COVID-19 Forecasting with California Mobility Data

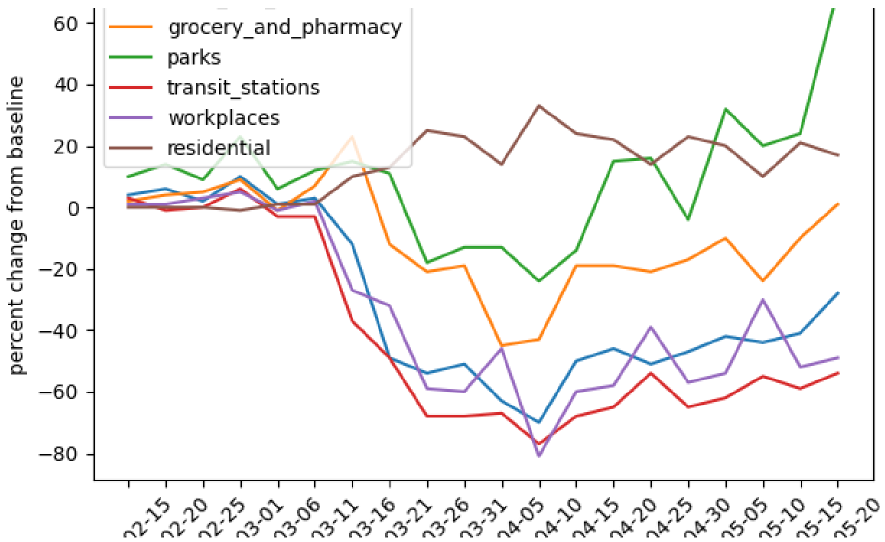
Yanshen Sun

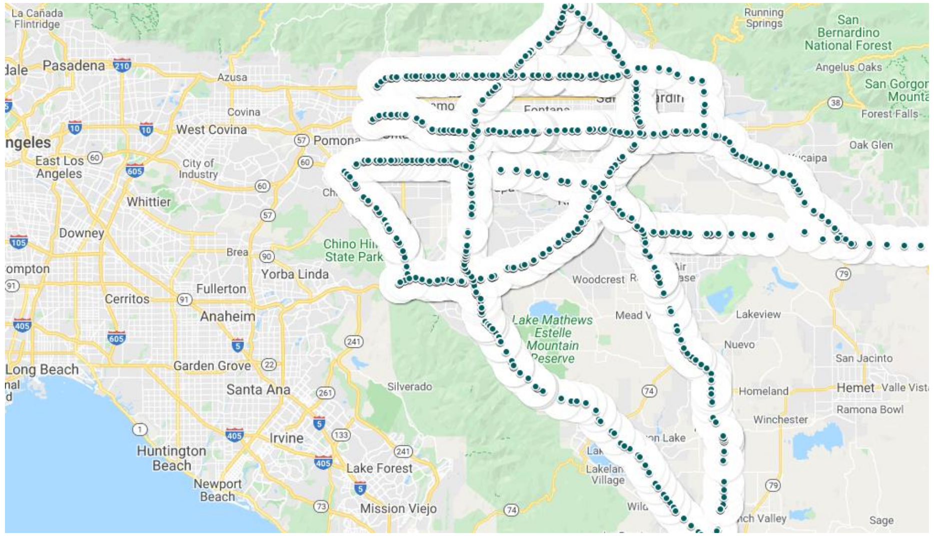
Problem statement

- California county-daily-level
- Human mobility data
  - => number of COVID cases forecasting
  - For day t, predict cases on day t+1 with day t-9, t-8, …, t-1, t

