Quality of Service Forecasting with LSTM Neural Network

International Symposium on Integrated Network Management (IM 2019)

May 11, 2019

Tomas Jirsik, Stepan Trcka, Pavel Celeda

Institute of Computer Science, Masaryk University, Brno

CSIRT-MU

Quality of Service Forecasting
what is it good for?

Quality of Service

- Abstract term used for comparing services
- Derived from measurable QoS attributes
- QoS Attributes
  - Application response time
  - Network response time

Applications

- Recommending systems for Web Pages

Forecasting

- Updates from service providers are sparse

Challenges

what do we research

How can be QoS attributes collected?

- Increase the frequency of the QoS attributes updates

How can we use Long Short-Term Memory Neural Network for QoS forecasting?

- How to create LSTM NN model?

What method should we use for QoS attribute forecasting?

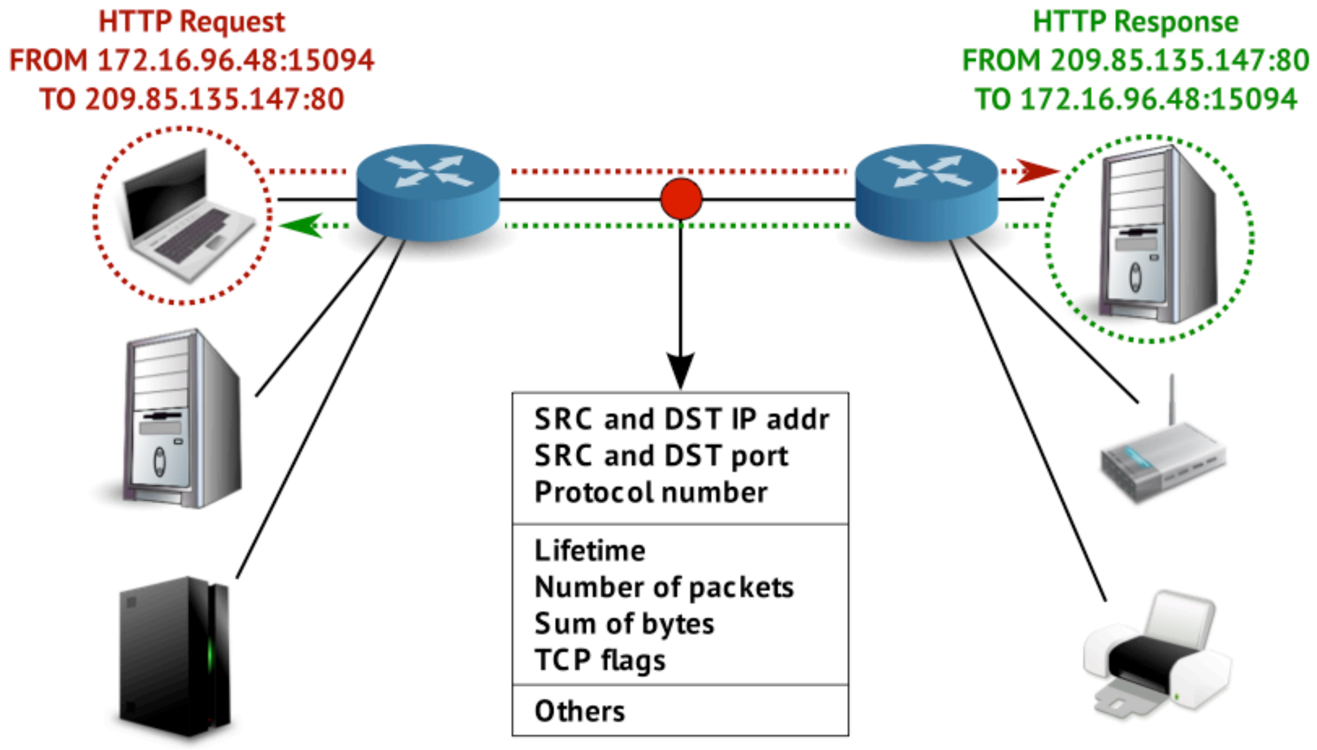
- Forecast precision
- Estimation time

Centralized QoS Attribute Collection

how to collect up-to-date data

IP flow network monitoring

- Passive approach to network traffic observation



Centralized QoS Attribute Collection

how to collect up-to-date data

Next-generation IP flow network monitoring

- Bi-flows
- Application layer information

IP flow monitoring for QoS Attributes collection

- Attributes
  - Round trip time
  - Number peers/users
  - Transport size
  - Application response time
- Passive, continuous observation
  - Observation point location makes the difference

