Uncertainty Quantification for Traffic Forecasting

Weizhu QIAN

Soochow University, China

September 14, 2023

Outline

1. Motivation
2. Problem Statement
3. Method
    - Aleatoric Uncertainty
    - Epistemic Uncertainty
    - Model Calibration
    - Theoretical Analysis: A PAC-Bayes Perspective
4. Experimental Results
5. Conclusion

Motivation

Motivation: Uncertainty quantification can estimate the possible minimum and maximum values of the predicted traffic flow/speed/volume. Such **reliability** information can be imperative for municipalities to manage urban traffic system in some critical real-world scenarios (e.g., emergency rescue and disaster evacuation), where unreliable point forecasting may lead to catastrophic consequences.

This issue is essential to Intelligent Transportation Systems (ITS) but has not been well explored.

Figure 1: Urban traffic flow, source: *https://accoladetechnology.com*.

Problem Statement

Goal: We aim to obtain both point prediction and uncertainty quantification (lower and upper bounds of the prediction) for the spatio-temporal time series prediction.

To this end, we will

1. Model spatio-temporal correlations;
2. Model uncertainties;
3. Reduce generalization gap (via model calibration).

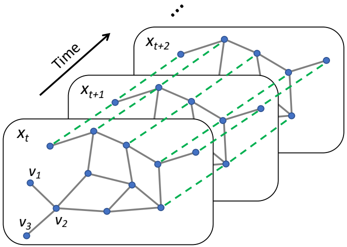


Figure 2: Traffic data can be described by graphs (nodes: sensors in road network, grey lines: spatial dependency, green dashed lines: temporal dependency. This is a multivariate time series forecasting problem.

Proposed Method: A Unified Approach

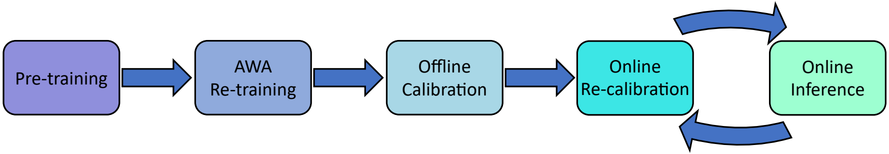


Figure 3: Pipeline of the proposed method.

Purpose of each step:

1. Pre-training: obtain crude predictive means and intervals;
2. Re-training: obtain better epistemic uncertainty estimation;
3. Offline calibration: mitigate the overconfidence of the trained model;
4. Online calibration: re-strengthen i.i.d. assumption.

Step 1: Modelling Spatio-Temporal Dependency

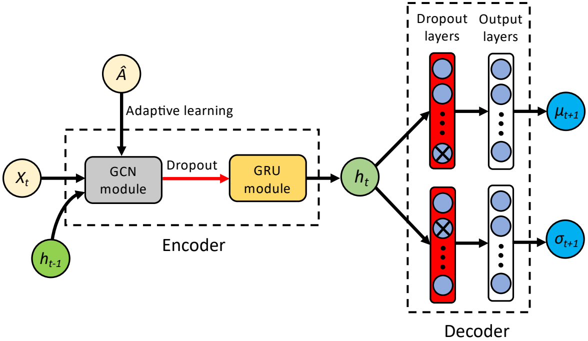


Figure 4: Architecture of the model.

1. Graph Convolutional Network (GCN): to model **spatial** correlations between sensors;
2. Gated Recurrent Unit (GRU): to model **temporal** correlations between time steps;
3. Dropout operations: to induce model uncertainty;
4. Output layers: to estimate aleatoric uncertainty.

Step 2: Estimating Uncertainty

The total uncertainty can be decomposed as follows:

$$\sigma^2_{\text{Total}} \approx \underbrace{\mathbb{E}_{\theta\sim p(\theta)}[\sigma^2_\theta]}_{\text{Aleatoric/data uncertainty}} + \underbrace{\mathbb{V}_{\theta\sim p(\theta)}[\mu_\theta]}_{\text{Epistemic/model uncertainty}}$$

1. The aleatoric uncertainty caused by sensor noise or system randomness can be modelled by estimating Gaussian likelihoods;
2. The epistemic uncertainty caused by lack of data or model misspecification can be modelled by variational inference and ensembling.

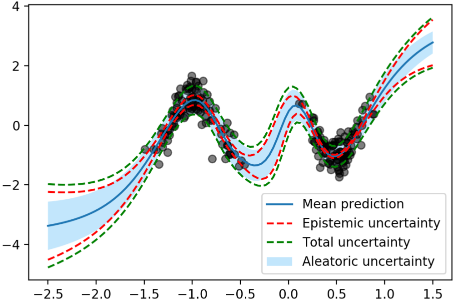


Figure 5: Different types of uncertainties, source: *https://arxiv.org/pdf/1905.09638.pdf*.

