Using Actively-learned Deep kernels to estimate epistemic uncertainty

Hena Ghonia
Department of Computer Science and Operations Research
Mila - Quebec AI Institute
Universite de Montreal
Montreal, Canada

Shanaya Mehta
Assert AI
Bangalore, India

Abstract

Accurately estimating the level of uncertainty plays a pivotal role in determining the reliability and effectiveness of machine learning models across a wide range of applications. The proposed approach leverages active learning techniques to select informative data points for labeling and incorporates scalable deep kernels to capture complex relationships in the data. By actively identifying and prioritizing samples that are likely to reduce uncertainty the most, our approach significantly enhances the efficiency of uncertainty estimation. We perform empirical evaluation on normal Deep kernel based method and Deep kernel learned in active learning framework to estimate epistemic uncertainty in regression tasks.

1 Introduction

Active learning is a machine learning approach that seeks to enhance the learning process's efficiency by guiding iteratively to select next unexplored data points using appropriate acquisition function.(Kumar & Gupta, 2020) Historically, active learning has primarily been employed in supervised learning situations where acquiring labeled data is limited or costly. However, it has also found applications in estimating epistemic uncertainty, which is crucial for reliable decision-making in many real-world domains.(Mamun et al., 2022). Epistemic uncertainty refers to the uncertainty arising from the lack of knowledge about the true underlying data distribution.

There has been increasing research interest in Gaussian processes after (Neal, 2012) demonstrated the characteristics of priors for infinite networks, showing that when the number of hidden neurons increases to infinity, the features of a neural network with one hidden layer converge to those of a Gaussian process. In order to combine the representational strength of neural networks with the accurate uncertainty estimates provided by Gaussian processes, Deep Kernel Learning (DKL) and related approaches have been developed (Ober et al., 2021). Adaptive deep kernel learning (Tossou et al., 2019) technique allows better performance than a single deep kernel learning deciding which kernel to use for each task during inference. Illustrated this by comparing it to current state-of-the-art algorithms using real-world, few-shot regression problems from the area of drug development. As a result, it is well suited for complicated task distributions in a few-shot learning scenario. It is feasible to establish a common prior that induces knowledge transfer if one has a lot of tiny but related tasks, as in few-shot learning. Given a new, unknown task, it is feasible to efficiently estimate the posterior distribution across a query set conditioned on a limited support set by using a deep kernel prior with parameters shared across tasks (Patacchiola et al., 2020).

Our method consists of Actively learned deep kernels which has Long short-term memory (LSTM) and feed forward network as feature extractor architectures with covariance kernel (Wilson & Adams, 2013) to create expressive and scalable closed form kernels (Wilson & Nickisch, 2015) that can be trained collectively with a single supervised goal within a non-parametric Gaussian process framework, all without the need for approximate Bayesian inference. By learning a set of covariance functions, we provide enhancements over conventional Deep kernel transfer for regression dataset. By selectively labeling informative samples, this approach enables better understanding and quantification of uncertainty, leading to more reliable and informed decision-making in real-world applications.

2 Approach

At the query stage in Figure 1(a), candidates are chosen from the pool of unlabeled samples based on pairwise distance and the variance reduction method. The chosen samples are checked in the lab and the results are added to the training data. This makes it easier to draw conclusions about the unlabeled samples in the pool. Variance reduction approach is used i.e. selecting the most uncertain sample by computing similarity score using pairwise distance between training batch and unlabled batch. In this workflow, base model is Deep kernels shown in Figure 1(b) which combines of feature extractor and gaussian process which selects the most useful samples from a pool of unlabeled samples. In active learning workflow, at the query stage in Figure 2, candidates are chosen from the pool of unlabeled samples based on pairwise distance and the variance reduction method.

For active learning, 20% data were used as test data and then 60% of the remaining data were used as the initial training data. Pairwise distance was used batch wise to select query points from remaining 20% of data(train_hold or unlabeled samples)Mamun et al. (2022):

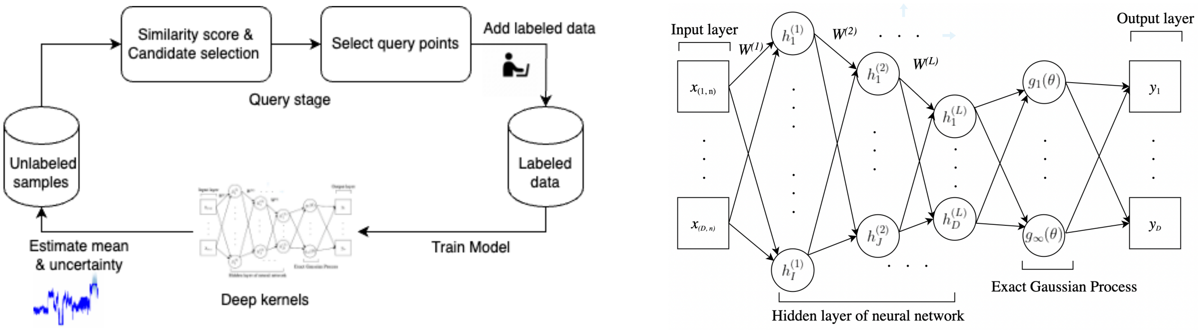
$$U = \sqrt{\text{cov}\left(\mathbf{f}_*\right)\dot{} E\left[\mathbf{f}_*\right]^2} \qquad (f_* \text{ from Gaussian process perspective 2.2.1})$$