Evaluated Forecasting Methods

three approaches to time series forecasting

ARIMA(p,d,q)

autoregression and moving average in one package

Auto-Regression

- evolving variable of interest is regressed on its own lagged (i.e., prior) values

Moving Average

- regression error is a linear combination of error terms whose values occurred at various times in the past

Integrated

- transformation applied to timeseries in order to make it stationary

$$(1-\phi_1 B - \cdots - \phi_p B^p)(1-B)^d y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t$$

Holt-Winters

seasonality included

Model

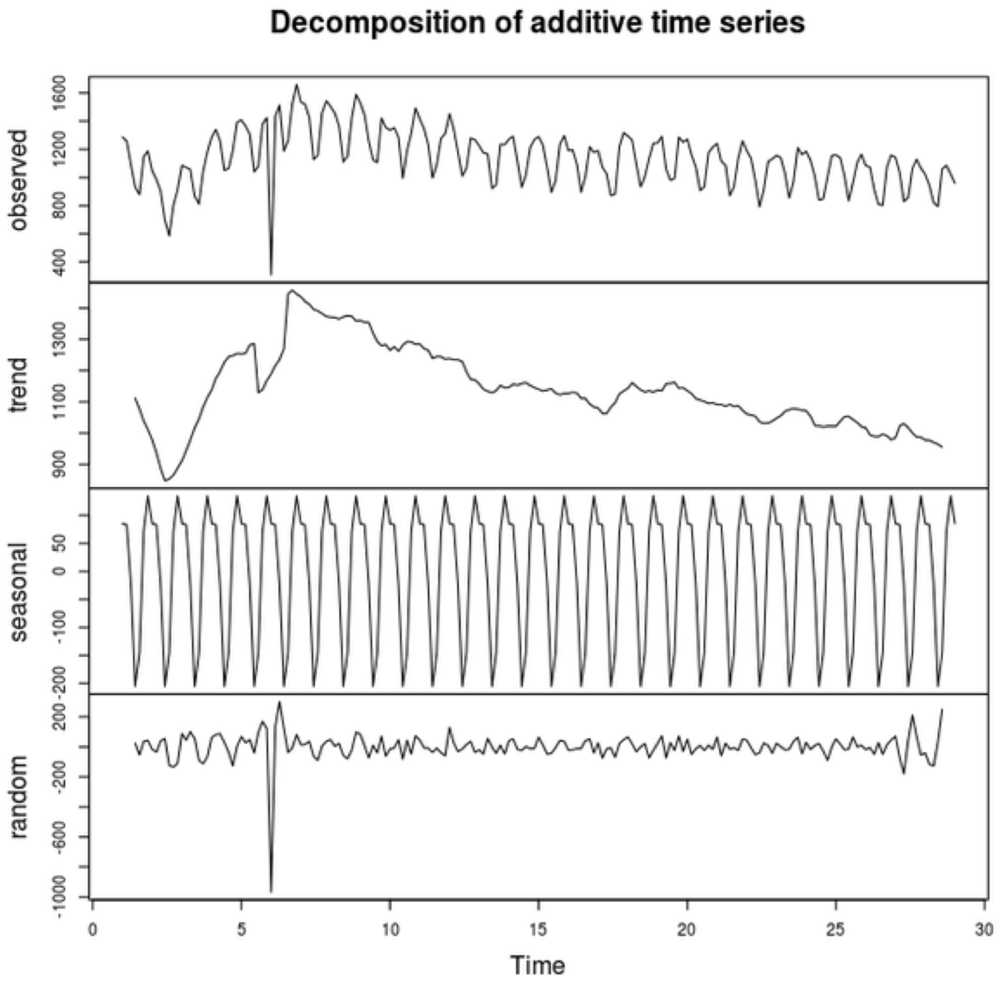
$$L_t = \alpha(y_t - I_{t-p}) + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$
$$I_t = \gamma(y_t - L_{t-1} - T_{t-1}) + (1-\gamma)I_{t-p}$$

