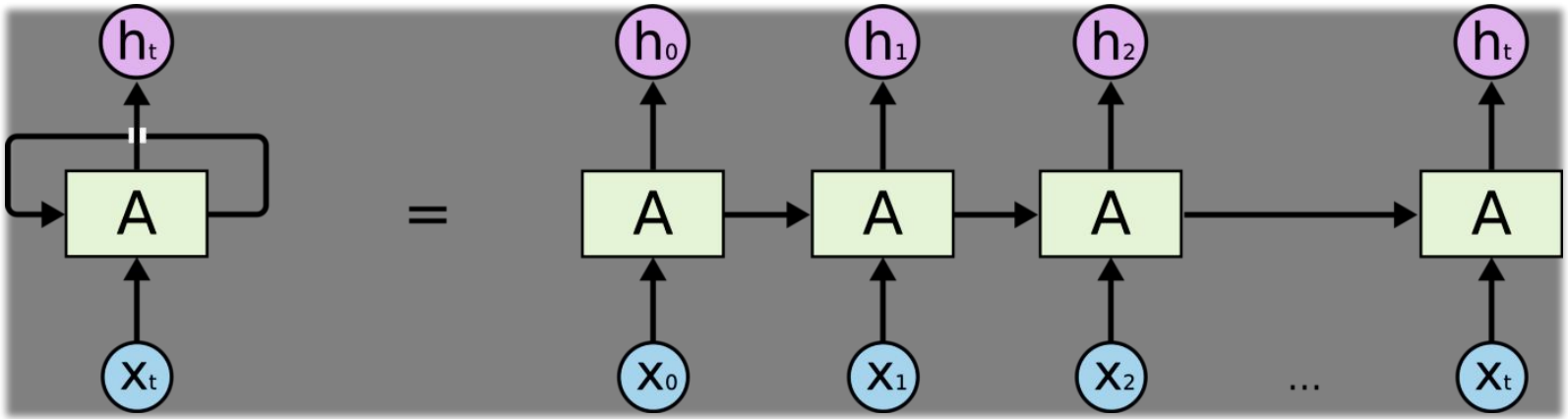
Original experiments columns
+
time series describe

Models

Long short-term memory (LSTM)