Data

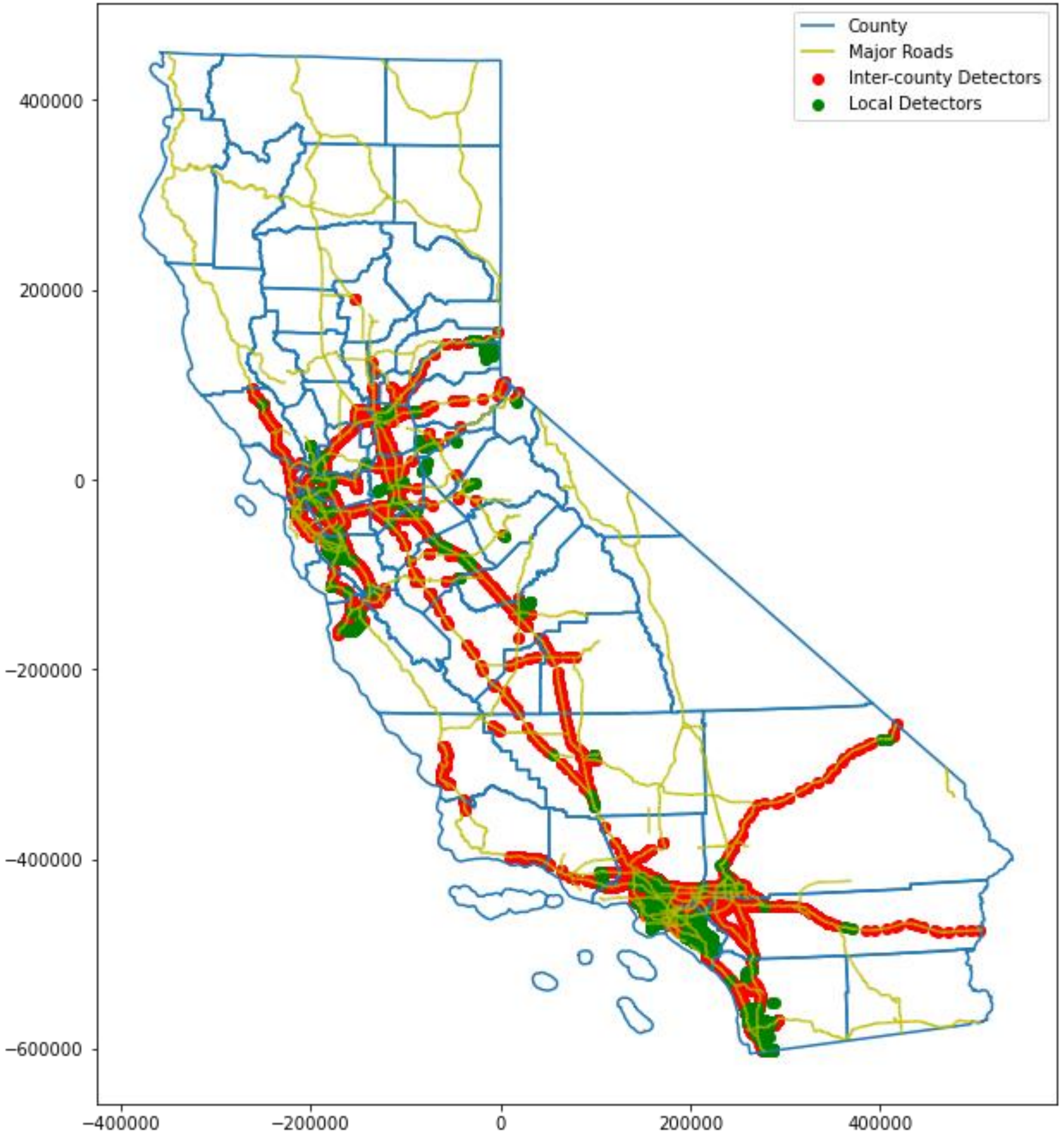
- Feb 1st ~ May 31st
- New York Times (NYT) COVID-19 dataset: daily case increment
- Google Community Mobility Report: human activities
- PeMS Caltrans traffic data: detector locations and # of cars (traffic volume)
- California county map and arterial road network
- 121 days * 58 counties