$$\text{sim} = \text{min\_pairwise\_distance(train\_batch, train\_hold)}$$

$$\text{scores} = \frac{1}{n}\sum\left(\alpha * \frac{1}{1+sim} + (1-\alpha) * \tanh(U)\right)$$

$$\text{Select\_batch\_query} = \text{argmax(scores)}$$

2.1 Active learning workflow



(a) Batch active learning cycle

(b) Architecture of Deep kernel

Figure 1: Active learning workflow

2.2 Deep kernels

To begin with, we conceptualise a deep kernel learning problem for regression task and discuss why it is more expressive than an existing deep learning problem in which we train a neural network on a multivariate attributes to learn meaningful feature representation by learning highly basis adaptive functions, and then apply a Gaussian process with a set of kernel functions. Through the marginal likelihood of a Gaussian process, we learn about the characteristics of these kernels.Wilson et al. (2016) One of the reason we choose Long Short Term Memory(LSTM) network as our feature extractor is due to time feature present in all the dataset for empirical evaluation and LSTM tends to perform well on time based datasets.

2.2.1 Gaussian Process

Gaussian process (GP) can be thought of as the generalization of a Gaussian distribution over a infinite vector space to a function space of infinite dimensionMacKay et al. (1998). Covariance kernel: The covariance function defines closeness or resemblance in the context of the Gaussian process perspective.

$$\mathbf{f} = f(X) = \left[f\left(\mathbf{x}_1\right), \ldots, f\left(\mathbf{x}_n\right)\right]^\top \sim \mathcal{N}\left(\boldsymbol{\mu}, K_{X,X}\right)$$

where $\boldsymbol{\mu}_i = \mu(x_i)$ and covariance matrix $(K_{X,X})_{ij} = k_{\boldsymbol{\gamma}}(\mathbf{x}_i, \mathbf{x}_j)$ obtained from gaussian process. Covariance kernel depends on hyperparameter γ. The predictive distribution of the GP assessed at the n_* test locations indexed by X_*, assuming additive Gaussian noise, is given byWilson et al. (2016):

$$\begin{aligned}
&\mathbf{f}_* \mid X_*, X, \mathbf{y}, \gamma, \sigma^2 \sim \mathcal{N}(\mathbb{E}[\mathbf{f}_*], \operatorname{cov}(\mathbf{f}_*)) \\
&\mathbb{E}[\mathbf{f}_*] = \boldsymbol{\mu}_{X_*} + K_{X_*,X}\left[K_{X,X} + \sigma^2 I\right]^{-1}\mathbf{y} \\
&\operatorname{cov}(\mathbf{f}_*) = K_{X_*,X_*} - K_{X_*,X}\left[K_{X,X} + \sigma^2 I\right]^{-1} K_{X,X_*}
\end{aligned}$$

where $K_{X_*,X}$ is $n_* \times n$ matrix of covariance obtained form gaussian process between X_* and X. $\boldsymbol{\mu}_{X_*}$ is $n_* \times 1$ vector and $K_{X,X}$ is $n \times n$ covariance matrix obtained form training data XWilson et al. (2016).

2.2.2 Deep kernel architecture

We start from base kernel $k(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta})$ with θ as hyperparameters, inputs $\mathbf{x}$ is transformed as below equation where $g(\mathbf{x}_i, \mathbf{w})$ is a non-linear mapping given by deep architecture such as Long short term memory(LSTM) network parameterized by $\mathbf{w}$ Wilson et al. (2016):

$$k(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta}) \rightarrow k(g(\mathbf{x}_i, \mathbf{w}), g(\mathbf{x}_j, \mathbf{w}) \mid \boldsymbol{\theta}, \mathbf{w})$$

Here $k(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta})$ is a RBF kernel.

$$k_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\| / \ell^2\right)$$

Spectral mixture kernel is also used in comparison with RBF kernel: $k_{\mathrm{SM}}(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta}) =$

$$\sum_{q=1}^{Q} a_q \frac{|\Sigma_q|^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\left\|\Sigma_q^{\frac{1}{2}}(\mathbf{x} - \mathbf{x}')\right\|^2\right) \cos\langle \mathbf{x} - \mathbf{x}', 2\pi\boldsymbol{\mu}_q\rangle .$$