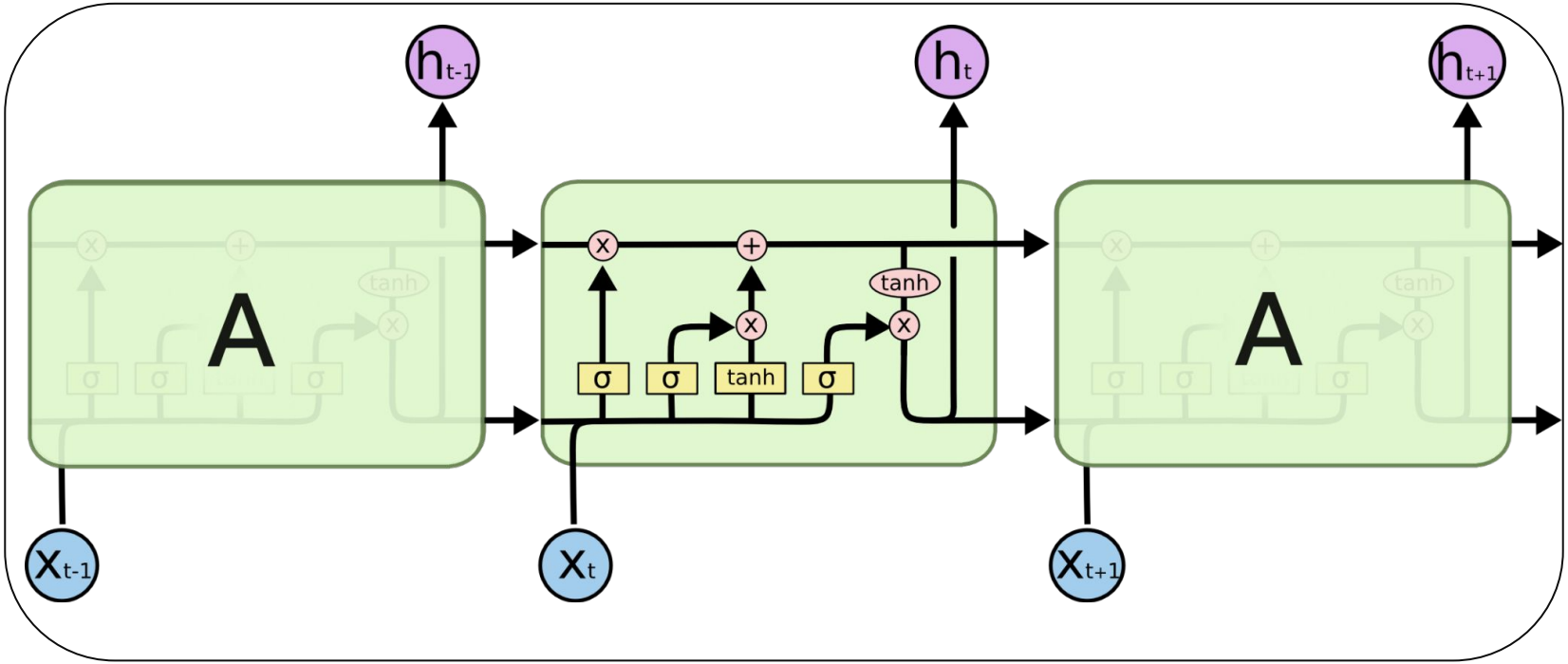


LSTM with several timesteps

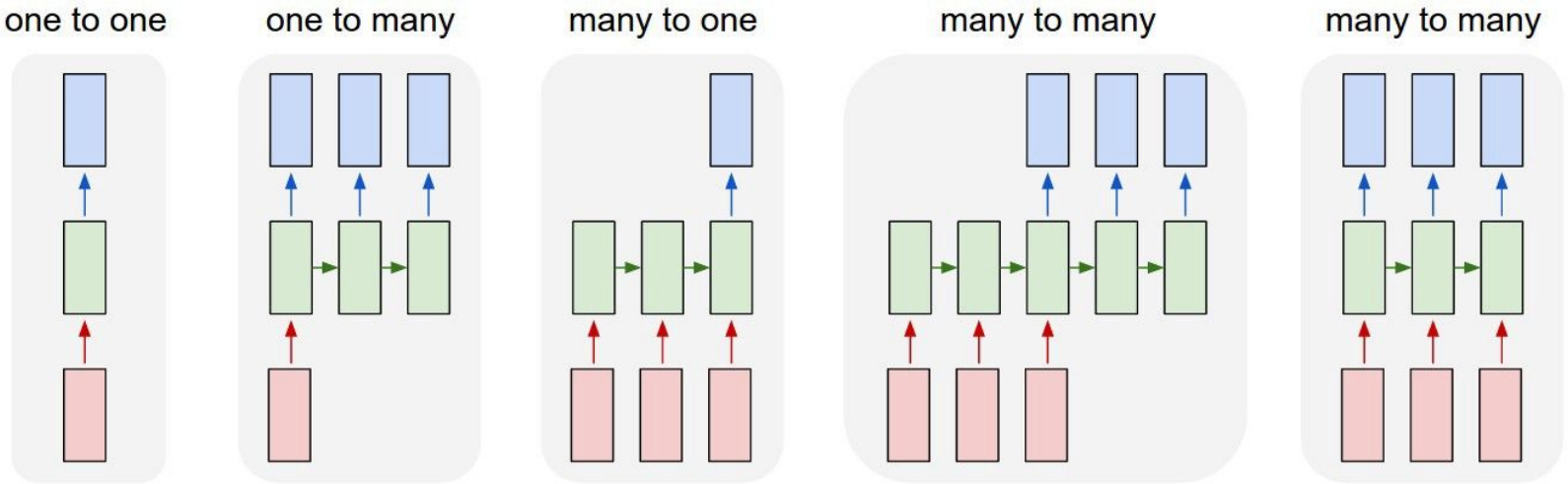