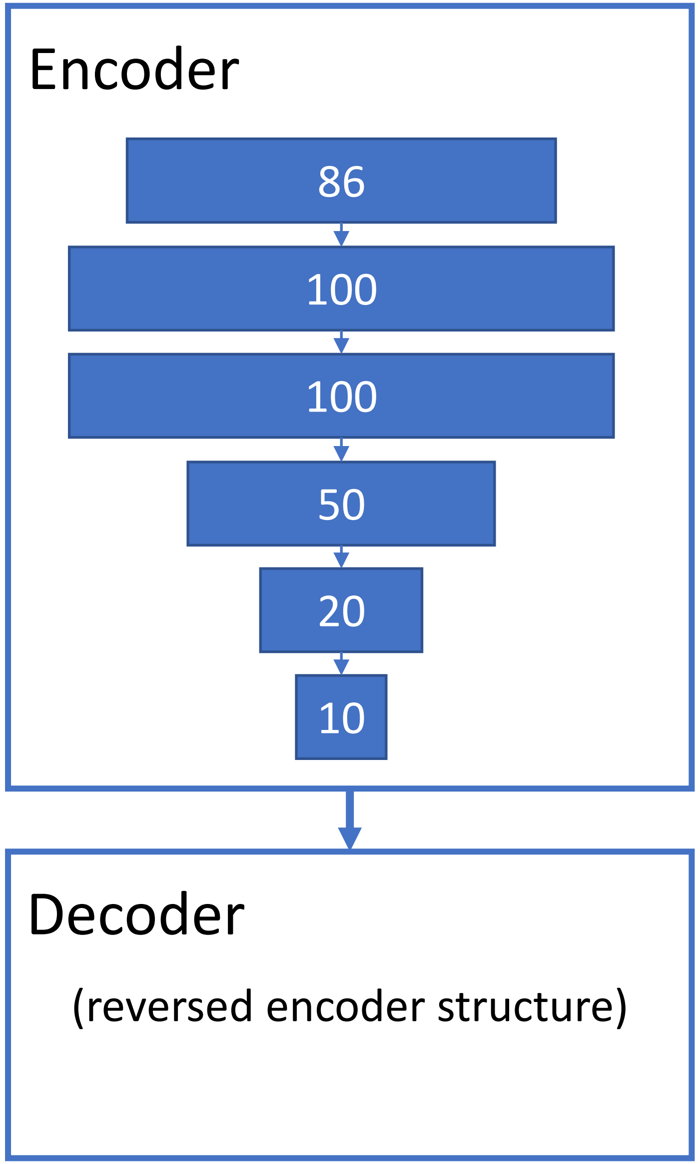
Input features

- Adjacent matrix
  - A: Adjacent counties – 0 or 1
  - B: Connected by inter-county roads – weighted by 1/distances
  - Final: A+B – rescaled to 0~1
- 86 attributes per day per county
  - Daily traffic volume (2*5*8=80)
    - Local road, inter-county (major) road
    - 4 directions + no direction
    - Count, mean, std, min, 25%, 50%, 75%, max
  - Mobility attribute (6)
- Historical Covid-19 case variation



Filling in missing features

- AutoEncoder
- 6-layer encoder, 6-layer decoder
- 100 epoch
- 0.001 learning rate
- Optimizer: SGD with momentum and weight decay
- Loss: MSE (missing values masked)
  - Train loss: 0.0542
  - Test loss: 3.72