Prediction

$$\hat{y}_t(k) = L_t + kT_t + I_{t-p+1+(k-1)modL}$$

Parameters

- Speed of learning/forgetting

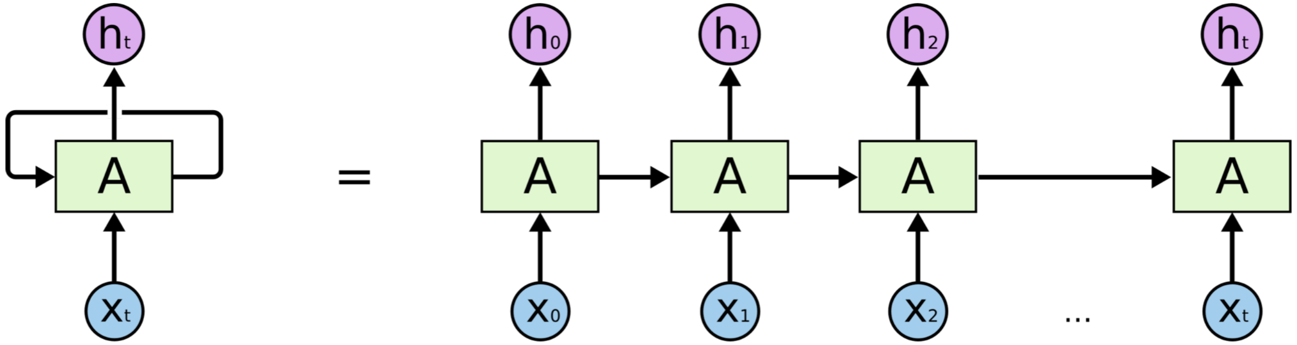


Long Short Term Memory Neural Network

recurrent neural network

Recurrent Neural Networks

- Text processing - understanding of the words based on the meaning of the previous ones.
- Classification events in the movie – previous events are necessary for reasoning
- Excellent for modelling sequences

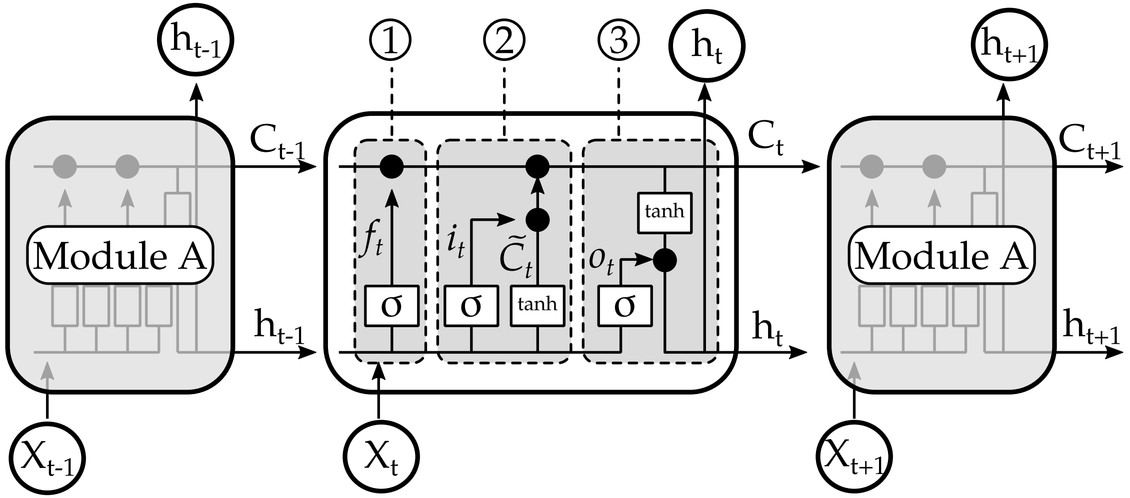


Long Short Term Memory Neural Network

recurrent neural network

Long Short Term Memory

- the context is more “far” in history
- specific function to determine what to remember
- gates
  - Forget
  - Input
  - Output