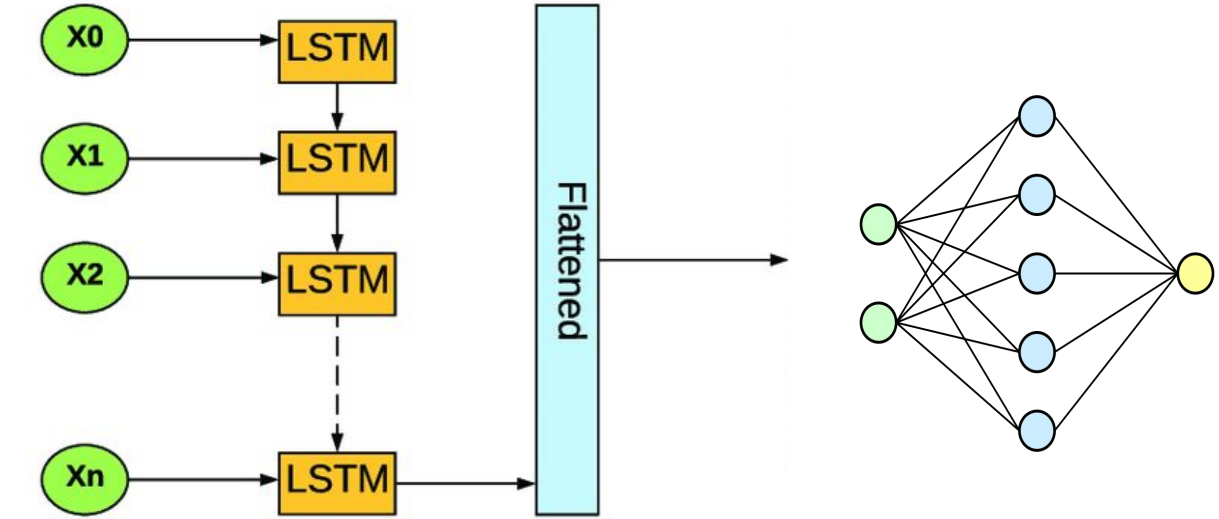
RNN unfolding from standard RNN for t timesteps

