Cloud Atlas

An LstmEncoder for UHECR AirShowers

G. Becuzzi L. Papalini

July 2022

Table of Contents

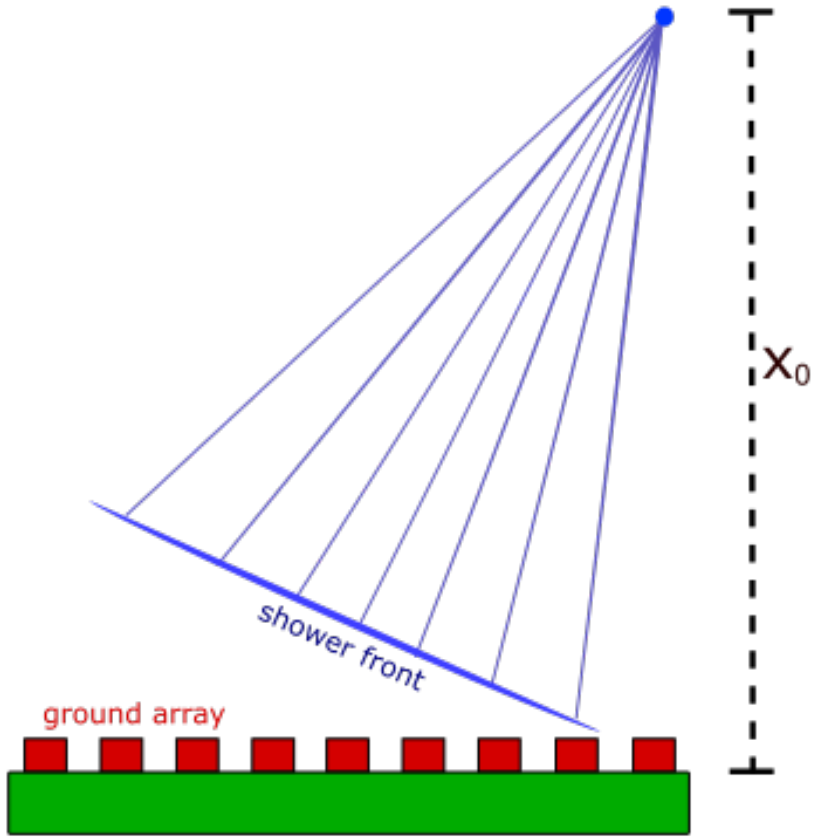
Table of Contents

UHECR Airshower

When *Ultra High Energy Cosmic Rays* (UHECR) enters the atmosphere they produce a particle cascade.

Detection: grid of water-Cherenkov ground based detectors.

Prediction: X_0 height at which the shower forms.



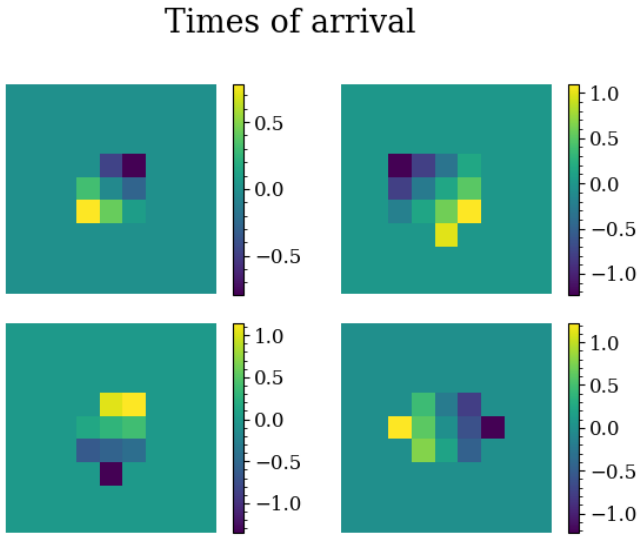
Dataset, first glance

Dataset: 10^5 simulated events:

- 9x9 grid of detectors
- most intense at center
- 80 frames of time series (40MHz sampling rate)
- 1 frame of times of first arrival

Single record shape: $(80 + 1, 81)$

`pd4ml` package splits by default in 70% train 30% test



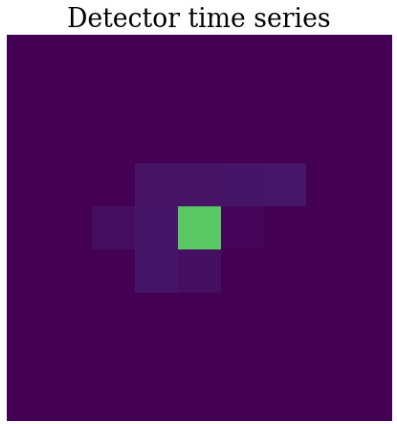


Table of Contents

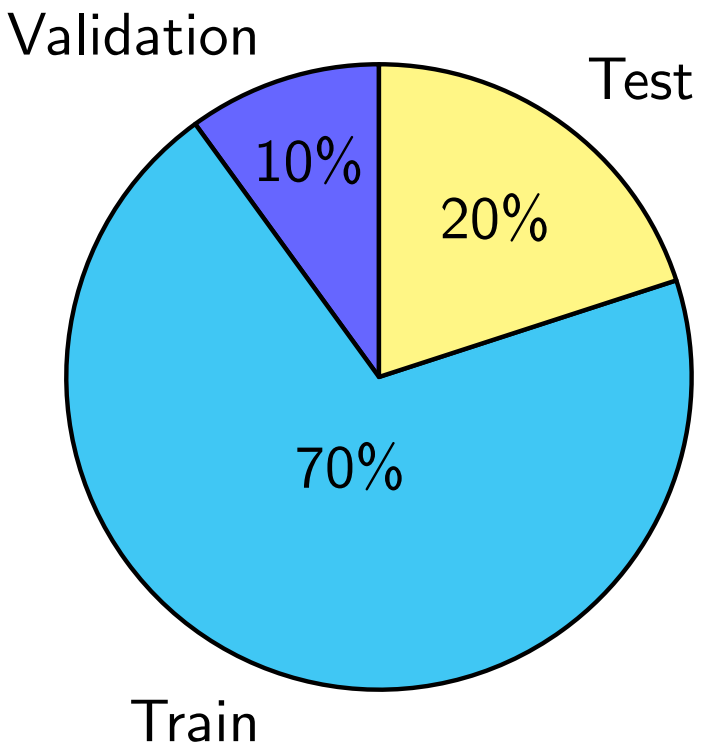
Split the dataset

Dataset was already split in test and train.