Methodology

how do we make the comparison

Dataset

real-world data shows the real performance

Two monitored services

- Access portal to information resources at university (libraries, datasets collections, …)
- Web presentation of the Faculty of Science

Observation period

- one month in 2018

Two granularities

- 5 minute => 8928 observations
- 1 hour => 744 observation

Missing values

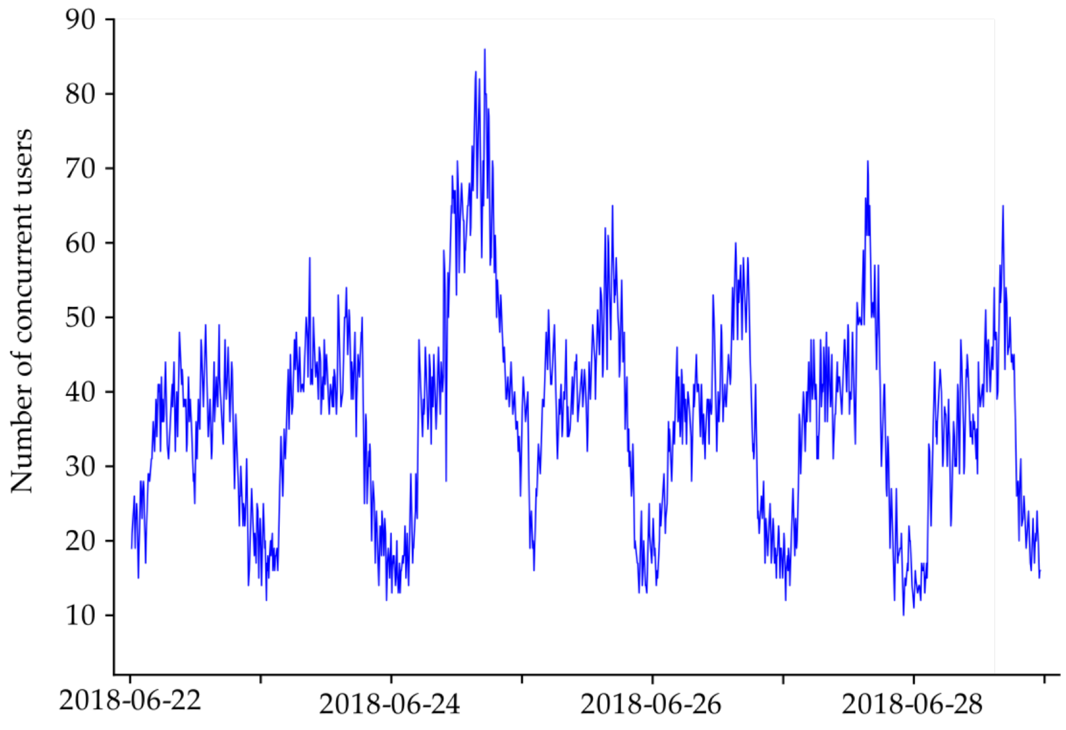
Dataset

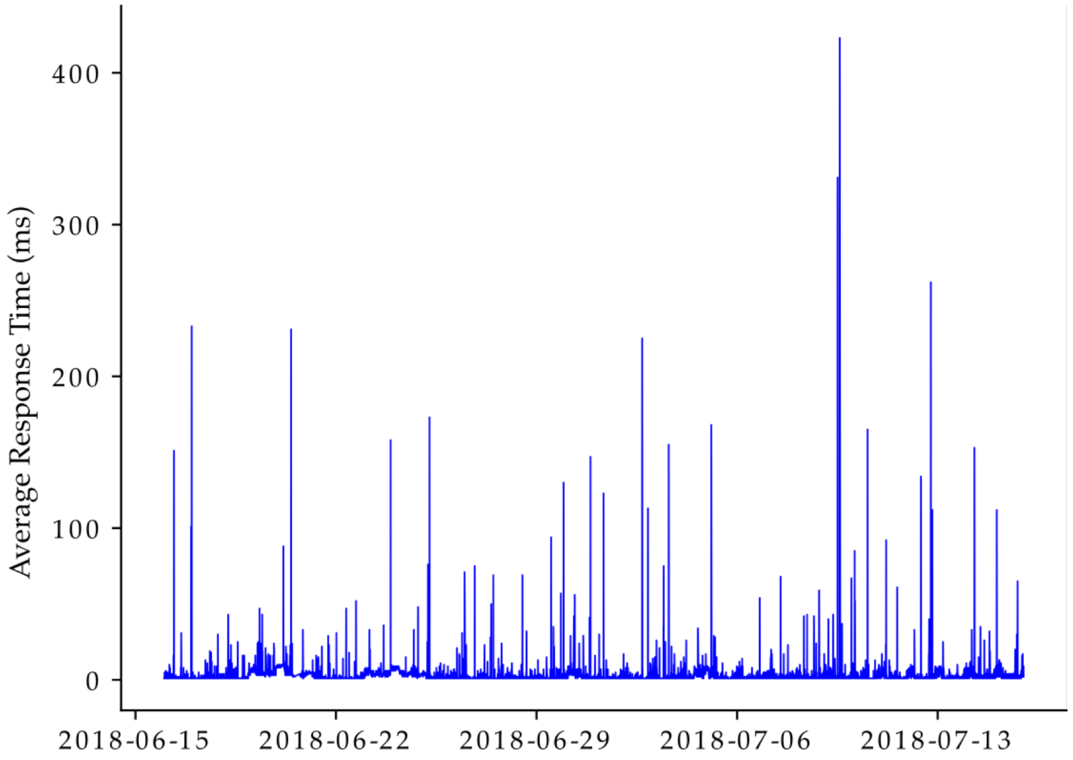
real-world data shows the real performance

QoS Attribute	Measured Statistics
Number of concurrent users (USR)	**count**
Application response time (s) (ART)	min, max, **avg**, p50, p90, **p99**
Transaction count (TC)	**count**
Network transport time (s) (NTT)	min, max, **avg**, sum
Transport size (s) (TS)	min, max, avg, **sum**

Dataset

real-world data shows the real performance





Forecast

there is not only one forecast

Time scale

- Real-time
- Short-term
- Middle-term
- Long-term

Number of forecasted observations

- One-step
- Multi-step

Forecast frequency

- One-time
- Continuous

Our goal

- One step, continuous, real-time/short-term

Models Construction

our approach to estimation

ARIMA(p,d,q)

- Box-Jenkins Methodology