LSTM with multiple units

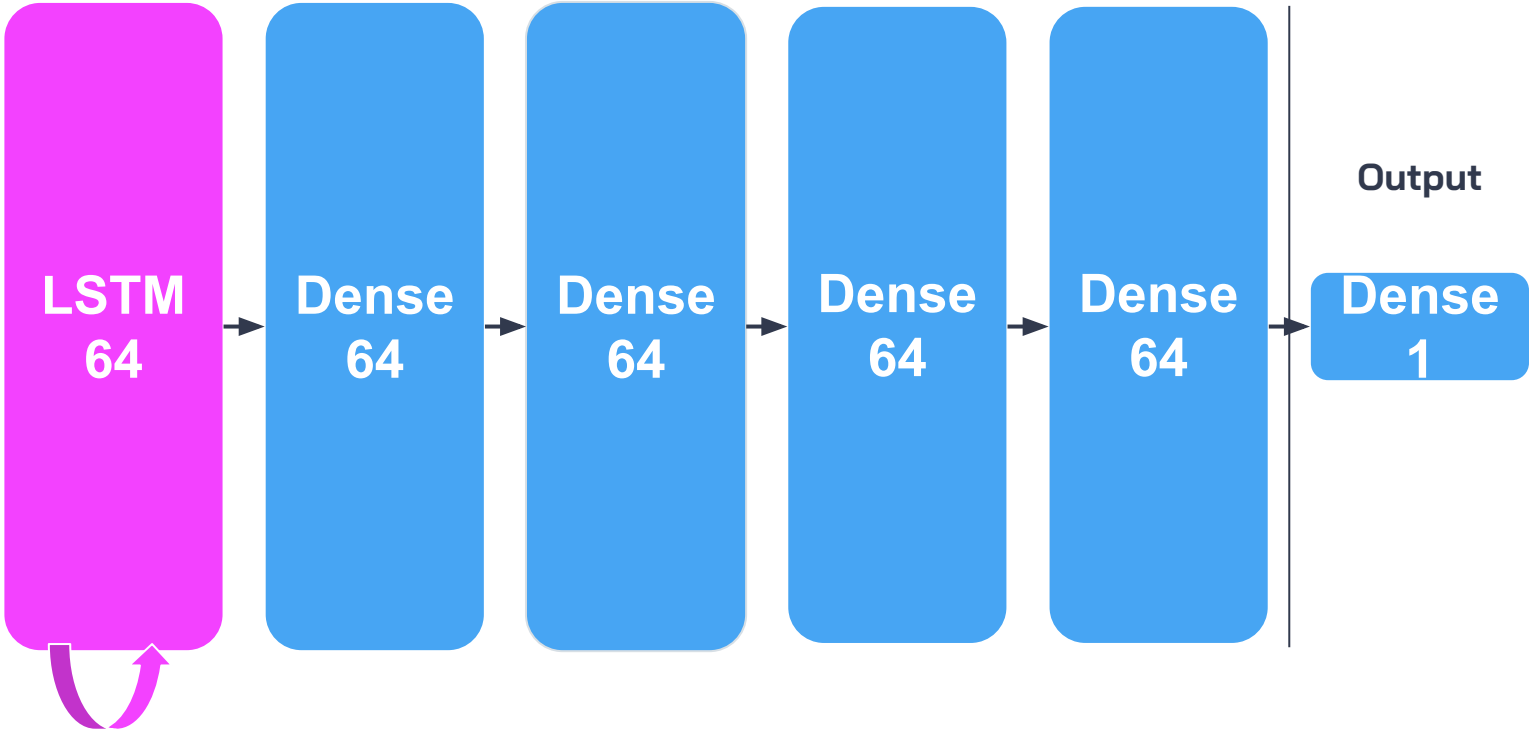


Connection between LSTM units

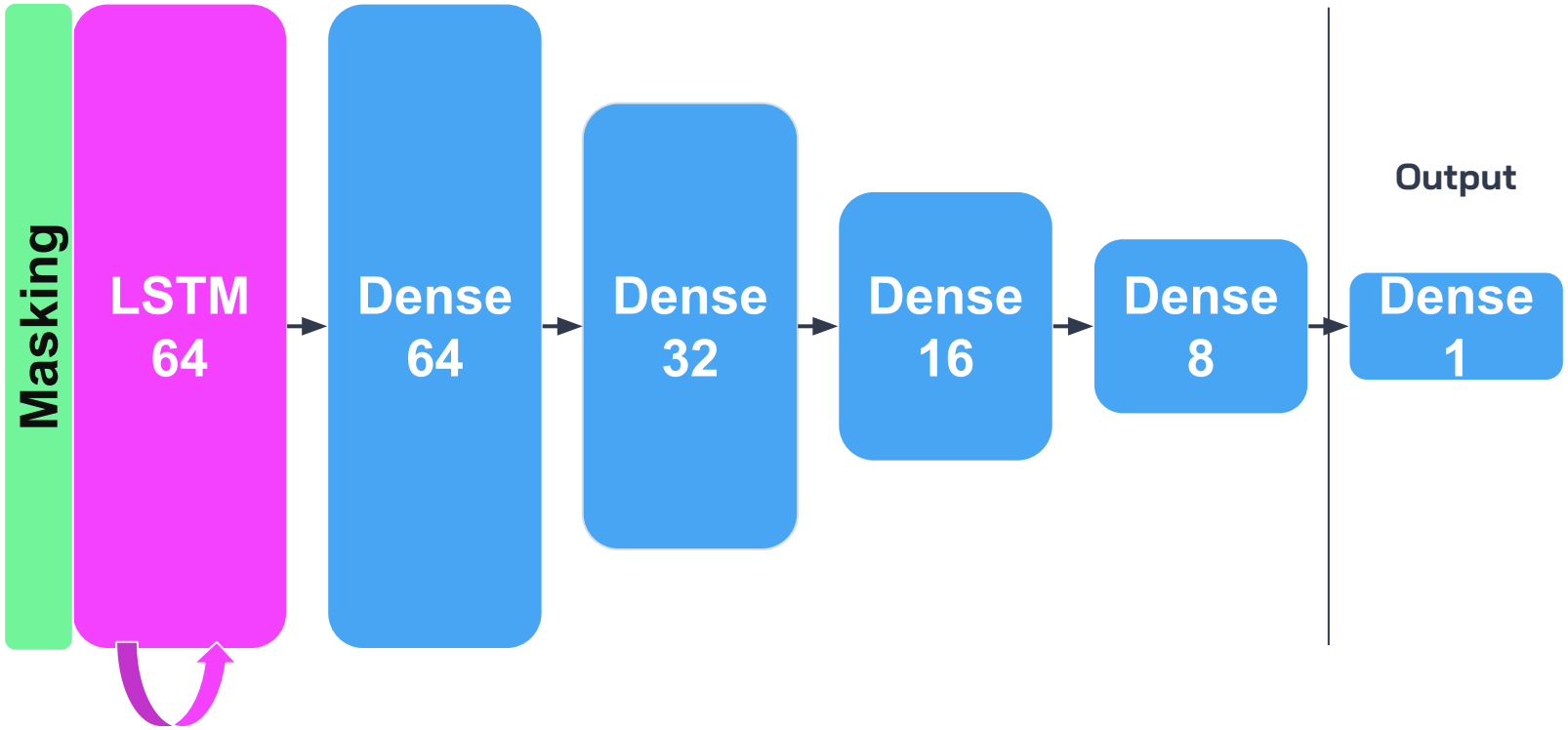

one to one
one to many
many to one
many to many
many to many


X0
X1
X2
Xn
LSTM
LSTM
LSTM
LSTM
Flattened

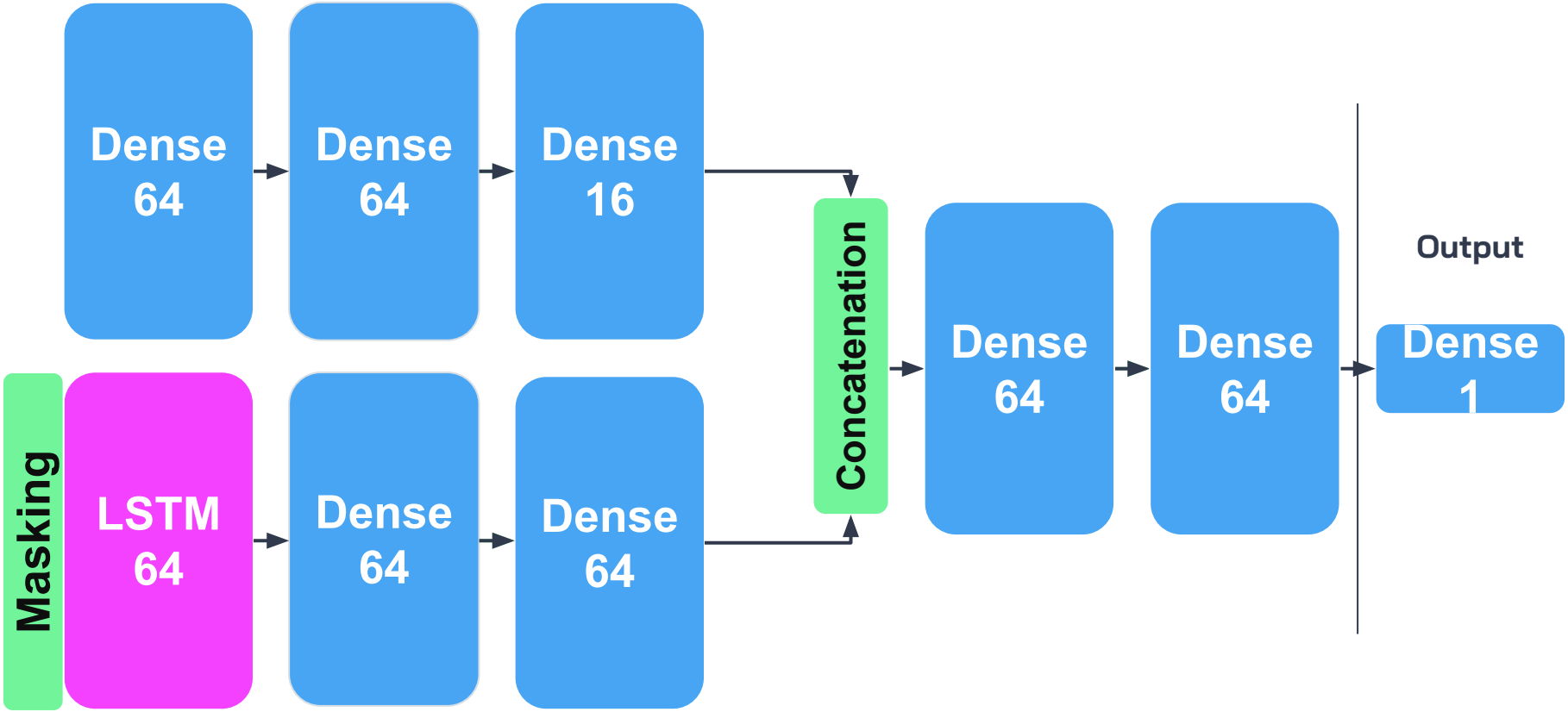
Baseline model



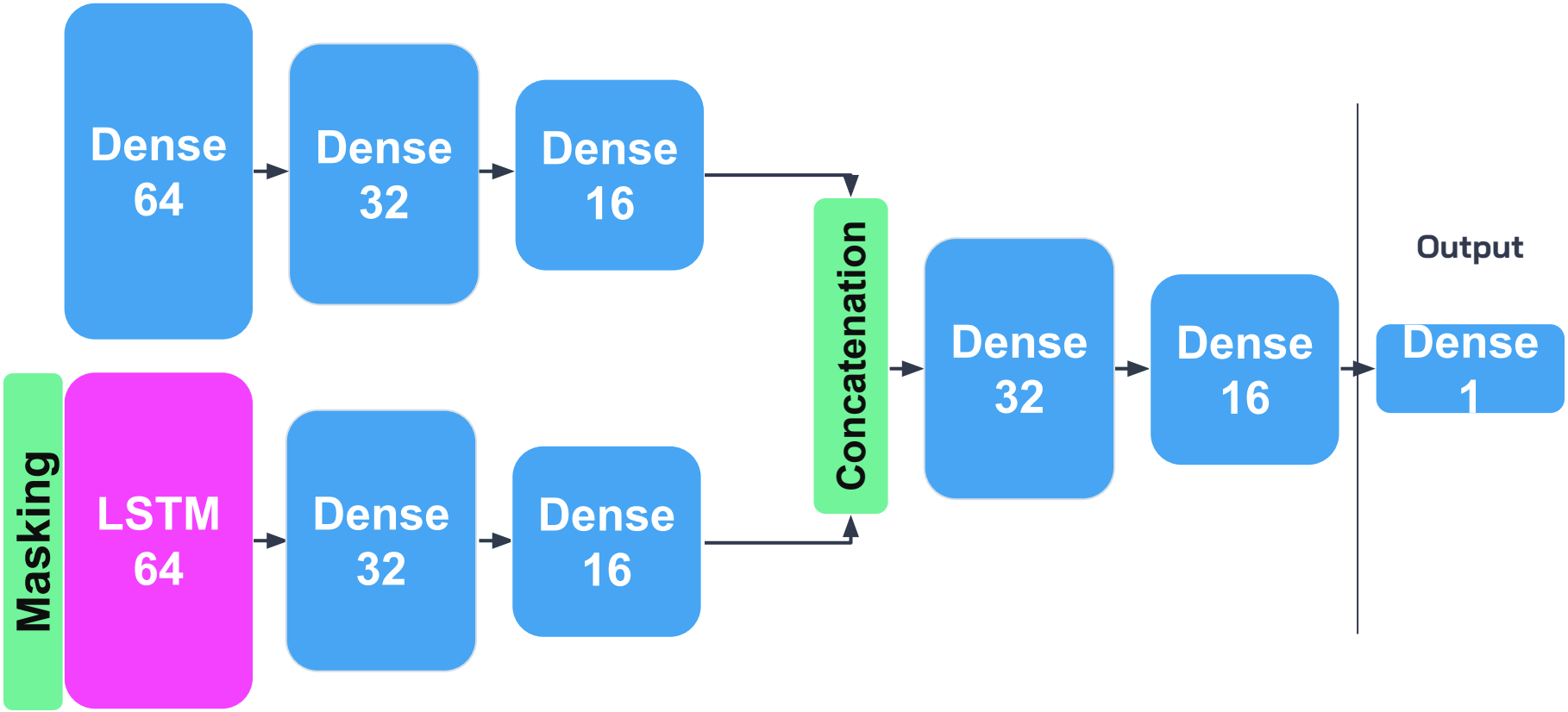
Optimized baseline model