Step 2.1: Estimating Aleatoric Uncertainty

Assume the conditional likelihood of each node at each step is Gaussian:

$$\theta = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log \mathcal{N}(\hat{X}^{i}_{>t}; \hat{\mu}_{\theta}(X^{i}_{<t}), \hat{\sigma}_{\theta}(X^{i}_{<t})^{2}), \tag{1}$$

where θ is the model parameters.

Corresponding hybrid loss function for numerical stability

$$\mathcal{L}_{\text{Aleatoric}} = \frac{1}{N} \sum_{i=1}^{N} \lambda \left\{ \log(\sigma(x_i)^2) + \frac{(y_i - \mu(x_i))^2}{\sigma(x_i)^2} \right\} + (1-\lambda)|y_i - \mu(x_i)|, \tag{2}$$

where λ is the relative weight with $0 < \lambda \leq 1$.

Note that Gaussianity assumption is reasonable for regression problems but still too strong in our case, it can be relaxed by using conformal calibration method.

Step 2.2: Estimating Epistemic Uncertainty

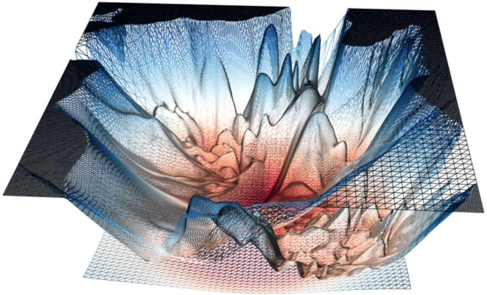


Figure 6: The highly non-convex loss landscape of a neural network, source: *Visualizing the Loss Landscape of Neural Nets*.

Learning strategy: combine the advantages of variational inference and deep ensembling to estimate epistemic uncertainty. **Essentially, we are performing Monte Carlo Integration:** $\mathbb{E}_w[f(x;w)]$**.**

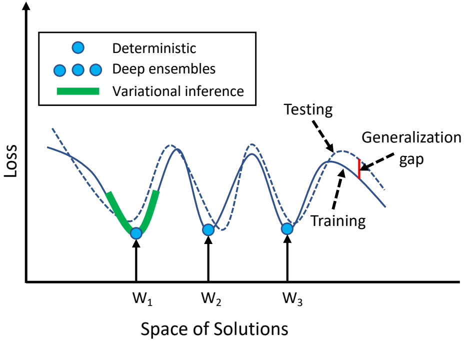


Figure 7: Deterministic: model trained by SGD, ensembles: set of Deterministic models, variational inference: uncertainty around uni-modal local minimums.

Step 2.2: Estimating Epistemic Uncertainty (Variational Inference)

Bayesian Neural Network (BNN): each weight is a Gaussian

$$D_{\mathrm{KL}}(q(w)||p(w|D)) = \int q(w)\log\frac{q(w)}{p(w)p(D|w)}dw = D_{\mathrm{KL}}(q(w)||p(w)) - \mathbb{E}_{w\sim q(w)}[\log p(D|w)], \tag{3}$$

where $p(w)$ is the prior and $\log p(D|w)$ is the predictive log-likelihood.

Monte Carlo Dropout (MCDO) as Bayesian inference

The dropout operations are used to induce uncertainty in **weight space** as it is flexible and fast.

$$q(W_i) = M_i \cdot (\mathrm{diag}[z_{i,j}]_{j=1}^{K_i}), \tag{4a}$$

$$z_{i,j} \sim \mathrm{Bernoulli}(p_i), \tag{4b}$$

$$\mathcal{L}_{\mathrm{Dropout}} = \mathbb{E}_{w\sim q(w)}E[Y, f_W(X)] + D_{\mathrm{KL}}(q(w)||p(w)) \approx \frac{1}{N}\sum_{i=1}^{N}E(y_i, f(x_i, w_i)) + \frac{\lambda_W}{2p_i}||w_i||^2. \tag{5}$$

Other approaches: Bayes By Backprop, Stochastic Gradient Langevin Dynamics, Hamiltonian Monte Carlo, etc.

Step 2.2: Estimating Epistemic Uncertainty (Deep Ensembling)

Standard ensembling needs to train multiple models, so it is slow. To tackle this, we proposed the Adaptive Weight Averaging (AWA) re-training process, the learning rate varies in an annealing manner to explore different local minimums (different weights in solution space) and average them to approximate ensembling.