  - Differencing order (Augmented Dickey-Fuller test for stationarity)
  - Autocorrelation plot to determine p,q (AIC if is unclear)
  - Maximum likelihood and Kalman Filter estimation

Holt-winters

- Additive vs multiplicative
- Season length identification (ACF, PACF)
- Parameters estimation (Maximum likelihood)

LSTM NN

- Standardization of time series
- One input, one hidden, one output layer
- MSE – stop loss function
- Stochastic gradient descent optimizer
- Number of iteration determined from learning curve

Models Evaluation

how do we compare

Training and testing dataset

Forecast Precision

- Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| * 100$$

Time complexity

- Time to estimate a model
- 6 AMD Ryzen5 CPUs 3.8GHz, 6GB RAM

Experiment Results

the data reveals the truth

Models Settings

given by the dataset

ARIMA

QoS Attribute	SERV-1		SERV-2	
	5 min	1 h	5 min	1 h
Number of concurrent users (USR)	(2,0,0)	(2,0,0)	(2,0,0)	(2,0,0)
Response time - avg (ART-avg)	(2,1,0)	(1,0,0)	(1,0,0)	(1,0,0)
Response time - p99 (ART-p99)	(1,1,0)	(2,1,0)	(0,0,1)	(1,0,0)
Transaction count (TC)	(3,0,0)	(2,0,0)	(4,0,0)	(3,0,0)
Network transport time (NTT)	(2,1,0)	(1,1,0)	(0,0,1)	(1,0,0)
Transport size (TS)	(3,0,0)	(2,0,0)	(3,0,0)	(3,0,0)