Refined model architecture



Optimized model architecture



Layers

Units

GRU

Dropout

Early Stopping

ReLU

Bi-Directional

Shuffle fold

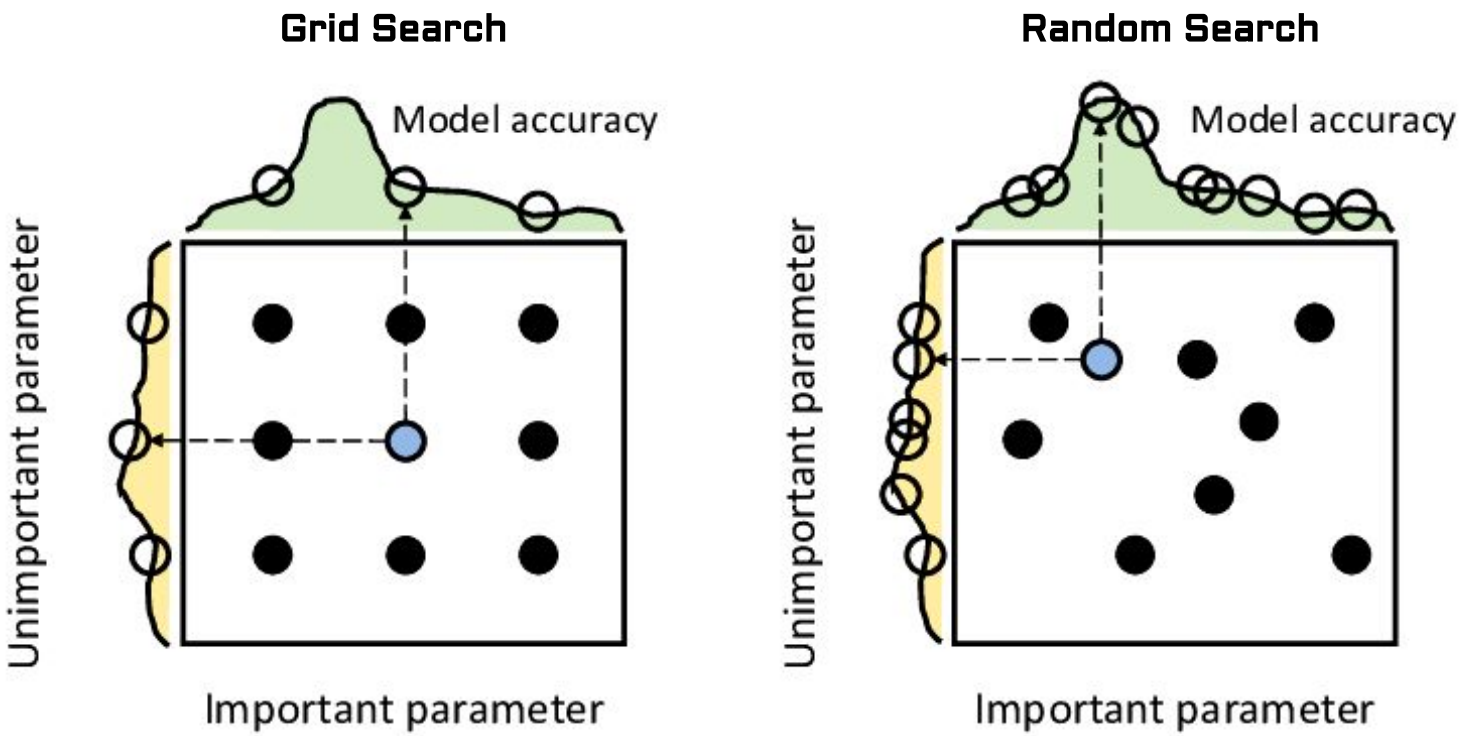
Batch size

Learning rate

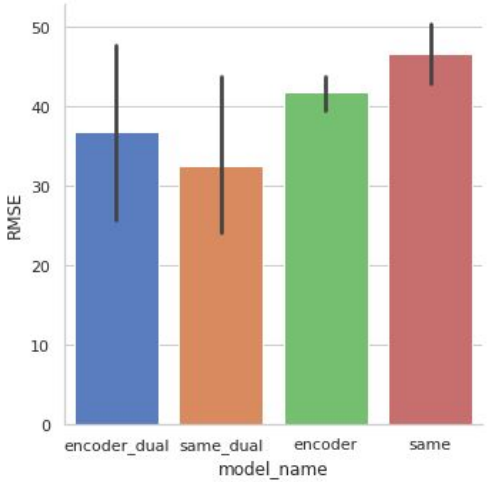
Weight initialization

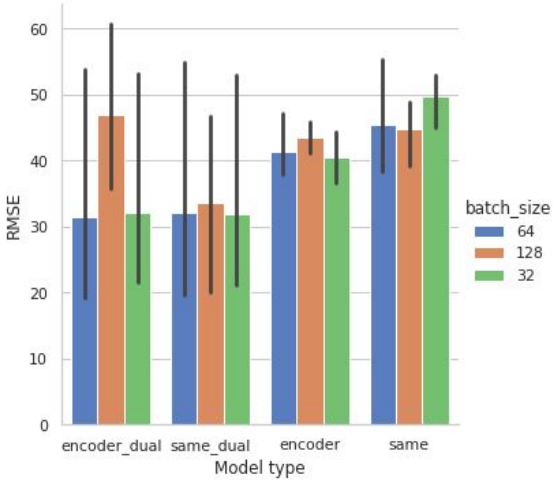
Hyperparameter optimization