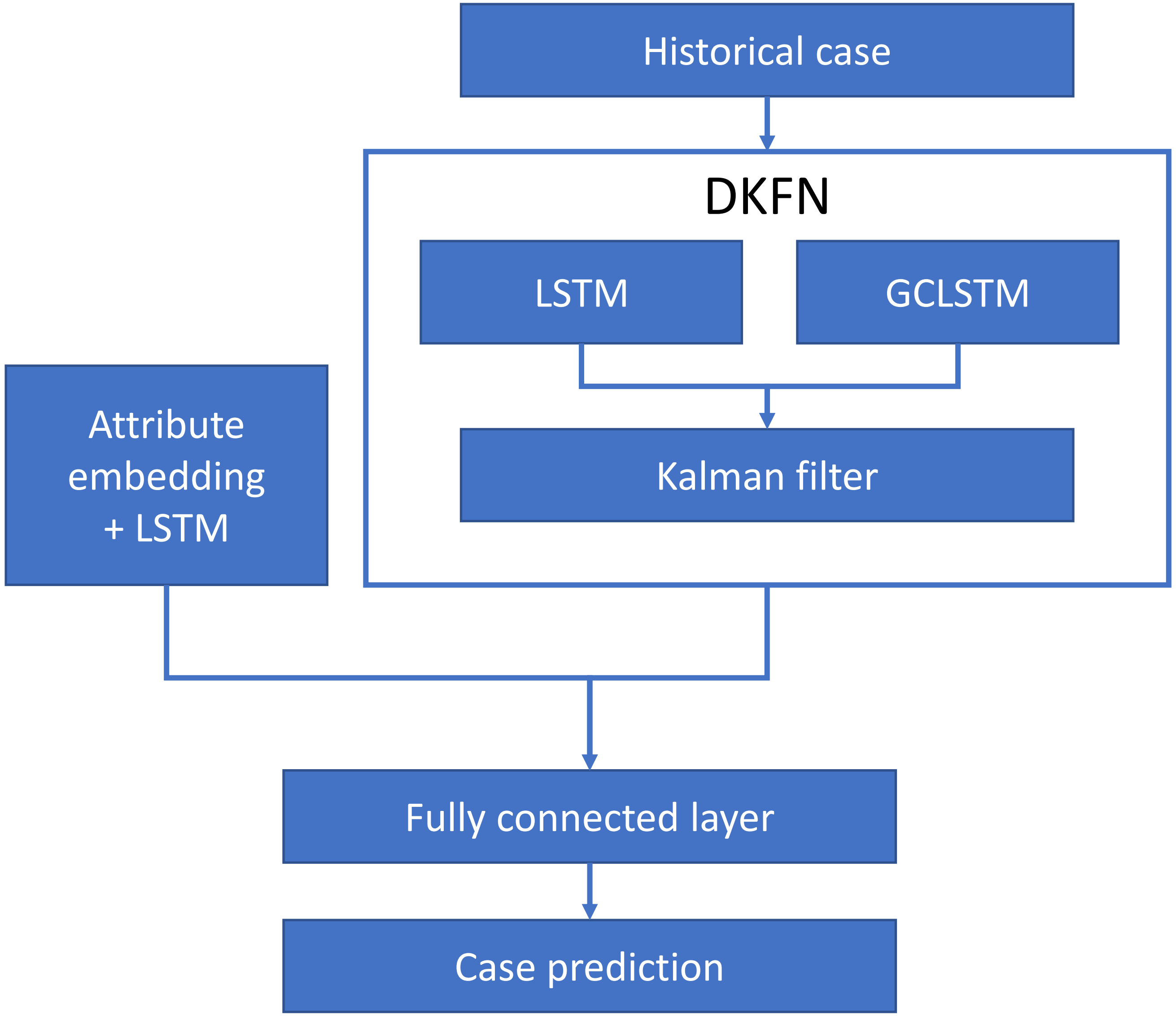


Prediction

- Attribute feature embedding and LSTM
  - Embedding: same as encoder (feature size 86->10)
  - LSTM
    - Input: 10 embedding features at each timestep
    - Output: 10 predicted features
  - FC layer: 10 -> 10 features as output
- Historical data spatial and temporal features
  - GCLSTM
    - GCN: weighted 3-NN adjacent matrix as output
    - LSTM: output 1 step prediction
  - LSTM
    - Merge with GCN+LSTM with Kalman filter