We put it all back together, shuffled it and divided with the following percentage:

- 70% *train*
- 20% *test*
- 10% *validation*

For the design of the net it is convenient using `numpy` structured arrays



Split the dataset: `funky_dtype`

```
    # custom numpy dtype
    funky_dtype = np.dtype(
        [
            ("outcome", np.float64),
            ("time_series", np.float32, (80, 81)),
            ("toa", np.float32, (9, 9, 1)),
        ])

```

All data relative to a single event is clustered in a single `numpy` object, transformation is:

$$(80+1, 81) \longrightarrow [("toa", (9,9,1)), ("timeseries", (80,9,9))]$$

Data can be accessed "as a dictionary", depending on what is needed.

DataFeeder class

Ensures an easy way to train the subnets separately Class `DataFeeder` main features:

- shuffles data randomly
- input fields can be specified
- can be extended to more complex training strategies
- returns a generator

Using a generator (`keras.utils.Sequence`)

- inherit multiprocessing features
- has default callbacks
- avoids memory overload

Resolution

The reference article suggests using the resolution:

Resolution

defined as the standard deviation of the distribution given by the difference between the predictions and the actual values of X_{max}

We point out that

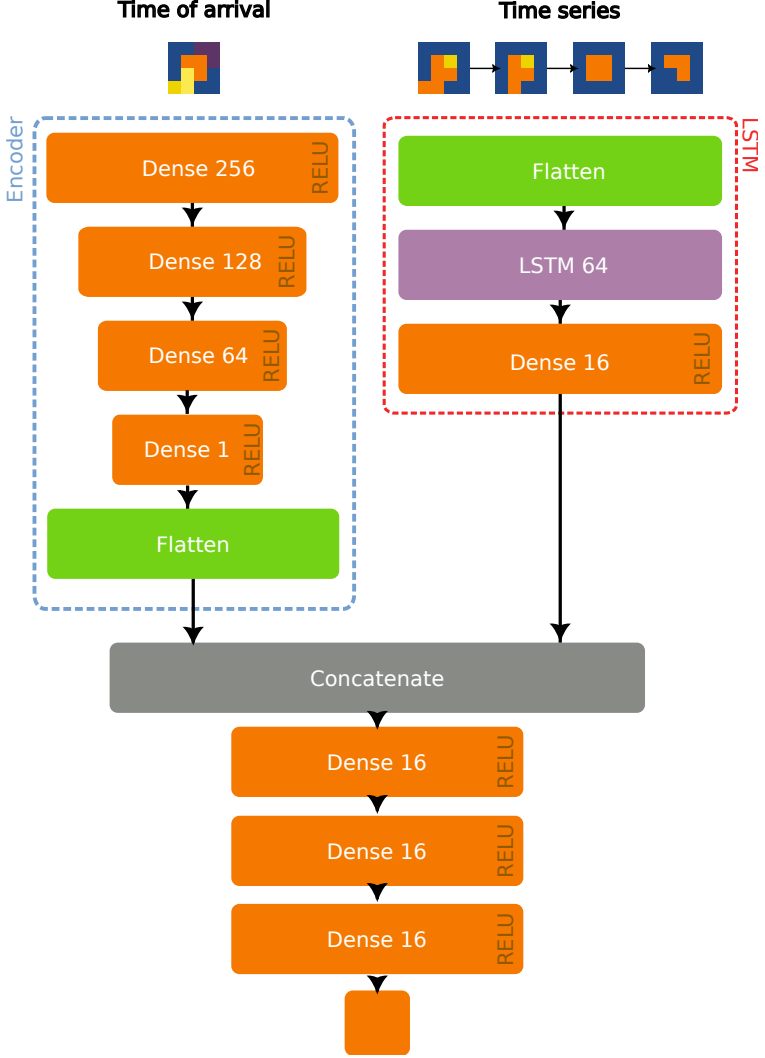
$$\sigma^2 = \frac{1}{N}\sum_i (\delta_i - \bar{\delta})^2$$

is a sensible estimator of the average error on predictions only if $\bar{\delta} = 0$, for which the adopted resolution is equal to the *RMSE* of the distribution

$$RMSE^2 = \frac{1}{N}\sum_i (x_i - \hat{x}_i)^2$$

Since (on a typical trained model) $\bar{\delta} \approx 10$m we preferred the RMSE.