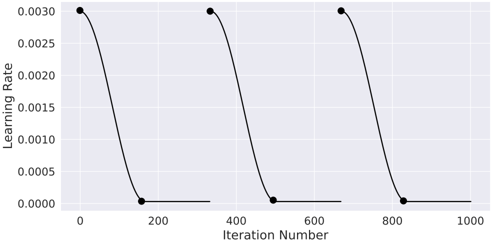


Figure 8: Learning rate change.

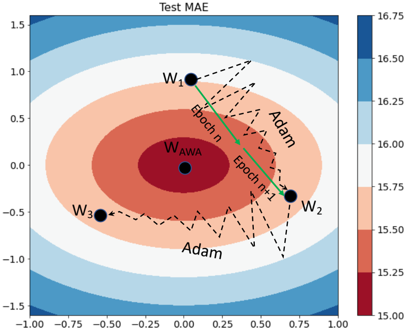


Figure 9: Test performance and weight space.

Aadm: approximate Fisher information by computing second moment of the gradients (therefore provide better metric for log-Gaussian loss than SGD).

Step 3: Calibrating Trained Model (Offline)

Can we do better even after re-training?
Remaining issues: (1) Gaussian likelihood assumption; (2) step-wise coverage, (3) overconfidence.

Distribution-Free Assumption: Conformal Inference

Conformal Inference uses the quantile of the errors with some target significance level α on the validation dataset to quantify uncertainty. Ideally, the empirical horizon-wise prediction interval coverage percentage (PICP), p_c^h should be greater than or equal to α.

$$p_c^h = \frac{1}{N_{\text{Cali}}} \sum_{i=1}^{N} k_i^h, \tag{6}$$

where N_{Cali} is number of datapoints of the calibration dataset, $k_i^h = 1$ if $\hat{X}_i^h \in C(X_i^h)$, otherwise, $k_i^h = 0$.

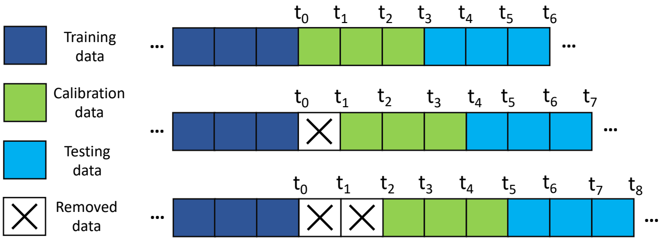
The horizon significance level α_c^h for calibration can be corrected by the proposed empirical equation:

$$\alpha_c^h = (p_c^h + 2\alpha - 1) + \gamma(p_c^1 - p_c^H)(h-1)^2, \tag{7}$$

where γ is a positive scalar.

Step 3: Calibrating Trained Model (Online)

Finally, the trained model is calibrated on the calibration/validation dataset using **conformal inference** (leveraging the quantile of the calibration error) to relax the Gaussian likelihood assumption and mitigate the overfitting issue.



Training data
Calibration data
Testing data
Removed data
t_0 t_1 t_2 t_3 t_4 t_5 t_6
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8
...

Figure 10: Online data splitting.

We can update the calibration dataset to re-strengthen the i.i.d assumption to proceed the calibration in an online learning manner.

Proposed Method: A Unified Approach

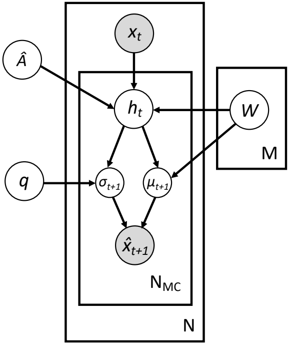


Figure 11: Graphical representation of the model (circles: variables; arrows: dependencies; plates: samples).

The final predictive means and variance can be computed as follows:

$$\hat{\mu}_{t+1} = \frac{1}{N_{MC}} \sum_{j=1}^{N_{MC}} \mu^j(x_t), \tag{8a}$$

$$\hat{\sigma}^2_{t+1} = q^2 \sum_{j=1}^{N_{MC}} \frac{\sigma^j(x_t)^2}{N_{MC}} + \sum_{j=1}^{N_{MC}} \frac{\left(\mu^j(x_t) - \hat{\mu}_{t+1}\right)^2}{N_{MC} - 1}, \tag{8b}$$

where $\hat{\mu}_{t+1}$ is used as the point prediction of the proposed approach, and q is the scalar obtained by calibration.