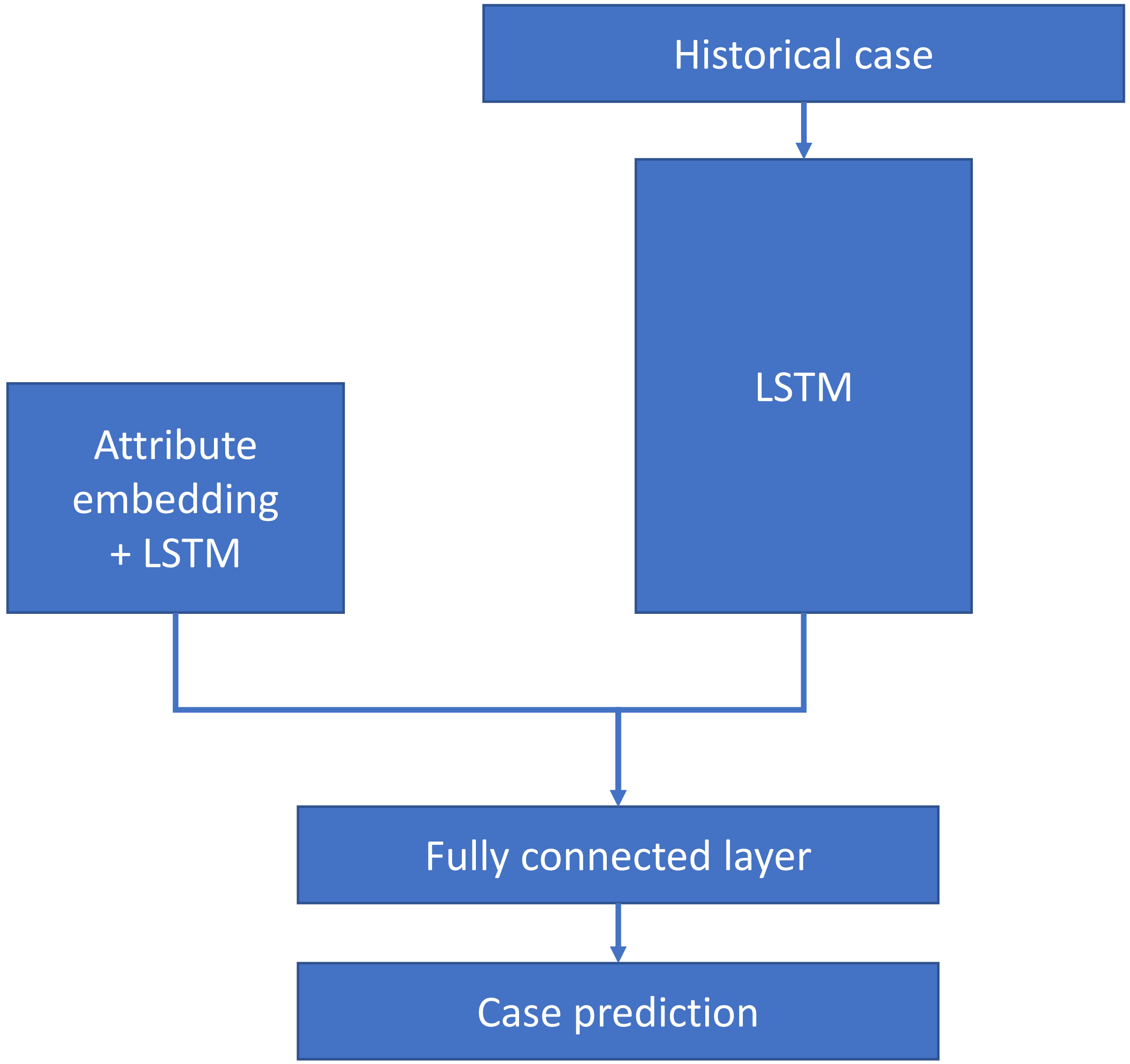
Prediction

- Node attributes embedding + DKFN[1]