Holt-Winters

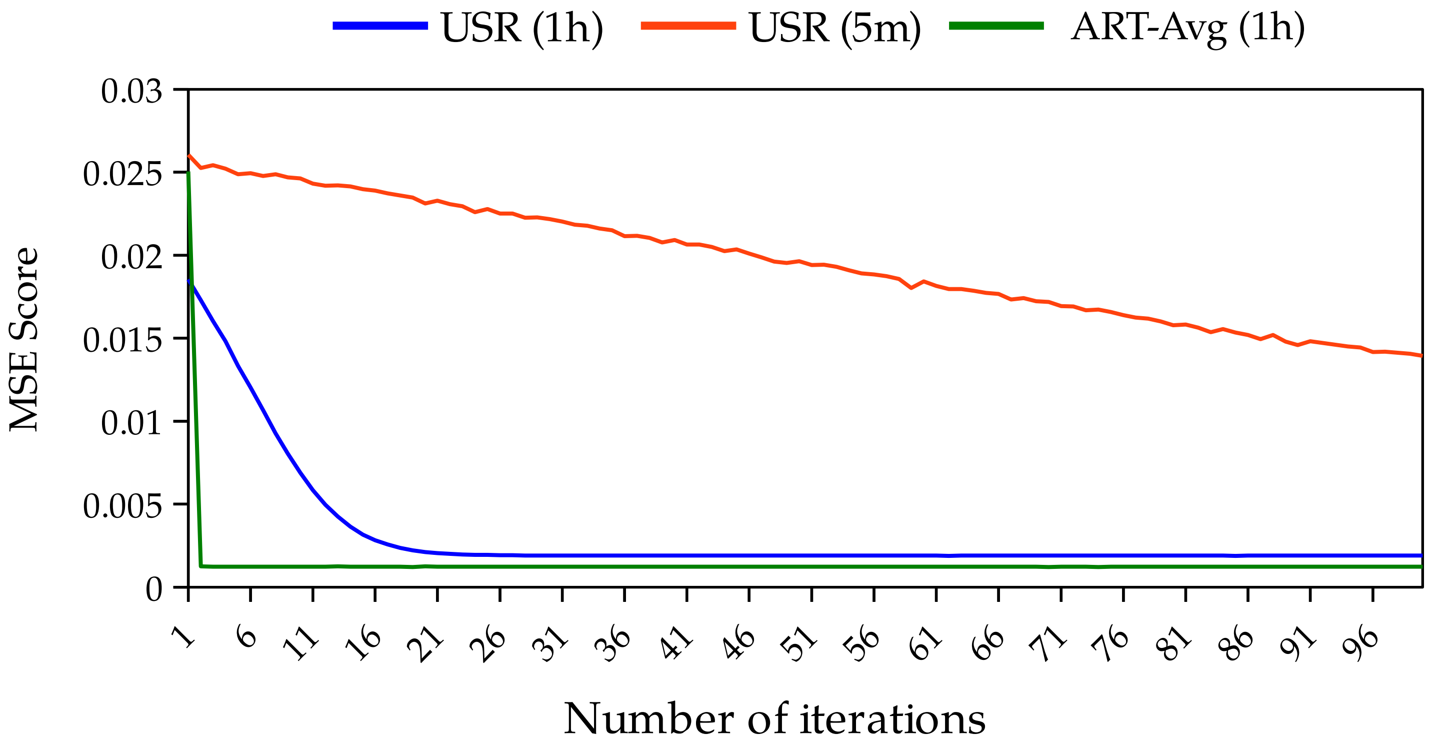
- USR, TC day-night, week pattern
- Season set to 7 days
- Parameter estimation
  - Level – varied over whole interval
  - Trend – no trend identified
  - Season – close to one – recent more weight