Theoretical Analysis: A PAC-Bayes Perspective

Probabilistic Approximate Correct (PAC)-Bayes theory

We first assume the datapoints in the training, calibration, and testing datasets are all i.i.d. Once the model architecture is specified, we have the hypothesis/parameter space $\mathcal{H}$. Consequently, the learning goal is to obtain a hypothesis $h \sim \mathcal{H}$. Let D be some unknown data distribution over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ be the loss function. Then true risk is defined as follow:

$$R(h) = \mathbb{E}_{(x,y)\sim D}\Big[\mathcal{L}\big((x,y),h\big)\Big]. \tag{9}$$

Theorem

For all probability measure Q supported on $\mathcal{H}$, with at least probability of $1-\delta$, the following PAC-Bayes bound holds :

$$R(h) \leq r(h) + \frac{1}{N_{Train}}\big(D_{KL}(Q||P) + \log(1/\delta)\big) + const \tag{10}$$

This is the reason why we model epistemic uncertainty $D_{\mathrm{KL}}(Q||P)$ to reduce generalization gap during training.

Theoretical Analysis: A PAC-Bayes Perspective

Once the training process is finished, a fixed hypothesis h is rendered. Let $D_{\text{Cali}} \sim D$ be a held-out calibration dataset, and h is independent of D_{Cali}. Then, we have

Theorem

With at least probability of $1-\delta$, the following validation PAC bound holds:

$$R(h) \leq r_{D_{Cali}}(h) + \sqrt{\frac{1}{2N_{Cali}} \log(2/\delta)}, \tag{11}$$

which means the model performance on the calibration dataset is close to the real model performance.

This is the reason why we leverage conformal calibration on the calibration dataset to reduce generalization gap.

Experimental Results

The proposed method is tested on 4 public traffic datasets, PEMS03, PEMS04, PEMS07, and PEMS08.

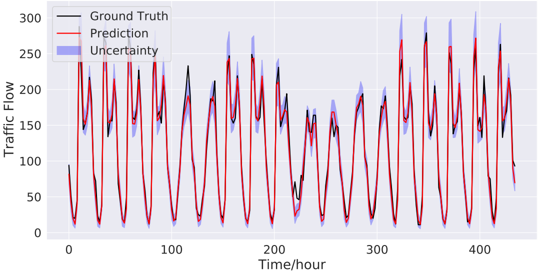


Figure 12: Uncertainty quantification results.

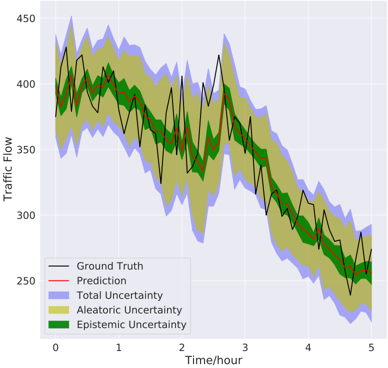
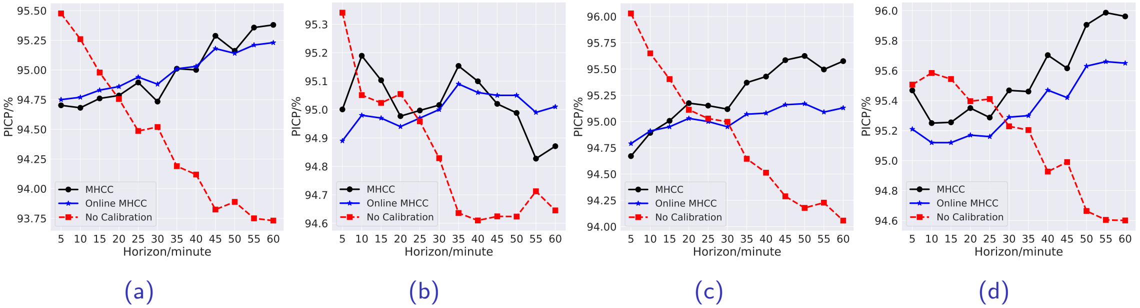


Figure 13: Different types of uncertainties.

Experimental Results

The proposed Multi-Horizon Conformal Calibration method improves the horizon-wise uncertainty quantification performance.



(a) (b) (c) (d)

Figure 14: Calibration results with respect to different forecast horizons.

Conclusion

The contributions are summarized as follows:

- We propose a hybrid loss to estimate the aleatoric uncertainty;
- We propose a training method combining the advantages of variational inference and deep ensembling;
- We propose a novel calibration method based on conformal inference;
- We provide a theoretical analysis based on the PAC-Bayes theory.

References

- **Weizhu QIAN**, Dalin Zhang, Yan Zhao, Kai Zheng and James J.Q. Yu, *Uncertainty Quantification for Traffic Forecasting: A Unified Approach*, 39th IEEE International Conference on Data Engineering, ICDE 2023.
- **Weizhu QIAN**, Yan Zhao, Dalin Zhang, Bowei Chen, Kai Zheng and Xiaofang Zhou, *Towards A Unified Understanding of Uncertainty Quantification in Traffic Flow Forecasting*, IEEE Transactions on Knowledge and Data Engineering (TKDE) 2023.

Thank you for your attention!