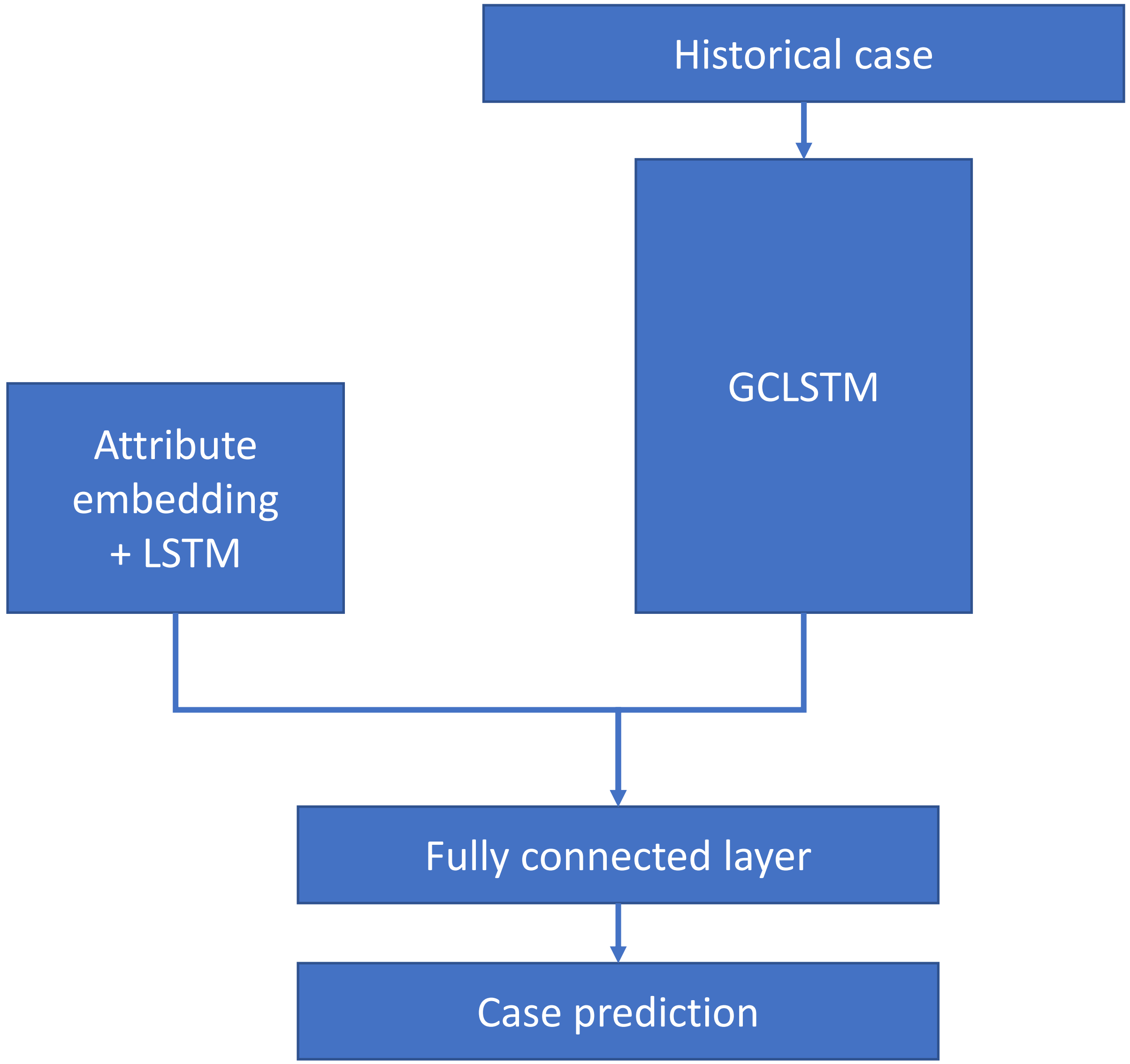
Prediction

- Node attributes embedding + LSTM



Prediction

- Node attributes embedding + GCLSTM[1]

