Głębokie modele uczenia maszynowego - czujniki zanieczyszczeń

Maciej Urbaniak 200842
Projekt realizowany wspólnie z Łukasz Drewniak

Abstract

Sprawozdanie skupia się na opisie danych oraz korelacji między danymi podanymi na wejście sieci rekurencyjnej.

1 Opis problemu

Problemem który staraliśmy się rozwiązać jest predykcja zanieczyszczenia powietrza przez występujące w powietrzu pyły zawieszone. W przypadku wykorzystanych danych z sieci czujników Luftdaten `https://luftdaten.info` są to pyły o średnicy 2.5μm oraz 10μm. Opieramy się głównie na trzech poniżej wymienionych dokumentach: "DeepAirNet: Applying Recurrent Networks for Air Quality Prediction."(1), "A deep learning model for air qualityprediction in smart cities."(2) oraz "Time Series Based LSTM Model to Predict Air Pollutant's Concen-tration for Prominent Cities in India."(3).

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Opis literatury

"DeepAirNet: Applying Recurrent Networks for Air Quality Prediction"(1) opisuje wykorzystanie 3 rodzajów komórek(RNN, LSTM i GRU) oraz wpływu liczby warstw od 1 do 4. Najlepsze wyniki uzyskali dla komórki GRU dlatego też również przychyliliśmy się w naszym projekcie do zbadania i wykorzystania właśnie jej.

"A deep learning model for air qualityprediction in smart cities."(2) wykorzystano sieci LSTM. Pokazano dodatkowo (zakładam że przypadkowo) jak można polepszyć celność modelu nie predykując dokładnych wartości a stosując klasy, które są pewnymi zakresami wartości.

"Time Series Based LSTM Model to Predict Air Pollutant's Concen-tration for Prominent Cities in India."(3) również LSTM dodatkowo skupiają się na opisie zgromadzonych danych szukając pewnych prawidłowości (w poprzednich pracach dane są zwyczajnie wrzucane do modelu). Korzystają z 2 warstwowej sieci LSTM z optymizerem Adam dla 100 epok.

3 Dane

3.1 Dane meterologiczne - Wrocław Open Data

Otwarte Dane Wrocław - to wrocławski serwis internetowy `https://www.wroclaw.pl/open-data/` umożliwiający szybkie i łatwe dotarcie do informacji publicznych gromadzonych przez Urząd Miejski Wrocławia oraz inne jednostki miejskie. Między innymi gromadzone są tam dane pogodowe. Czujniki pogodowe są w następujących lokalizacjach:

- ul. Sulowska / Most Widawski, (czujnik uszkodzony)
- ul. Lotnicza / ul. Kosmonautów,
- Most Romana Dmowskiego,
- al. Jana III Sobieskiego,
- ul. Opolska / ul. Katowicka,
- Estakada Gądowianka,
- Most Milenijny,
- Most Warszawski

oraz mierzą następujące parametry:

- wilgotność
- temperatura gruntu
- temperatura powietrza
- kierunek wiatru
- prędkość wiatru
- typ opadu

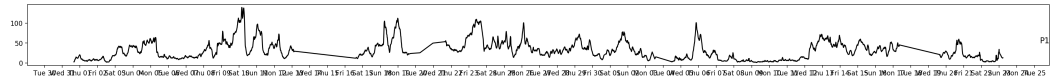
3.2 Dane zanieczyszczenia powietrza - Luftdaten

Luftdaten to niemiecki serwis internetowy `https://luftdaten.info` oraz projekt społeczny umożliwiający dołączanie własnych czujników do wspólnie tworzonej sieci oraz pobieranie danych z dostępnych czujników w ramach tej społeczności. Charakterystyka gromadzonych danych:

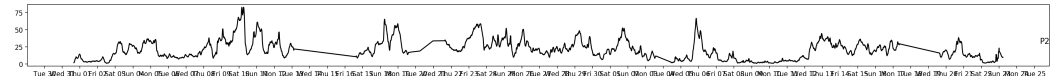
- PM10
- PM2.5
- czas pomiaru

3.3 Rozkłady danych

Rozkłady zanieczyszczeń są do siebie bardzo zbliżone co widać na poniższych rysunkach oraz na macierzy korelacji w późniejszym rozdziale. W procesie zbierania danych wystąpiły przerwy w dostępie do pomiarów co można zauważyć na wykresach jako stale malejącą linię trendu z powodu zastosowania średniej kroczącej. Przy niedługo trwającej przerwie w dostępie do danych pomaga to w wygładzeniu nagłych skoków w wartościach. Niestety jak można zauważyć okresy niedostępności rozłożyły się nawet na kilka dni. Wynika to częściowo z faktu iż inicjatywa jest społeczna i niektórzy wyłączają czujniki np. na czas wyjazdu z domu.