$\boldsymbol{\theta} = \{a_q, \Sigma_q, \boldsymbol{\mu}_q\}$ are the parameters of spectral mixture kernel where a_q is mixture weights, bandwidths (inverse length-scales) Σ_q, and frequencies $\boldsymbol{\mu}_q$.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}}{\partial K_{\boldsymbol{\gamma}}} \frac{\partial K_{\boldsymbol{\gamma}}}{\partial \boldsymbol{\theta}}, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}}{\partial K_{\boldsymbol{\gamma}}} \frac{\partial K_{\boldsymbol{\gamma}}}{\partial g(\mathbf{x}, \mathbf{w})} \frac{\partial g(\mathbf{x}, \mathbf{w})}{\partial \mathbf{w}}$$

This is the implicit derivative of log marginal likelihood with regard to our $n \times n$ data covariance matrix K_γ:

$$\frac{\partial \mathcal{L}}{\partial K_{\boldsymbol{\gamma}}} = \frac{1}{2}\left(K_\gamma^{-1}\mathbf{y}\mathbf{y}^\top K_\gamma^{-1} - K_\gamma^{-1}\right)$$

$\sigma^2 I$ is covariance matrix which is part of base kernel hyperparameter $\boldsymbol{\theta}$. $\frac{\partial K_{\boldsymbol{\gamma}}}{\partial \boldsymbol{\theta}}$ is derivatives of the deep kernel with respect to $\boldsymbol{\theta}$ conditioned on fixed transformation of inputs $g((\boldsymbol{x}, \boldsymbol{w}))$. Also, $\frac{\partial K_{\boldsymbol{\gamma}}}{\partial g(\mathbf{x}, \mathbf{w})}$ is implicit derivative of deep kernel with respect to g, keeping $\boldsymbol{\theta}$ fixed. And $\frac{\partial g(\mathbf{x}, \mathbf{w})}{\partial \mathbf{w}}$ is the derivative with respect to weight which is computed using backpropogation.

3 Experiment

3.1 Dataset details

Shift Dataset : The dataset is part of Shift Challenge (Malinin et al., 2022) The Shifts vessel power estimation dataset is based on minute-by-minute sensor observations from a merchant ship over four years, cleaned and supplemented with third-party meteorological data. From the vessel's speed, draught, time since last dry dock cleaning, and weather and sea conditions, anticipate the main engine shaft power, which can be used to predict fuel consumption given an engine model.
Shift-21K is subset of the Shifts vessel power estimation dataset. Air Quality dataset: This dataset is part of UCI Machine learning repository. (Vito, 2016). Conatins data from five metal oxide chemical sensors device such as CO, tungsten oxide, indium oxide, benzene, titania and humidity. The target variable used to estimate uncertainty is temperature. Hyperparameters choosen for each experiment can be found in Appendix A.1.

Dataset	Instances	Attributes
Shift	559406	10
Shift-21K	21800	10
Air Quality	9357	14

Table 1: Dataset details

3.2 Results

We compare error metrics such as Mean absolute error(MAE), Root Mean Square error(RMSE), Symmetric Mean absolute percentage error(sMAPE) and Correlation metric(R2 score) for Deep kernel learning trained without active learning strategy which is denoted as DKL and with Active learning which is denoted as Active-DKL in table 2. For all six empirical experiments, spectral kernel performed better than RBF. From 2, it seems Active learning with Deep kernel performs well for small and Large size dataset. But for Shift-21 Dataset it seems training on subset of data which is enough to capture distribution shift performs well for Deep kernels (not trained with active learning).

Method	Dataset	MAE	RMSE	sMAPE	R2 score
DKL	Shift	2784.4573	3412.79	0.10	0.59
Active-DKL	Shift	**1260.52**	**1578.82**	**0.04**	**0.91**
DKL	Shift-21K	**1108.18**	**1439.54**	**0.03**	**0 .92**
Active-DKL	Shift-21K	1716.35	2116.19	0.03	0.83
DKL	Air Quality	3.94	4.55	0.35	0.42
Active-DKL	Air Quality	**2.68**	**3.59**	**0.18**	**0.64**