Table of Contents

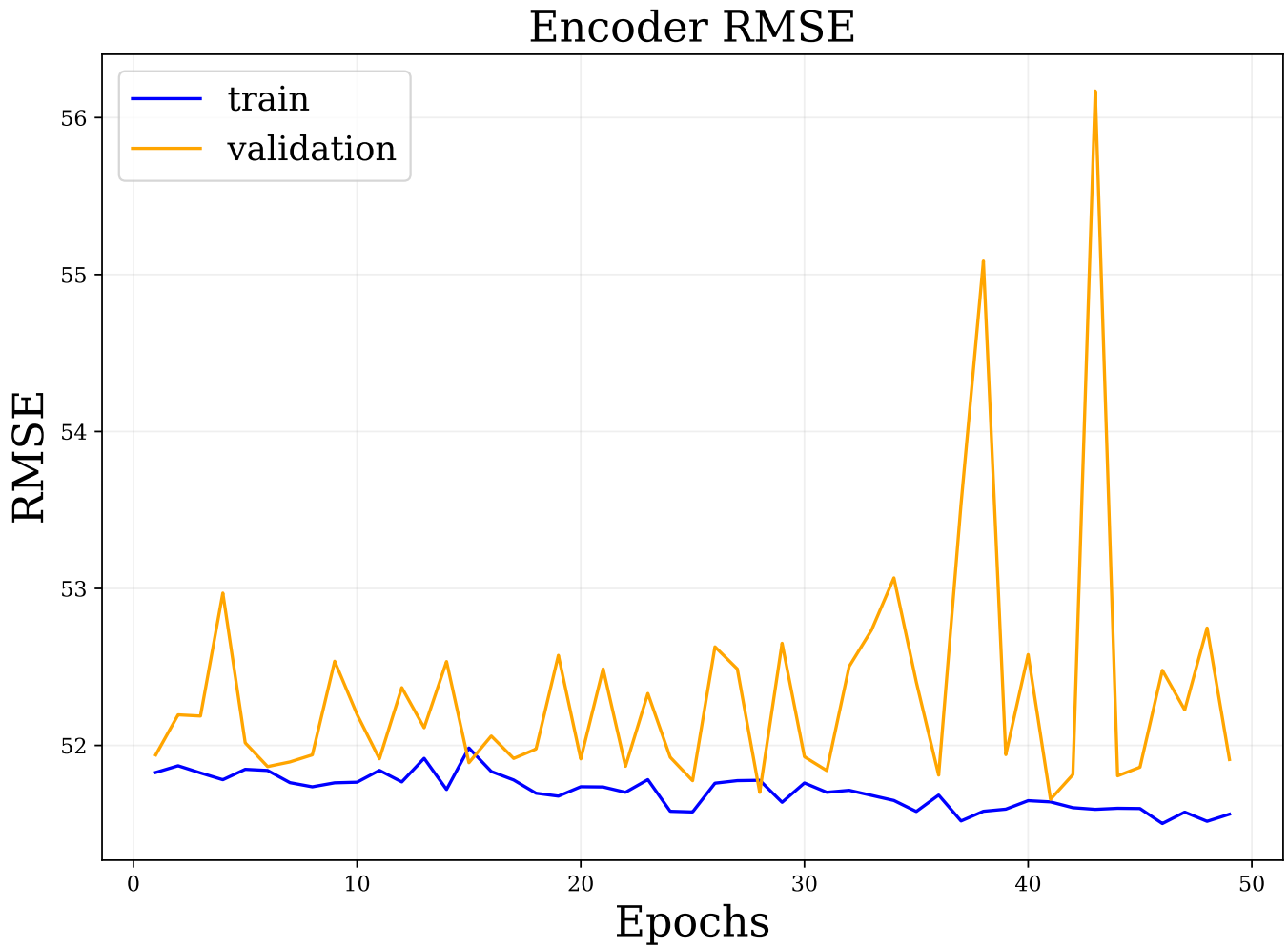

Time of arrival
Time series
Encoder
Dense 256
RELU
Dense 128
RELU
Dense 64
RELU
Dense 1
RELU
Flatten
LSTM
Flatten
LSTM 64
Dense 16
RELU
Concatenate
Dense 16
RELU
Dense 16
RELU
Dense 16
RELU

Overview on the network

The assumption that lead to this design is that from the time of arrival matrix it is possible to infer some kind of "homogeneous" shower parameters (incidence angle, spread, etc.) while the time series can be processed by a recurrent network.

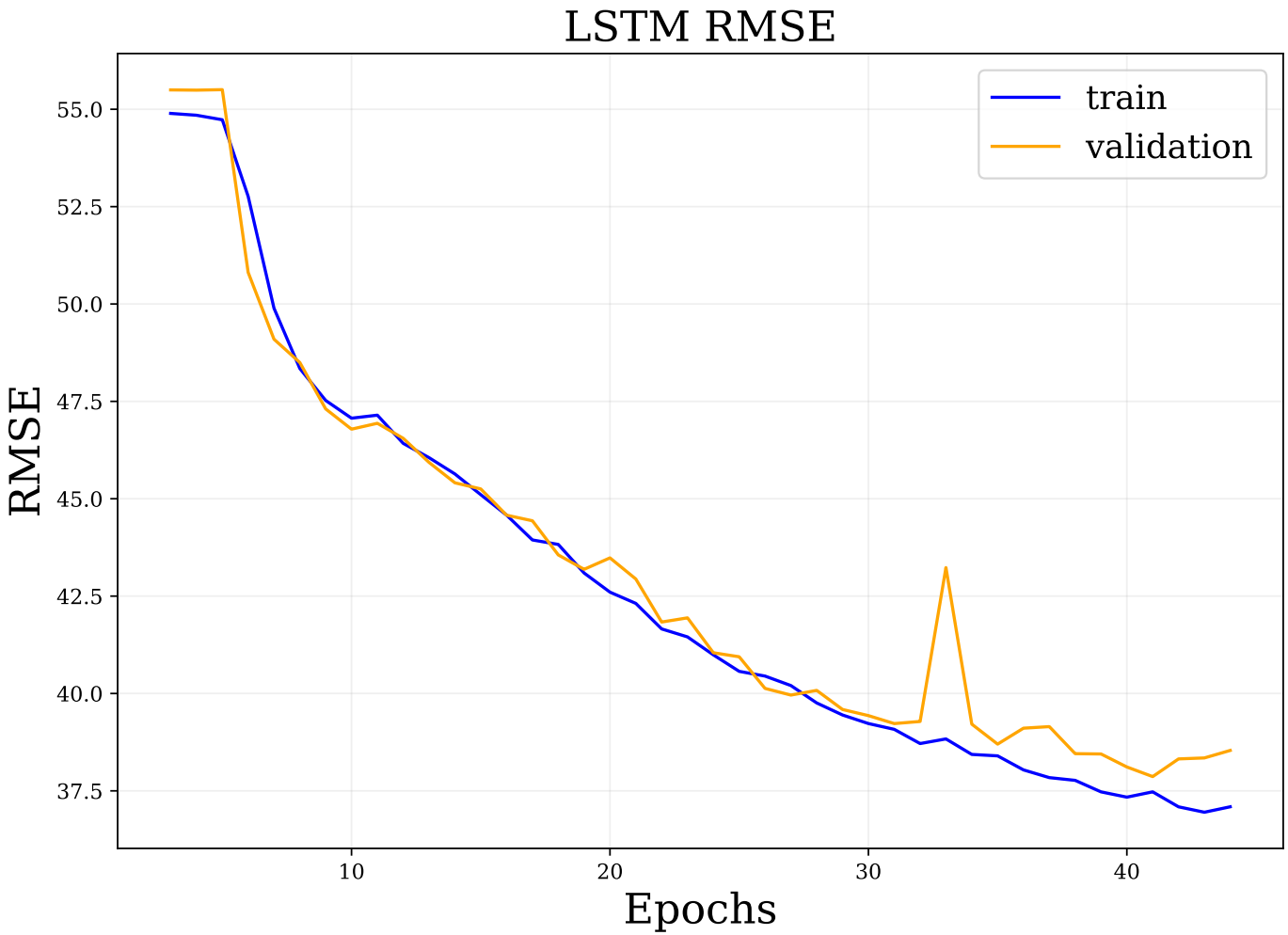
Encoder for the time of arrival

We tried to train just the Encoder with input time of arrival matrices.