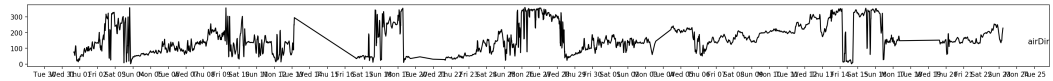


Rysunek 1: Dystrybucja PM10

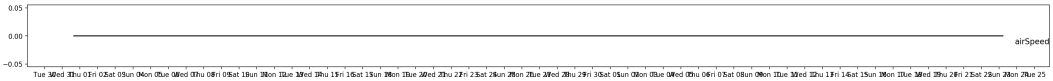


Rysunek 2: Dystrybucja PM2.5

Poniżej zostały przedstawione wykresy dotyczące kierunku i siły wiatru. Jak można z łatwością zauważyć siła wiatru jest stała z powodu prawdopodobnego uszkodzenia czujnika(tyczy się to aż połowy stacji we Wrocławiu).

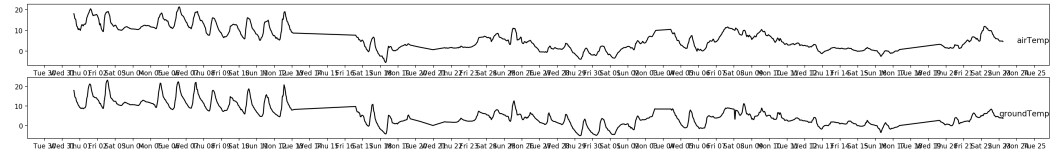


Rysunek 3: Dystrybucja kierunku wiatru



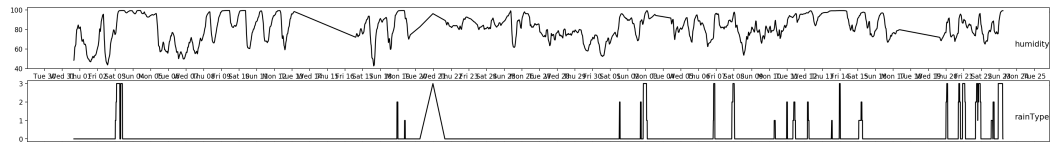
Rysunek 4: Dystrybucja siły wiatru

Na rysunku niżej widać wykresy temperatur zależności między tymi wskazaniami różnią się w zależności od pory roku, ale prawie zawsze są ściśle ze sobą skorelowane(co jest dość oczywiste).



Rysunek 5: Dystrybucja temperatur powietrza i gruntu

Oraz ostatnie wskazanie typu opadu oraz wilgotności. Jak widać nie zawsze wzrost wilgotności spowodowany jest występującymi opadami. Okresowość wzrostów i spadków sugeruje występowanie porannej rosy.



Rysunek 6: Dystrybucja wilgotności i typu opadów

3.4 Korelacja w danych

W zebranych pomiarach występuje znaczna korelacja w wartościach pomiędzy pyłami zawieszonymi PM10 oraz PM2.5 co może sugerować, że pochodzą z tego samego źródła. Dlatego też predykujemy tylko jedno wskazanie pyłu zawieszonego(szczegóły w drugim wypracowaniu). Inną korelacją tym razem negatywną są wskazania pomiędzy zanieczyszczeniami a temperaturą co sugeruje wpływ czynnika jakim jest stosowanie opału do ogrzewania budynków w zimie. Pomiar cząsteczek zanieczyszczeń odbywa się z użyciem laserowego czujnika pyłu SDS011, przez co możliwe są wzrosty wskazań w przypadku zassania cząsteczek wody do układu. W macierzy korelacji widać, że właśnie taka sytuacja mogła wystąpić.