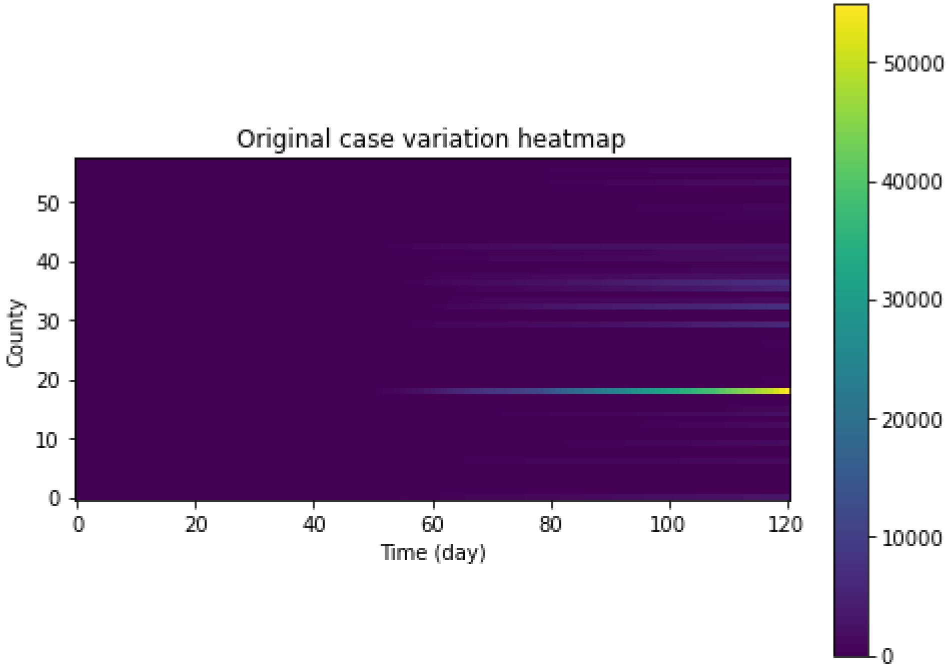


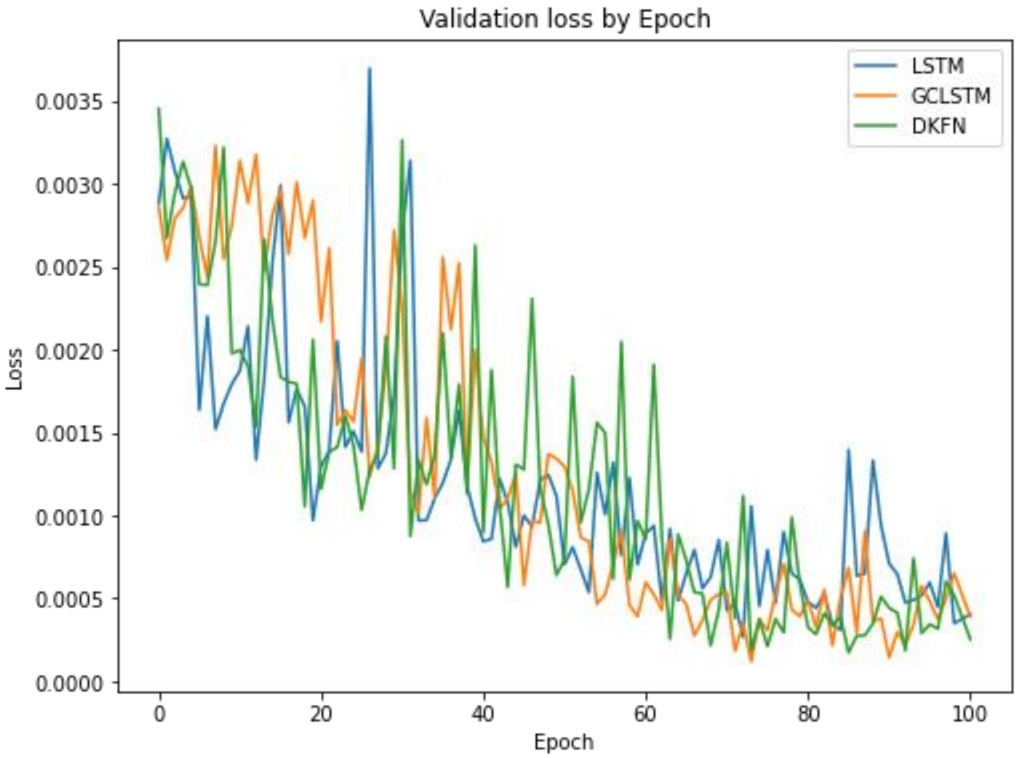
Prediction

- Experiment
  - 10-day window for prediction
  - Input: Attributes (A), Historical cases (C)
    - $\{A_{t-9}, A_{t-8}, \ldots, A_{t-1}, A_t\}, \{C_{t-9}, C_{t-8}, \ldots, C_{t-1}, C_t\}$
  - Output: case for time step t+1 (C_{t+1})
  - Predict #10 - #120 day (110 days): 80 train + 8 validation + 20 test
  - Adam optimizer (learning-rate 0.0001)
  - MSE Loss
  - Batch size is 4