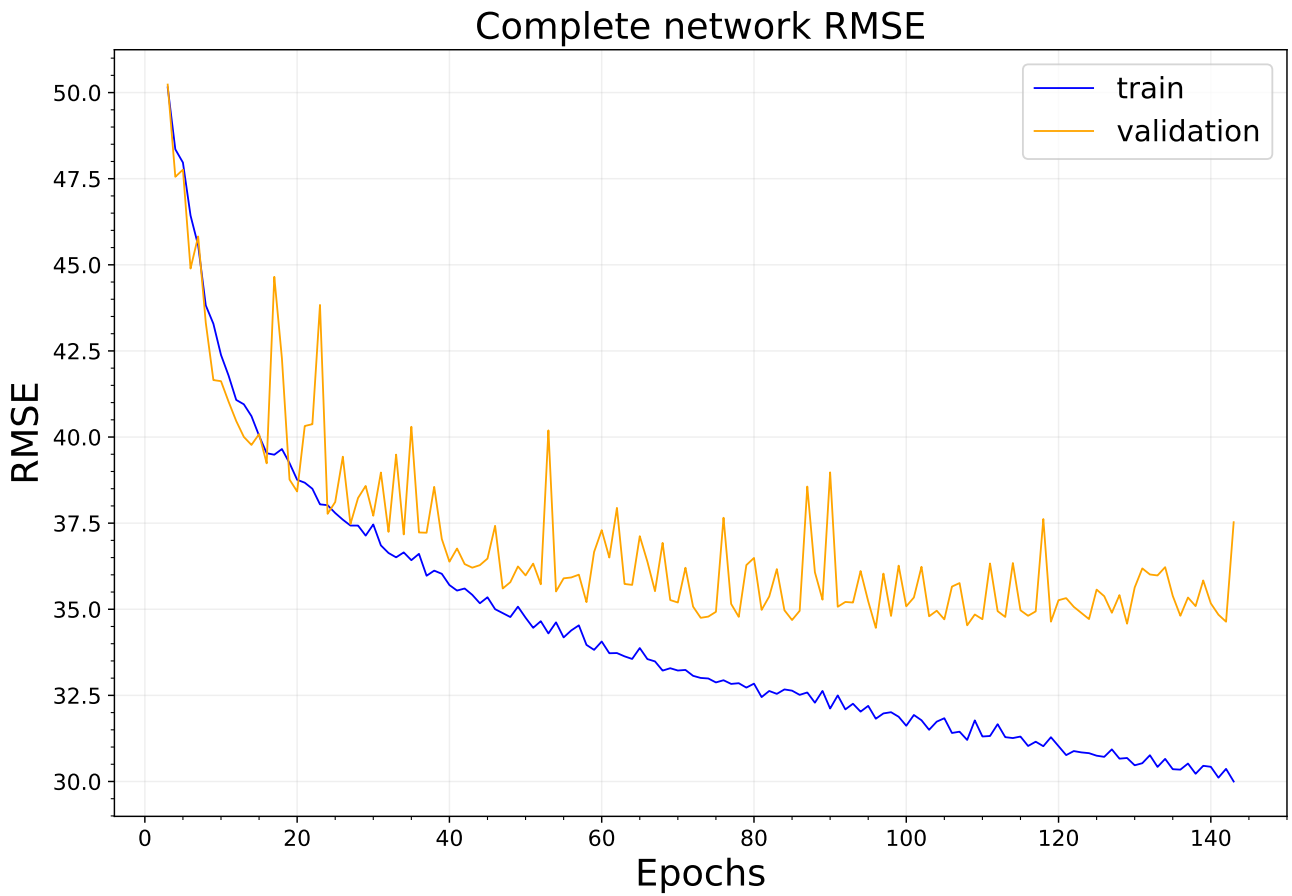
LSTM for the time series

We did the same thing with the LSTM network with the time series.



Whole Network performance

This is a training of the whole network on 144 epochs.



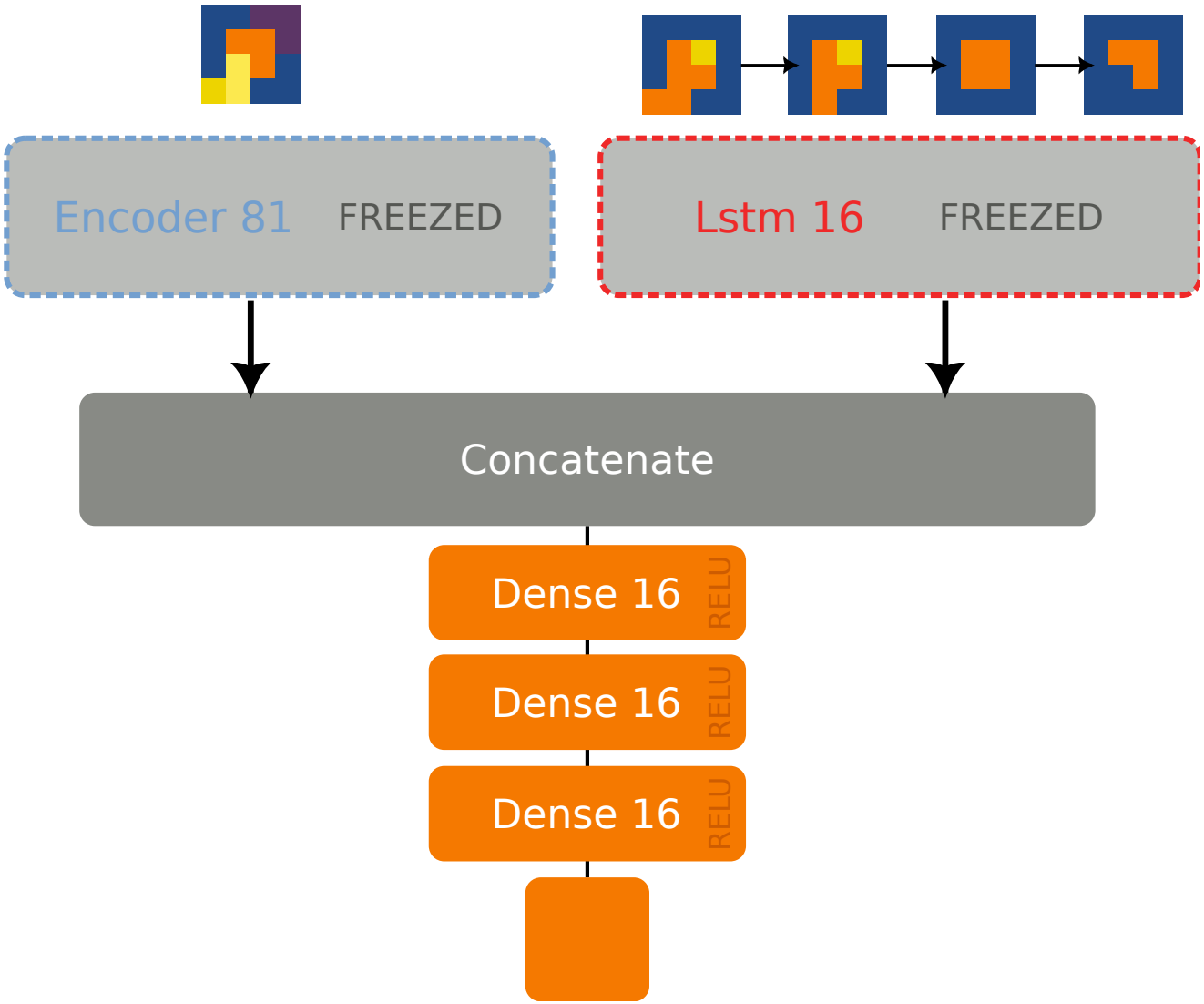
Uncertainties

Strategies for Performance Improving

We tried 3 strategies to improve network's performance:

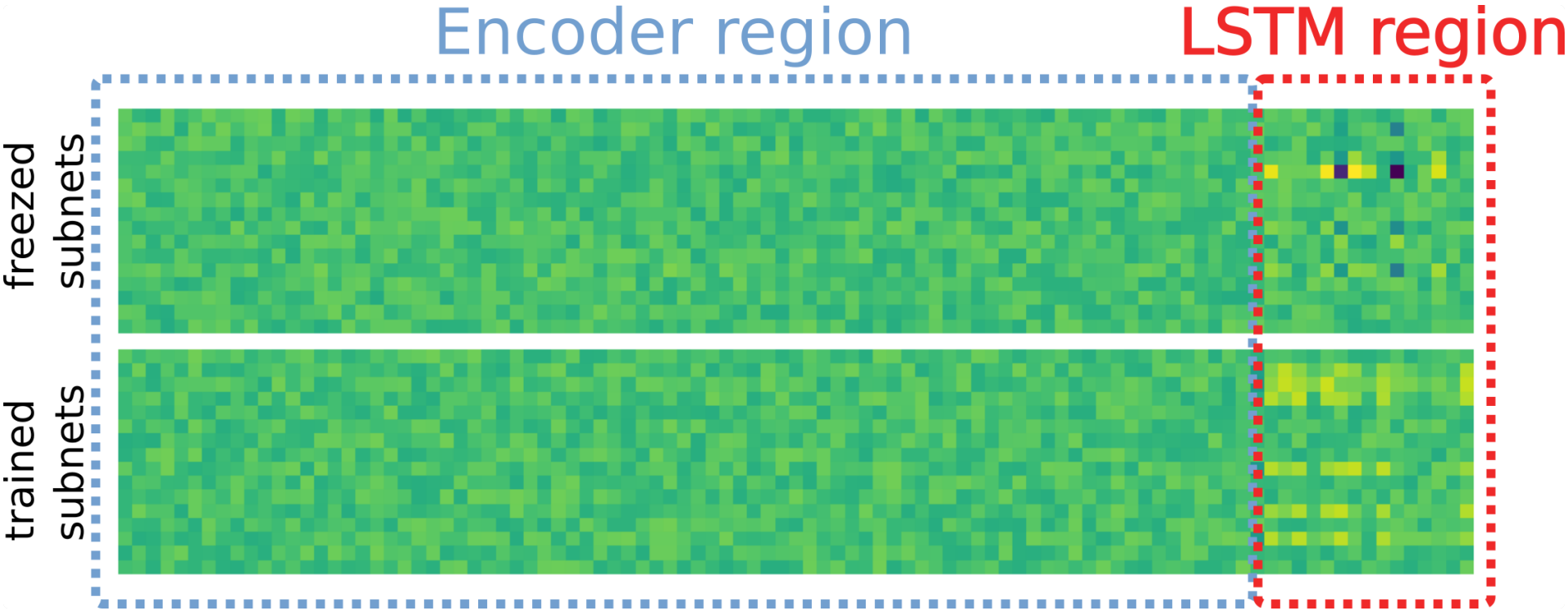
1. **Transfer Learning**
2. **Data Augmentation**
3. **Curriculum Learning**

Transfer Learning Strategy



Weights at concatenation layer

Is the encoder really useful?



Data Augmentation Strategy

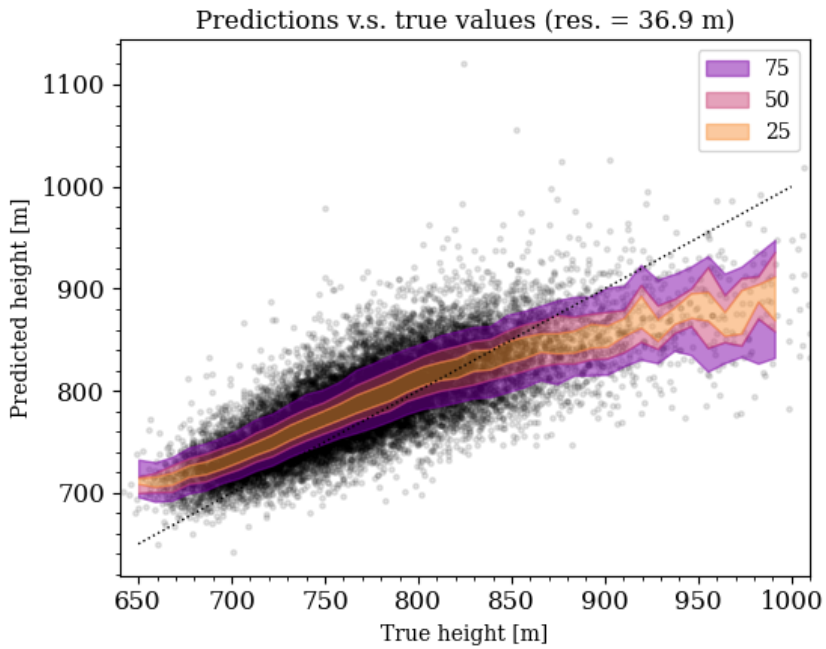
Dataset has a lack of high events ($X > 850$m) so a first network training showed a worse resolution for samples corresponding to this range

Strategy

Increase the number of samples that overcome a certain heigth threshold using the symmetries of the problem