	P1	P2	airDir	airSpeed	airTemp	groundTemp	humidity	rainType
P1	1	0.979775	-0.0274091	nan	-0.319105	-0.251368	0.467102	-0.0339331
P2	0.979775	1	-0.0418078	nan	-0.316738	-0.246976	0.455993	-0.0227802
airDir	-0.0274091	-0.0418078	1	nan	-0.200514	-0.169696	0.130294	0.0395581
airSpeed	nan	nan	nan	nan	nan	nan	nan	nan
airTemp	-0.319105	-0.316738	-0.200514	nan	1	0.964986	-0.365063	-0.00841263
groundTemp	-0.251368	-0.246976	-0.169696	nan	0.964986	1	-0.363516	-0.0144523
humidity	0.467102	0.455993	0.130294	nan	-0.365063	-0.363516	1	0.1604
rainType	-0.0339331	-0.0227802	0.0395581	nan	-0.00841263	-0.0144523	0.1604	1

Rysunek 7: Macierz korelacji

4 Opis wektora wejściowego

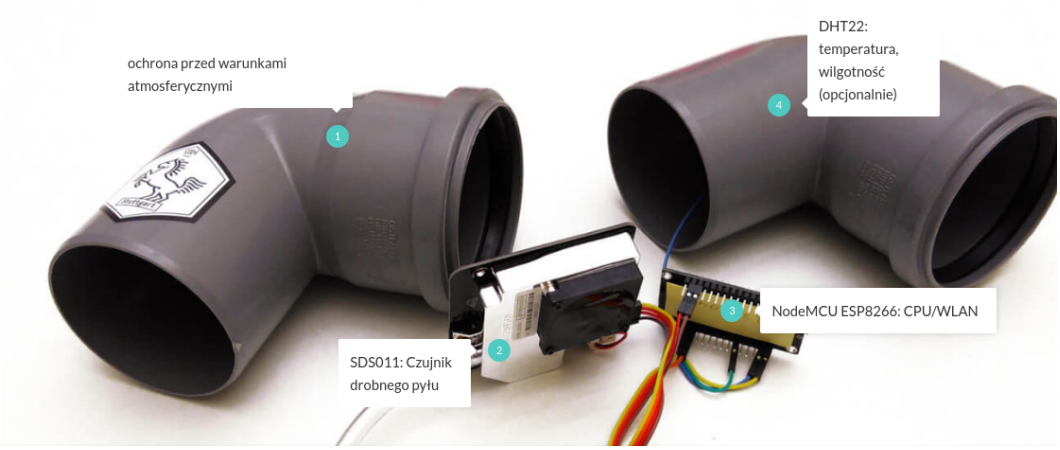
W celu stworzenia wektora który podajemy na wejście sieci połączono dane czujnika z najbliższą stacją pogodową tworząc następującą strukturę:

P1	P2	airDir	airSpeed	airTemp	date	groundTemp	humidity	rainType

Kolejne komórki przekładają się kolejno na wskazania: PM10, PM2.5, prędkości wiatru (m/s), temperatury powietrza (°C), daty (RMD HH:MM:SS), temperatury gruntu (°C) wilgotności (%) oraz typu opadu (69-Brak opadu 70-Opad przelotny 71-Opad ciągły 72-Opad intensywny)

5 Czujnik

Dane zbierane są z czujników takich jak na rysunku poniżej. Z racji że jest to inicjatywa społeczna często czujnik jest nieprawidłowo instalowany, nie posiada dodatkowych modułów jak pomiar temperatury czy wilgotności oraz często jest wyłączany przez właścicieli na czas wyjazdu. Dodatkowo polega na połączeniu z siecią wifi użytkownika co również wpływa na czasowe zaniki wskazań pomiarów. Mimo tego i tak nie przebiły rekordu w czasie niedostępności danych z otwartych danych Wrocławia trwających od 24.12.2018r. do dnia dzisiejszego czyli 27.01.2019r. wskazania czujników pogodowych wciąż pokazują pomiary z Wigili Bożego Narodzenia.



Rysunek 8: Czujnik źródło zdjęcia:`https://luftdaten.info/`

6 Wyniki

6.1 Opis wykresów

Predykowane są **wartości PM10** które zostały przeskalowane tak aby przyjmowały wartości od 0 do 1 (sigmoidalna warstwa aktywacji) i znajdują się na **osi Y**.