Models Settings

given by the dataset

LSTM NN

- Two hidden cells
- Number of iterations
  - ART, NTT, TS – rapid drop
  - USR, TC – 1 hour
  - Other linear descend
  - Set to 100



Models Comparison

MAPE performance

QoS Attribute	Service	ARIMA		Holt-Winters		LSTM NN	
		5 min	1 h	5 min	1 h	5 min	1 h
Number of concurrent users (USR)	SERV-1	7.79	*13.70*	28.09	38.89	**2.16**	20.27
	SERV-2	5.44	*10.02*	24.41	32.11	**1.61**	20.84
Response time - avg (ART-avg)	SERV-1	119.04	*113.01*	212.61	141.45	**100.99**	116.52
	SERV-2	103.44	41.24	66.48	45.08	**40.39**	*30.87*
Response time - p99 (ART-p99)	SERV-1	**250.42**	*110.83*	504.83	195.52	497.23	153.43
	SERV-2	205.54	84.62	165.77	126.70	**106.58**	*71.70*
Transaction count (TC)	SERV-1	**76.28**	*50.80*	310.98	272.62	252.68	119.89
	SERV-2	36.23	28.75	226.91	198.95	**28.07**	*11.40*
Network transport time (NTT)	SERV-1	288.63	96.53	**238.26**	99.16	460.22	*69.96*
	SERV-2	**374.92**	81.40	394.67	96.04	409.73	*34.31*
Transport size (TS)	SERV-1	46.82	*25.99*	51.79	160.12	**46.40**	39.72
	SERV-2	**112.90**	48.84	210.06	154.127	386.56	*35.73*

Time Complexity

how long does it take

Granularity	ARIMA	Holt-Winters	LSTM NN
5 minutes	574.56 ± 509.06	44.04 ± 4.17	397.48 ± 43.63
1 hour	30.21 ± 30.42	2.92 ± 0.82	33.70 ± 1.61

Further Notes

what can be improved

Initial weights for LSTM NN

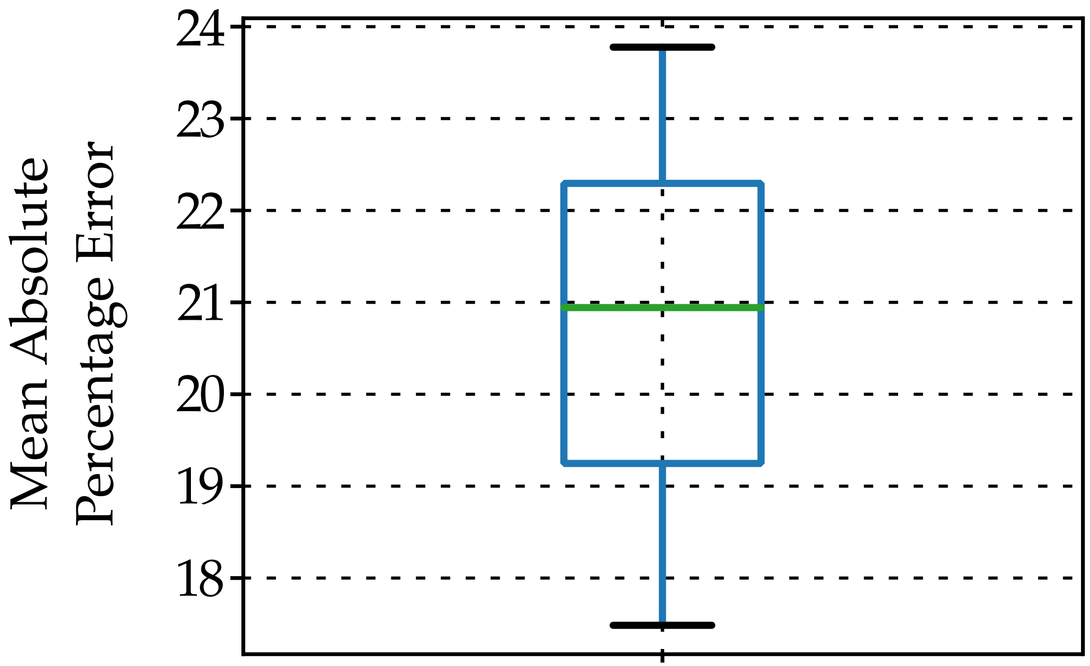
Outliers present

- Use Symmetric Mean Absolute Percentage Error instead MAPE

LSTM Time complexity

- Adam or RMSProp optimizer instead SGD

Data preprocessing



Number of concurrent users
(SRV-2, 1hour granularity)