Data is augmented using

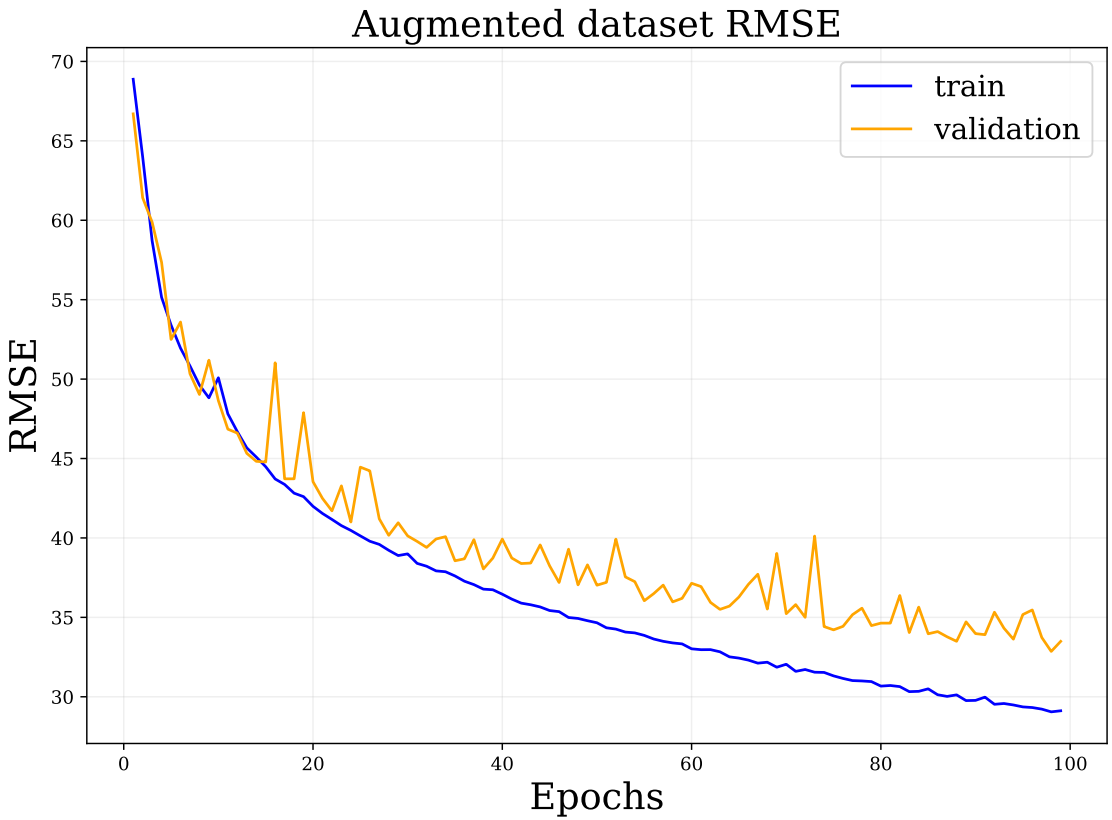
- flip up-down
- flip left-right
- diagonal flip
- rotation of 90°



Augmented dataset performance

The network trained with an augmented dataset gave the following performance on test dataset (not augmented):