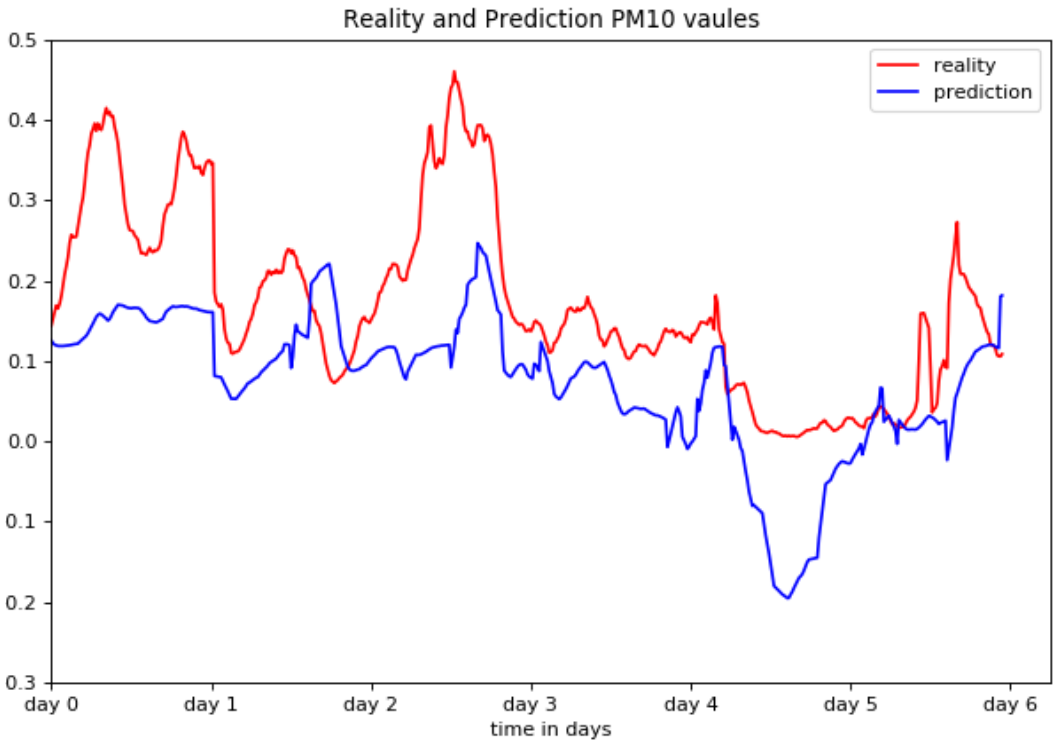
Oś X wyznacza **czas**, gdzie każdy krok jest co 15 minut, łatwo można wyliczyć że zaznaczone na osi X punkty 100, 200,(...), 600 odpowiadają kolejno 25, 50, (...), 150 godzinom predykcji w przyszłość.

Czerwony kolor linii na wykresie oznacza jakie wartości PM10 były w rzeczywistości odczytane z czujnika. Dla czytelności wykresu wartości te zostały połączone linią ciągłą.

Niebieskim kolorem linii na wykresie oznaczono natomiast predykcję PM10.

6.2 Regresja liniowa

Jako punkt odniesienia wykorzystano regresję liniową widoczną na rysunku poniżej. Widać że niedopasowanie zdarza się często dla wartości które są mocno odchylone od centrum wykresu. Dodatkowo po 400 kroku z powodu gwałtownego spadku wartości, po okresie względnej stabilności, następuje wyolbrzymiona odpowiedź dla predykcji. Inną negatywną cechą zastosowania regresji jest spadek wskazań poniżej minimalnych. W przełożeniu na dane rzeczywiste spadek po kroku 400 oznaczałby ujemne zanieczyszczenie, co jest niemożliwe z technicznego punktu widzenia.



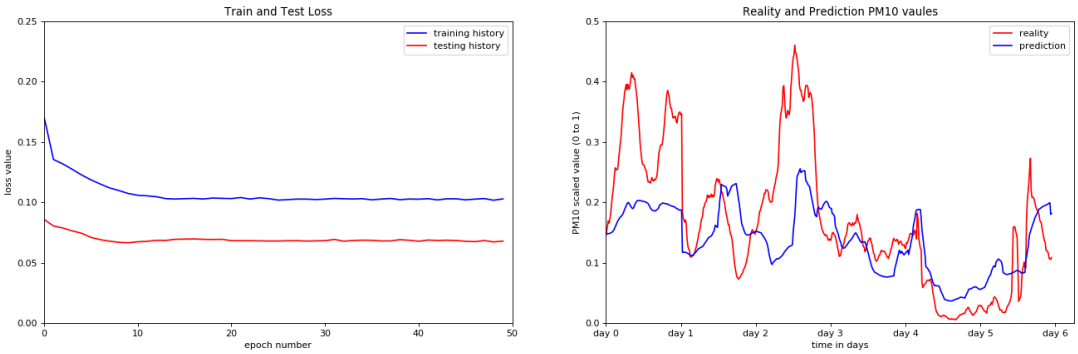
Rysunek 9: Wyniki dla regresji liniowej i faktyczne pomiary