Summary
and future work

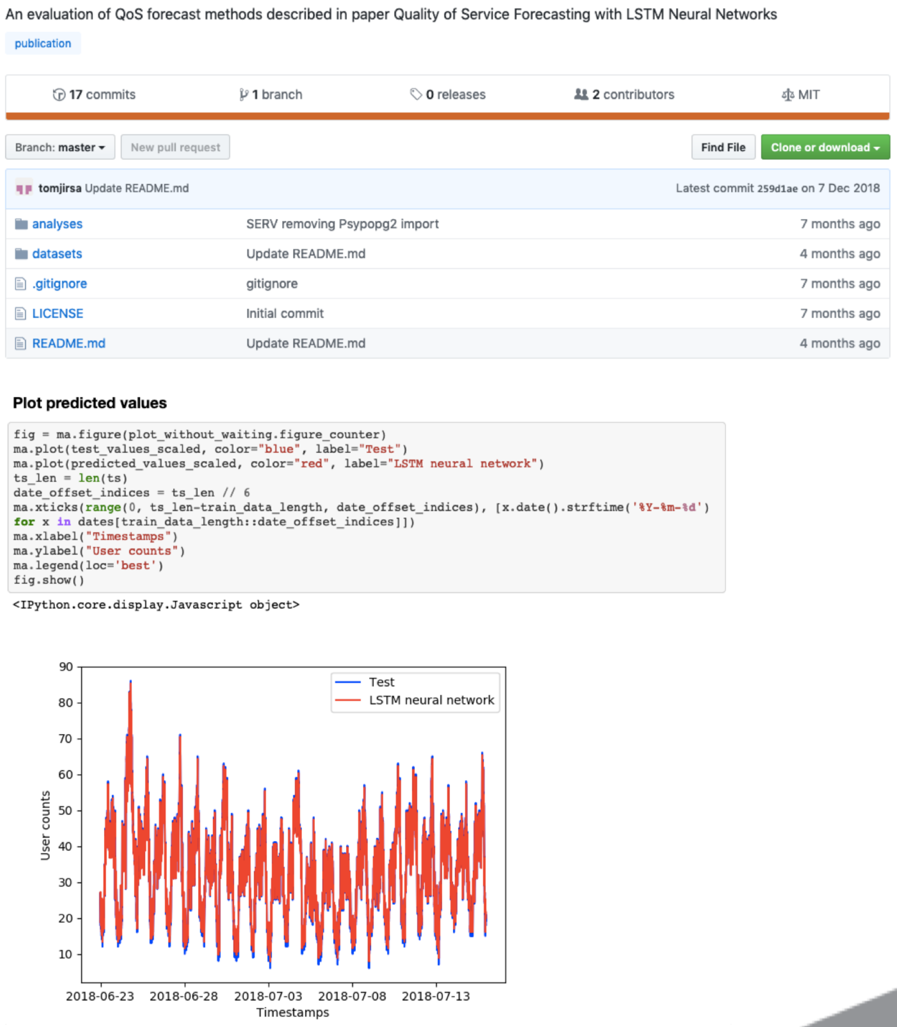
Centralized monitoring of QoS

Comparison of methods for QoS timeseries forecasting

- ARIMA vs. Holt-winters vs. LSTM NN
- LSTM NN better for high granular data
- Dataset and experiment released for public

Future work

- K-step prediction
- Optimization of LSTM NN performance
- Data preprocessing



Thank you for your attention

https://csirt.muni.cz/

https://github.com/CSIRT-MU/QoSForecastLSTM

@csirtmu

Tomas Jirsik et al.

jirsik@ics.muni.cz

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

MUNI ICS

CSIRT-MU