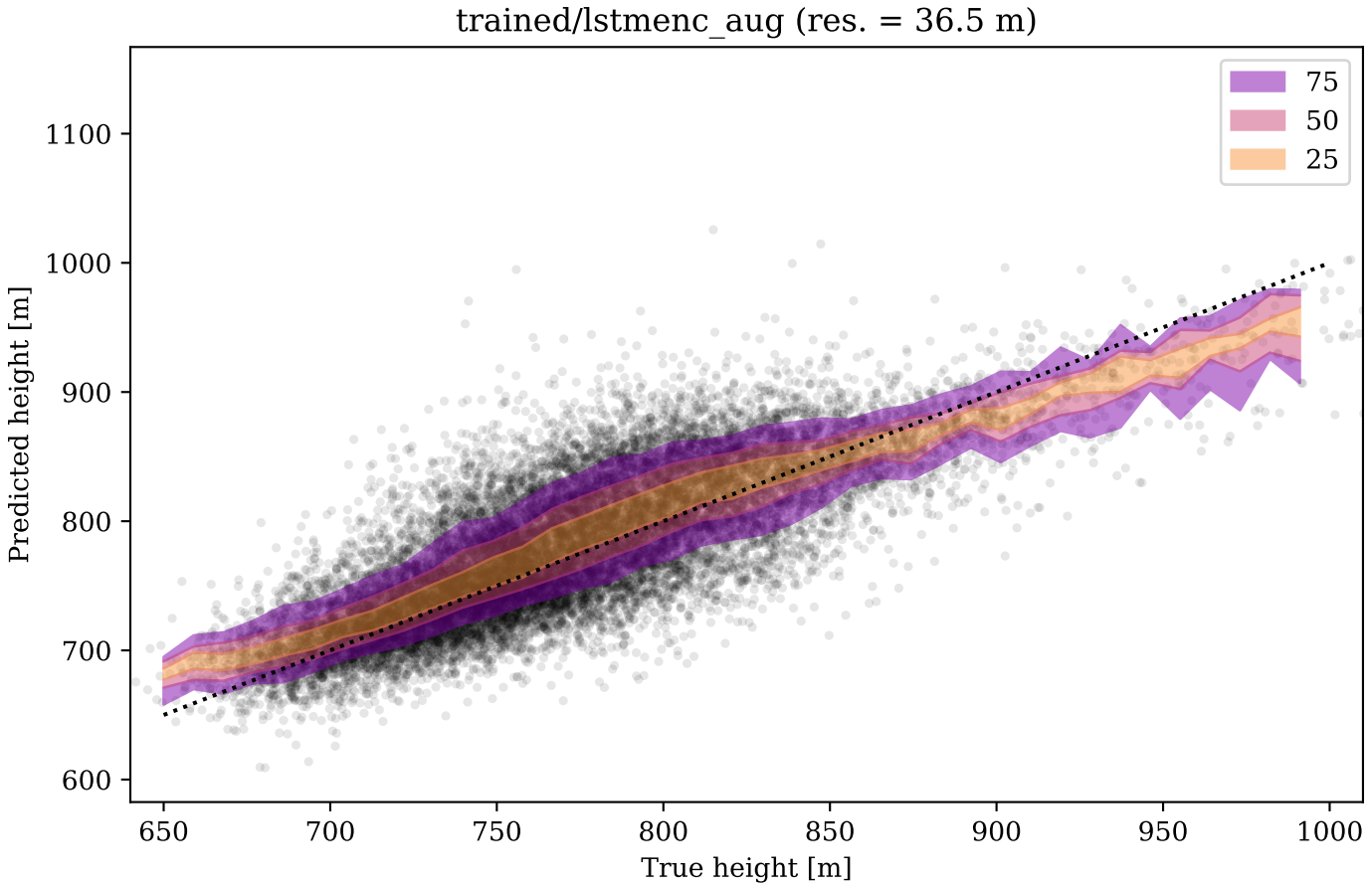
- **loss**= 1744.27
- **rmse**= 36.48



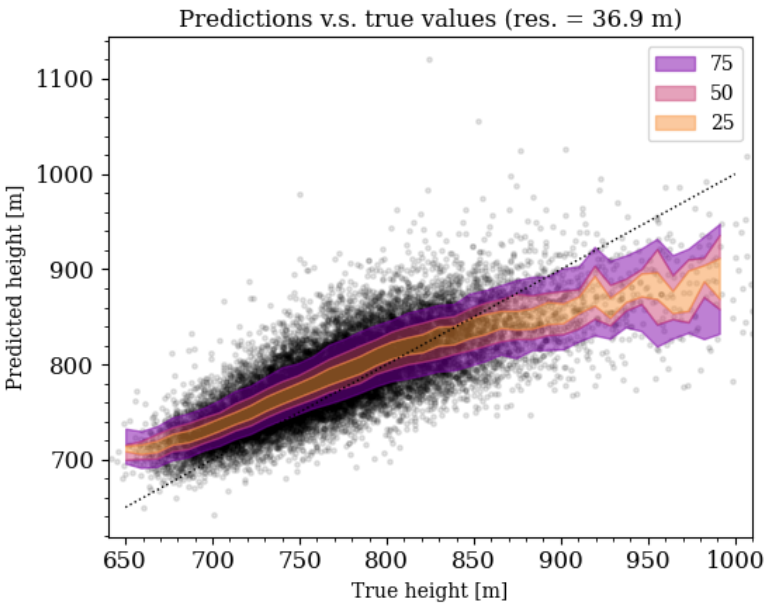
Augmented dataset performance

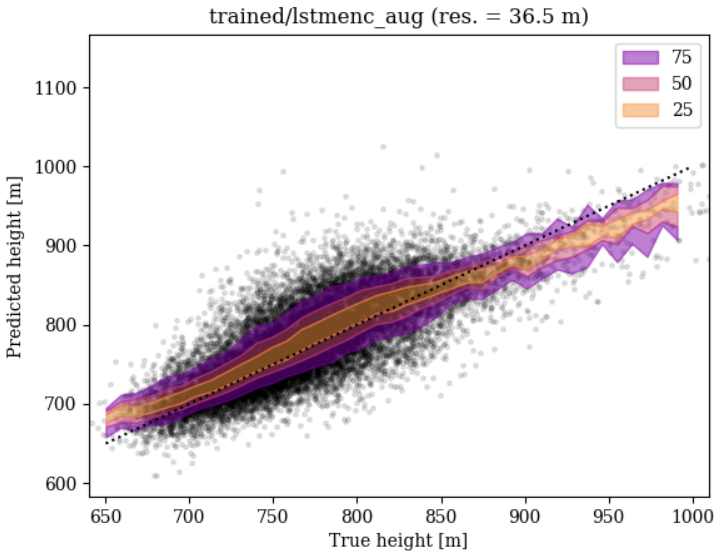
The augmented-dataset-train gave a Pearson correlation coefficient of:

$$\rho_{X,Y} = 0.80 \pm 0.02$$



Comparison with vanilla network





Curriculum Learning Strategy

- Normal training
- Use the trained NN to score the difficulty of samples
- Schedule a training for another NN

We also tried to generalize `Augment` class to augment more difficult data.

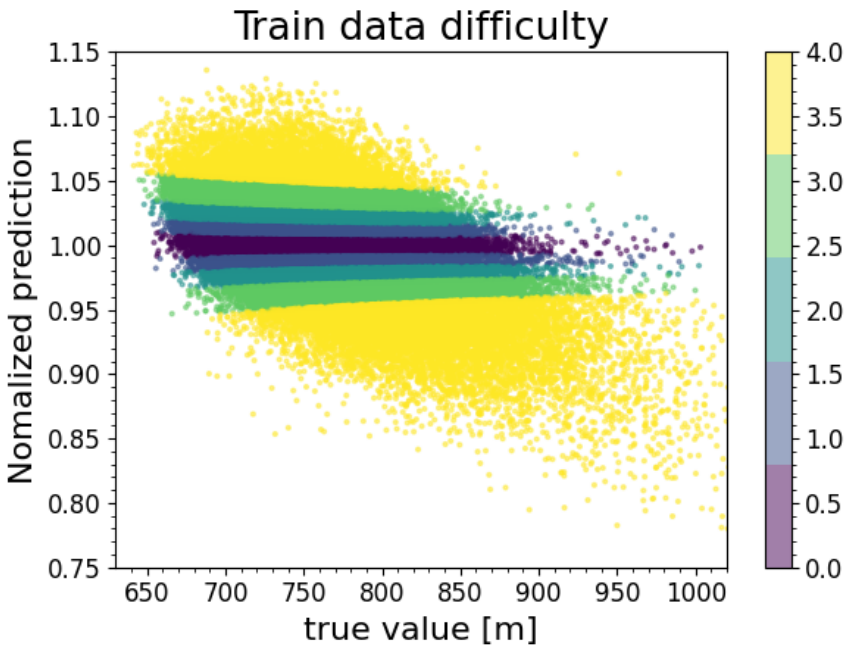
Maybe useful on classification tasks, but for regressions..

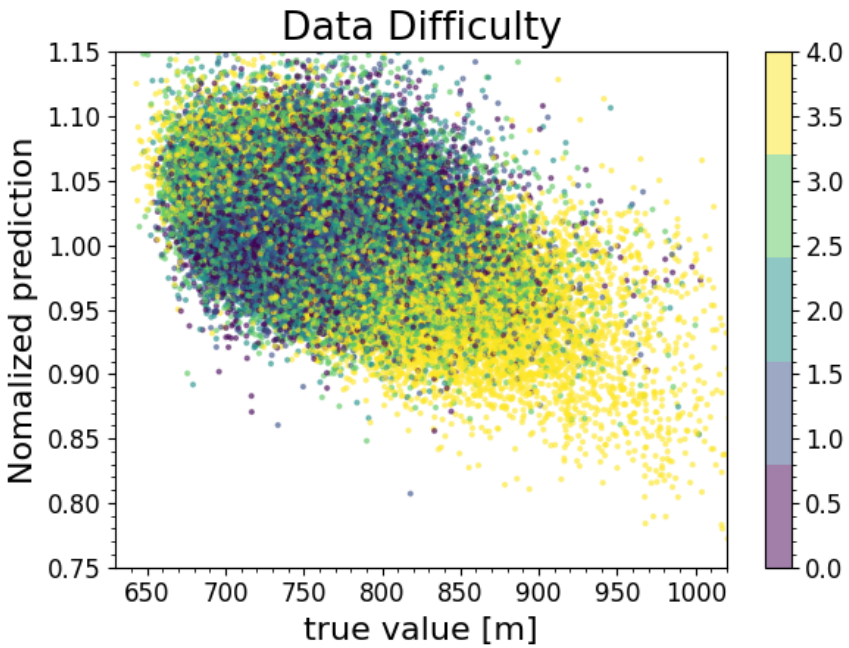
Curriculum learning

Training over "difficult" data makes the net forget the easier samples

Sort in ascending error and divide in equal-length groups that the "prof" feeds to the next network

The "student" network is trained increasing in difficulty but spends more time on hard samples