6.3 Podstawowa wersja sieci

Trenujemy dla 50 epok z batchem o rozmiarze 72.

Została zaimplementowana sieć rekurencyjna. Zawsze korzystamy z sigmoidalnej warstwy aktywacji(z tego powodu przeprowadzona została normalizacja danych do zakresu od 0 do 1). Kolejne warstwy to:

- GRU with 64 cells
- Dropout with rate 0.3
- GRU with 64 cells
- Dense with 1 unit



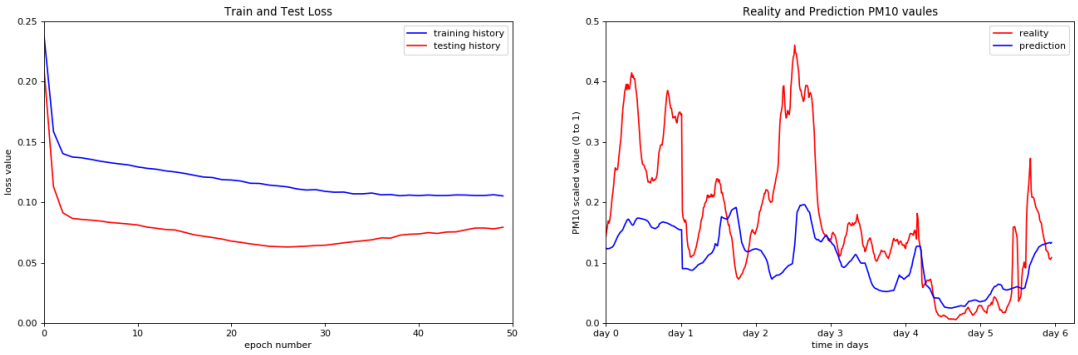
Rysunek 10: Wyniki dla predykcji sieci RNN i rzeczywiste pomiary

Podobnie jak przy zastosowaniu regresji problematyczne jest przewidzenie nagłych dodatnich skoków wartości, natomiast nie występuje już spadek predykcji pomiarów poza minimalną i wykresy bardziej do siebie przylegają.

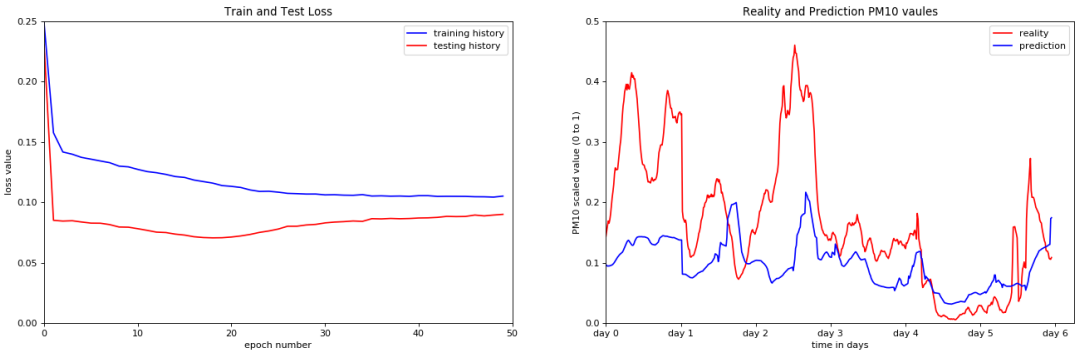
6.4 Porównanie optymalizatora

Został zbadany wpływ optymalizatorów:

- RMSprop:
- learning_rate 0.001
- rho 0.9
- epsilon None
- decay 0
- Adam:
- learning_rate 0.001
- beta_1 0.9
- beta_2 0.999
- epsilon None
- decay 0
- amsgrad False