Results

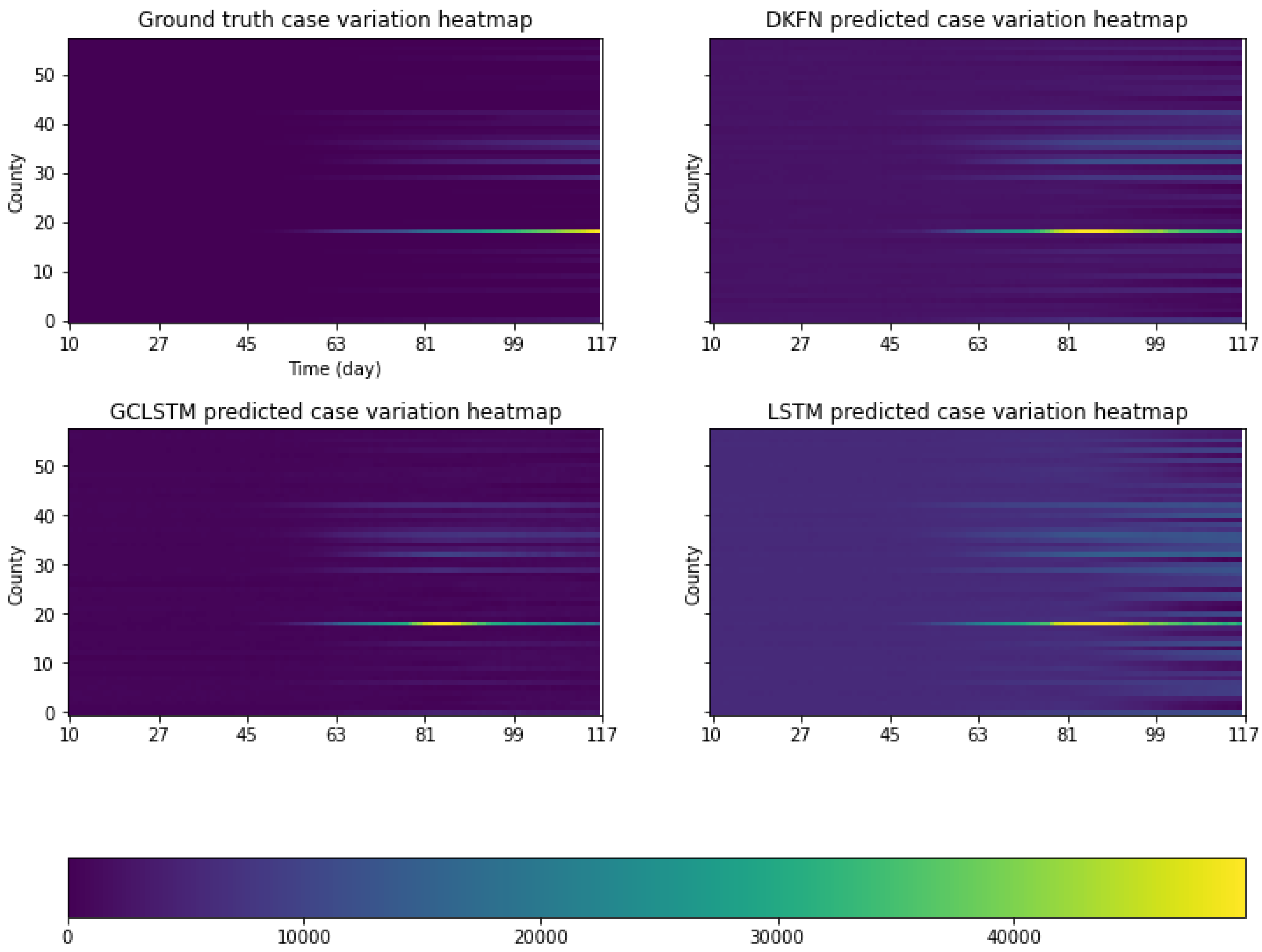
	MAE	RMSE	R^2
LSTM	1631.44	4198.08	0.42
GCLSTM	1109.66	4506.14	0.33
neDKFN	1335.57	4926.02	0.13
DKFN	**1015.51**	**3808.50**	**0.52**





Results

- The county got most cases:
  - Los Angeles County (54996)
- Trend is correct
- Counties of high occurrence is correct
- Values of cases are close
- Peak is earlier



Conclusion

- Goals achieved
  - Spatial-temporal model performs better than spatial/temporal models
  - Spatial relations of counties does improve the results
- Drawbacks
  - Accumulated error produced by the autoencoder
  - Lacking training and testing data
  - Better way to merge county attributes with historical data
  - Need more proof that mobility attributes are related to the results

Reference

- [1] Chen, F., Chen, Z., Biswas, S., Lei, S., Ramakrishnan, N., & Lu, C. T. (2020, November). Graph Convolutional Networks with Kalman Filtering for Traffic Prediction. In Proceedings of the 28th International Conference on Advances in Geographic Information Systems (pp. 135-138).

Thank!