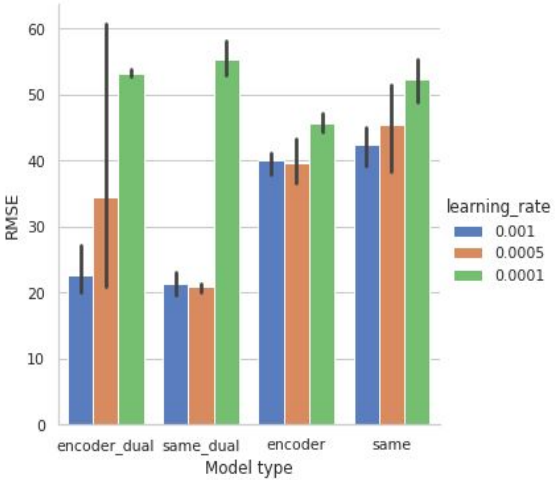
K Tuner


Grid Search
Model accuracy
Unimportant parameter
Important parameter
Random Search
Model accuracy
Unimportant parameter
Important parameter

GridSearch results





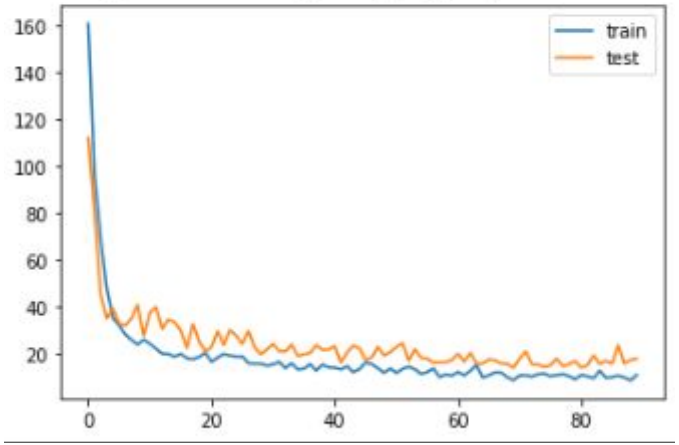


Top 10 models GridSearch

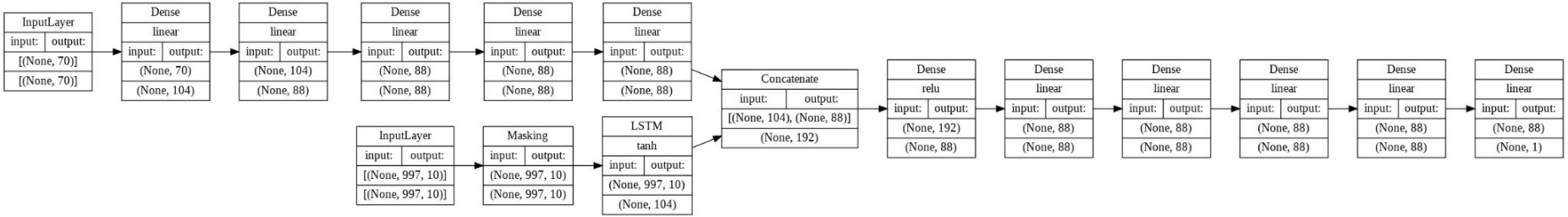
RMSE	model_name	batch_size	learning_rate	epochs
19.2808	encoder_dual	64	0.001	50
19.7086	same_dual	64	0.001	46
19.9849	same_dual	128	0.0005	99
21.0189	encoder_dual	64	0.0005	102
21.1027	same_dual	32	0.0005	42
21.4186	same_dual	64	0.0005	63
21.5796	encoder_dual	32	0.0005	105
21.5999	encoder_dual	32	0.001	54
21.7104	same_dual	32	0.001	39
22.9838	same_dual	128	0.001	42

Batch_size = [32, 64, 128] *Learning_rate = [0.001, 0.0005, 0.0001)*

Top model RandomSearch



Score: 18.5575



06

Future work

Other machine learning models

More hyperparameter tuning

Transformers

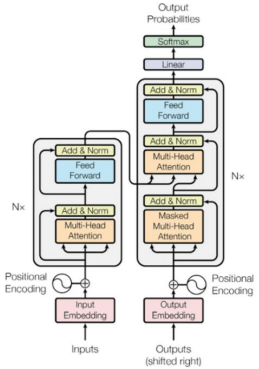
XGBoost

Auto-Sklearn

K Tuner

Transformer

Attention Is All You Need



Many thanks

sergio-sanchez-valles

alvarocampillosdelgado

SergioSV96

acampillos