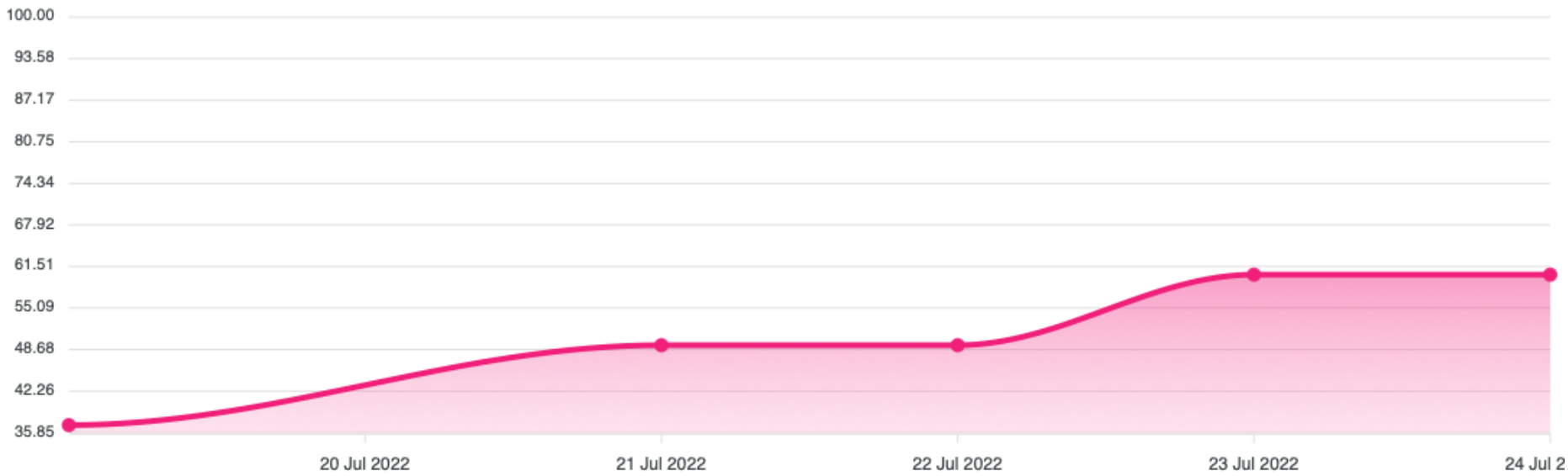
Test setup on CircleCI

We covered $\sim 60\%$ of the whole project, main test ideas are for training classes and functions with a fititious smaller dataset.
Main tests:

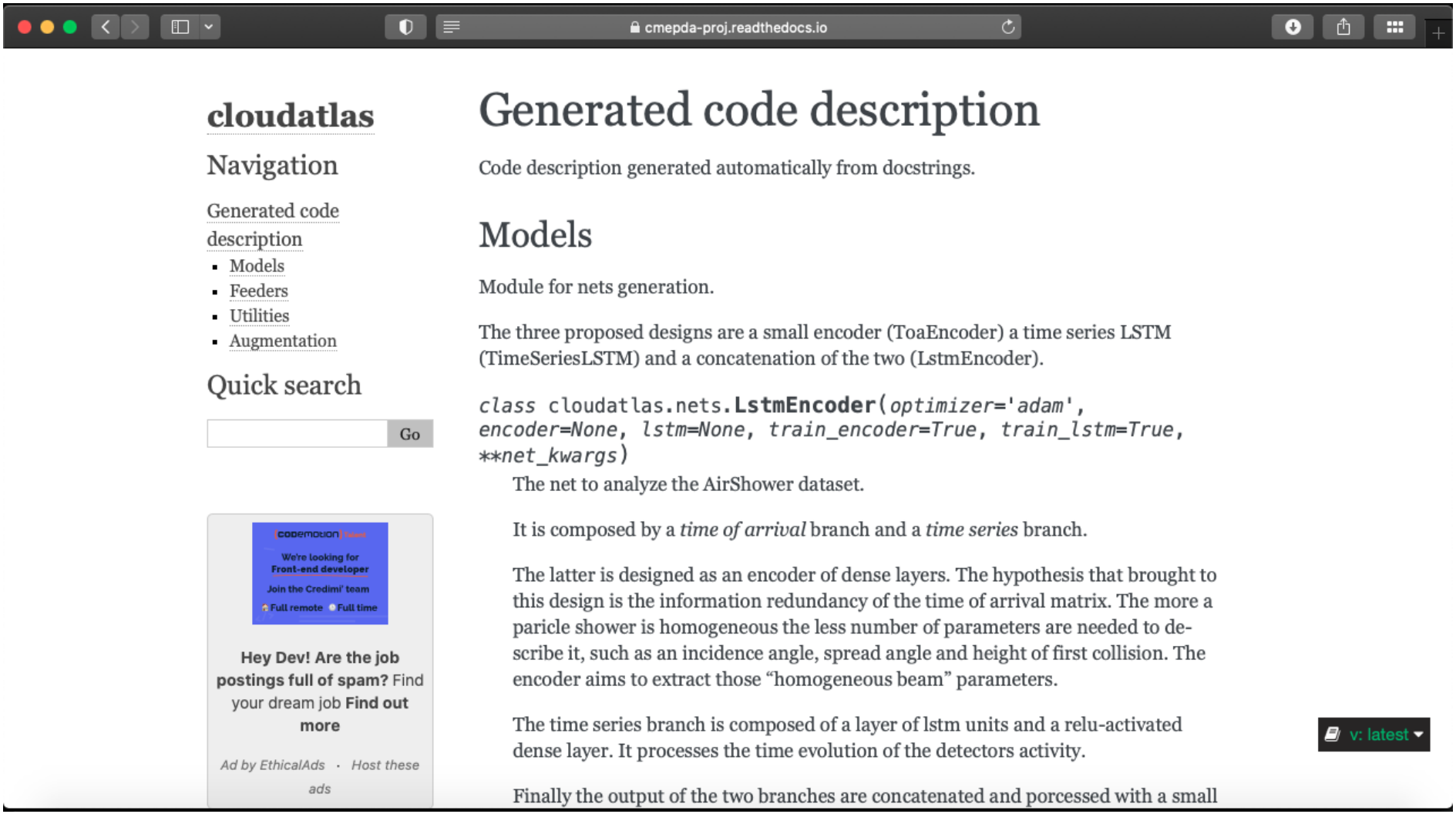
- **Net building**: checks if networks are correctly built
- **DataFeeders**: builds a fictitious dataset and checks whether `DataFeeders` work fine
- **Augmentation**: tests if an augmented matrix is effectively "rotated", "flipped", etc.

Some tests required a trained model so they were not executed in CI.

Code coverage



Documentation



Danke e bibliography

Danke Schon