Rysunek 11: Wyniki dla RMSprop



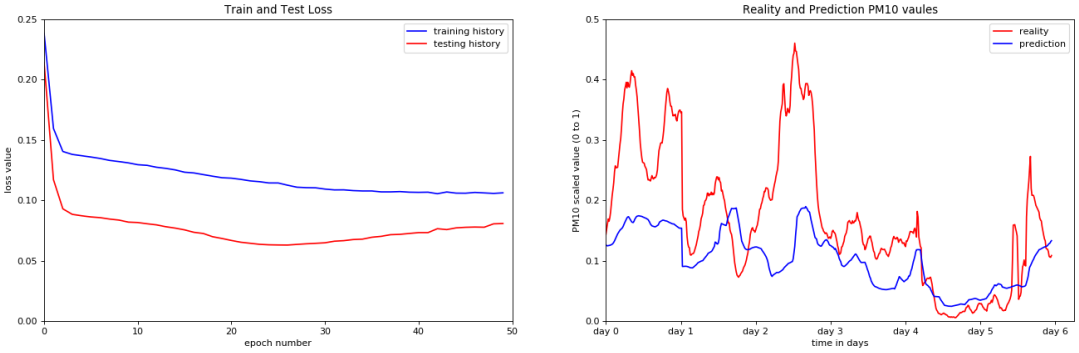
Rysunek 12: Wyniki dla Adam

W kolejnych badaniach wykorzystano optimizer RMSprop.

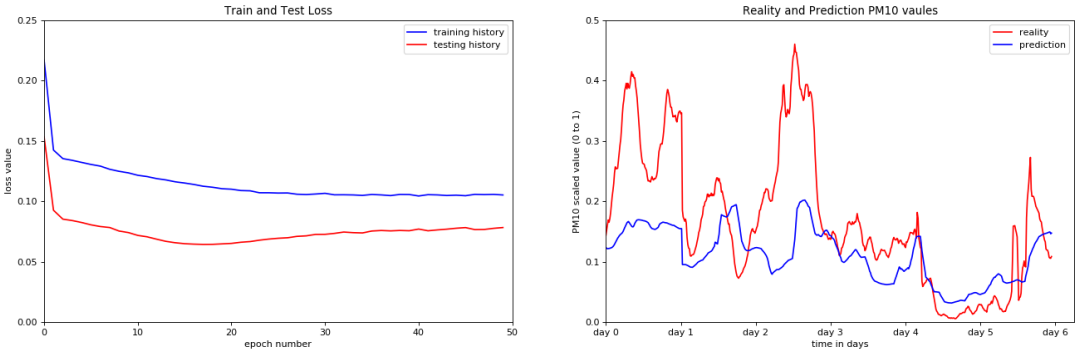
6.5 Porównanie rodzaju komórek

Został zbadany wpływ rodzaju komórki:

- LSTM
- GRU



Rysunek 13: Wyniki dla LSTM



Rysunek 14: Wyniki dla GRU

Stwierdzenie która komórka jest bardziej odpowiednia przypadła w głównej mierze w stronę zawyżenia wyników z powodu typu projektu, jakim jest predykcja szkodliwych substancji w powietrzu. Dlatego wybrano komórki typu GRU.

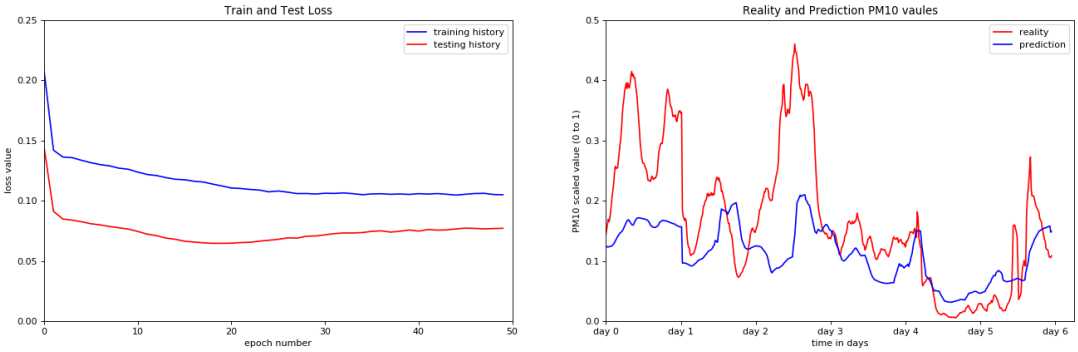
6.6 Porównanie liczby warstw

Został zbadany wpływ liczby warstw z podstawowego modelu

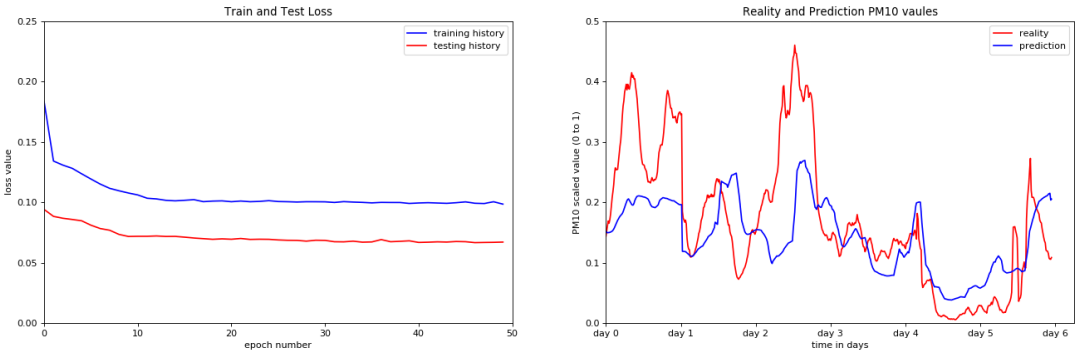
- GRU with 32 cells
- Dropout with rate 0.3
- GRU with 32 cells
- Dense with 1 unit

oraz wyniki po dodaniu kolejnych 2 warstw:

- GRU with 32 cells
- Dropout with rate 0.3
- GRU with 32 cells
- GRU with 32 cells
- GRU with 32 cells
- Dense with 1 unit



Rysunek 15: Wyniki dla 2 warstw



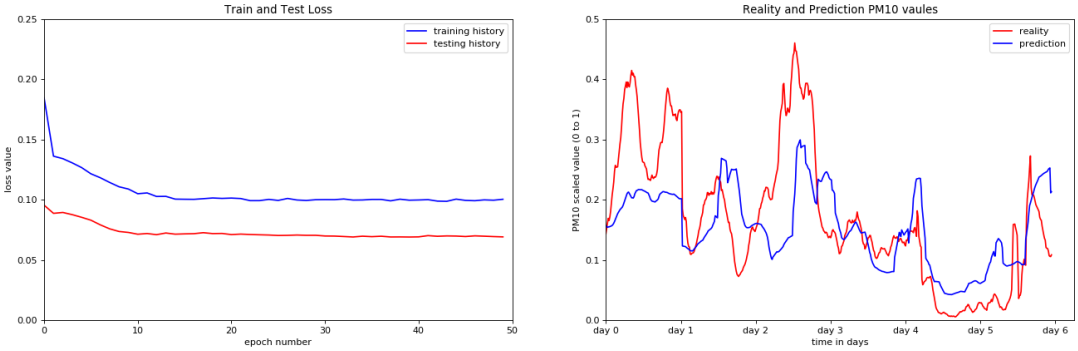
Rysunek 16: Wyniki dla 4 warstw

Zwiększenie liczby warstw pozwoliło na lepsze dopasowanie predykcji do realnych wskazań.

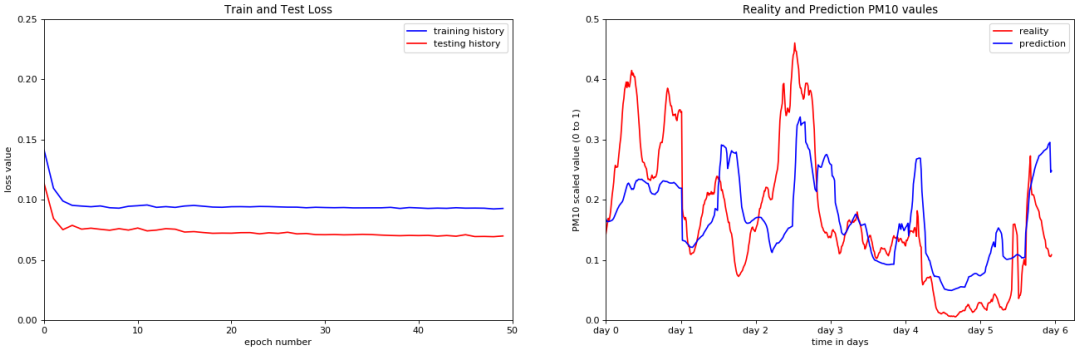
6.7 GRU - porównanie liczby neuronów

Został zbadany wpływ liczby neuronów na wyniki dla:

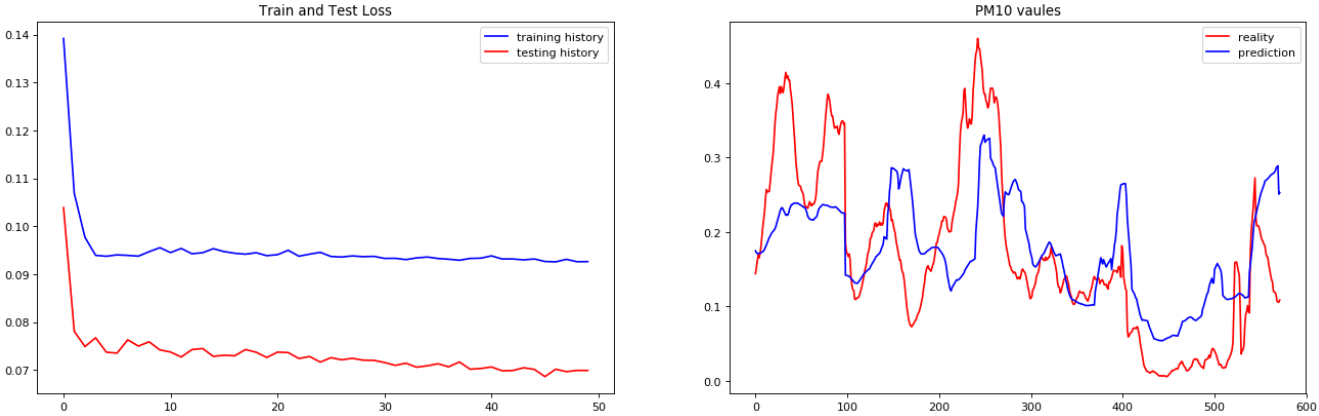
- 32 neuronów
- 64 neuronów
- 128 neuronów



Rysunek 17: Wyniki dla 32



Rysunek 18: Wyniki dla 64



Rysunek 19: Wyniki dla 128

Liczba neuronów wpływa na wysokość wskazań predykcji, może to wynikać z faktu iż na początku występują lokalne maksima odczytów zanieczyszczenia z czujników.

7 Wnioski

Jak można zauważyć ze zgromadzonych danych sieć czujników zarówno Luftdaten jak i pogodowych często ulega problemom takim jak przerwy w dostępie do pomiarów lub ciągłego braku poprawnych wskazań danego sensora. Tyczy się to w głównej mierze siły wiatru, ale może również występować dla danych o pyłach zawieszonych z powodu nieprawidłowego umieszczenia urządzenia przez użytkownika np. w domu lub blisko ścian budynków. Dodatkowo w przypadku wystąpienia problemów z serwerem po "naszej" stronie, bezpowrotnie tracimy dostęp do danych z czujników pogodowych, ponieważ nie istnieje archiwum pomiarów. W przypadku natomiast Luftdaten jest możliwość uzyskania wcześniejszych wskazań sensorów poprzez stronę projektu.

Najlepszy model (jeśli chcemy zawyżyć predykcję z uwagi na to że lepiej pomylić się o zbyt duże niż małe wskazanie, należałoby zwiększyć liczbę komórek z 32 do 128):

- GRU with 32 cells
- Dropout with rate 0.3
- GRU with 32 cells
- GRU with 32 cells
- GRU with 32 cells
- Dense with 1 unit (sigmoid)

Optymizer: RMSprop (learning_rate 0.001, rho 0.9, epsilon None, decay 0)
Liczba epok: 50
Wielkość batchy: 72

Literatura

[1] Athira, V., et al. "DeepAirNet: Applying Recurrent Networks for Air Quality Prediction." Procedia Computer Science 132 (2018): 1394-1403.

[2] Kök, İbrahim, Mehmet Ulvi Şimşek, and Suat Özdemir. "A deep learning model for air quality prediction in smart cities." Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017.

[3] Chaudhary, Vidushi, et al. "Time Series Based LSTM Model to Predict Air Pollutant's Concentration for Prominent Cities in India." (2018).