Table 2: Performance error metric on test set

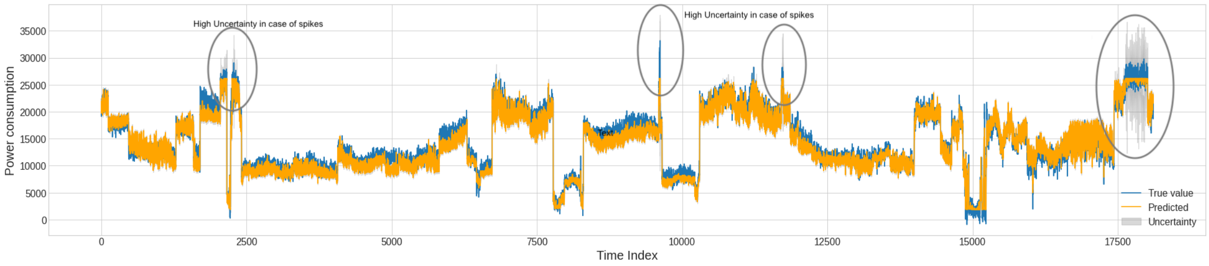


Figure 2: Uncertainty estimation using Active learning approach for Shift Dataset

Comparing Figure :2 and Figure :7, it seems that spikes are highly uncertain while performing Active learning with Deep kernels rather than Figure: 6 For Air Quality dataset where we estimate uncertainty for temperature, actively learned kernels perform better but as higher of temperature frequently occures in dataset, model cannot generalize for lower temperature values as seen in Figure 8 and 9. More graphs for empirical results can be found in A.2.

4 Discussion

In this work, two approach are described to estimate epistemic uncertainty. Actively learned deep kernel performs well where the model needs to be aware of its own limitations and provide meaningful uncertainty estimates for decision-making purposes. However, it's worth noting that active learning with deep kernels can be computationally expensive since it requires training deep neural networks and evaluating uncertainty estimates for a large number of unlabeled examples. Additionally, careful consideration should be given to the selection of informative examples and the design of the active learning strategy to ensure effective uncertainty estimation and model improvement. We select query points using pairwise distance, other approach such as using K-means clustering can also be used. Here we test for regression tasks and structured data, whereas it would be interesting future direction to explore similar approach for classification and unstructured data.

Author Contributions

Hena Ghonia worked on developing idea and writing code for Actively-learned Deep kernels to estimate epistemic uncertainty. Shanaya Mehta worked on debugging the code and examining correctness of algorithm.

Acknowledgments

We acknowledge guidance of Luc Vouligny, Martin Gagnon and Luc Cauchon during internship at Hydro-Quebec for work done in deep kernel learning.

References

Punit Kumar and Atul Gupta. Active learning query strategies for classification, regression, and clustering: a survey. *Journal of Computer Science and Technology*, 35:913–945, 2020.

David JC MacKay et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168: 133–166, 1998.

Andrey Malinin, Andreas Athanasopoulos, Muhamed Barakovic, Meritxell Bach Cuadra, Mark Gales, Cristina Granziera, Mara Graziani, Nikolay Kartashev, Konstantinos Kyriakopoulos, Po-Jui Lu, Nataliia Molchanova, Antonis Nikitakis, Vatsal Raina, Francesco La Rosa, Eli Sivena, Vasileios Tsarsitalidis, Efi Tsompopoulou, and Elena Volf. Shifts Marine Cargo Vessel Power Consumption Prediction Dataset, September 2022. URL https://doi.org/10.5281/zenodo.7057666. This work is supported by the Hasler Foundation, Cambridge University Press and Cambridge Assessment and DeepSea.

Osman Mamun, MFN Taufique, Madison Wenzlick, Jeffrey Hawk, and Ram Devanathan. Uncertainty quantification for bayesian active learning in rupture life prediction of ferritic steels. *Scientific Reports*, 12(1):2083, 2022.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Sebastian W Ober, Carl E Rasmussen, and Mark van der Wilk. The promises and pitfalls of deep kernel learning. In *Uncertainty in Artificial Intelligence*, pp. 1206–1216. PMLR, 2021.

Massimiliano Patacchiola, Jack Turner, Elliot J Crowley, Michael O'Boyle, and Amos J Storkey. Bayesian meta-learning for the few-shot setting via deep kernels. *Advances in Neural Information Processing Systems*, 33:16108–16118, 2020.

Prudencio Tossou, Basile Dura, Francois Laviolette, Mario Marchand, and Alexandre Lacoste. Adaptive deep kernel learning. *arXiv preprint arXiv:1905.12131*, 2019.

Saverio Vito. Air Quality. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C59K5F.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning*, pp. 1067–1075. PMLR, 2013.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International conference on machine learning*, pp. 1775–1784. PMLR, 2015.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pp. 370–378. PMLR, 2016.

A Appendix

A.1 Hyperparameter details

For all the experiments batch size was 256 and kernel selected was spectral for Exact gaussian process. Loss function is Marginal log likelihood. Number. of epochs for Deep Kernel learning method was 50 and for active learning it was 5, 5, and 50 for Shift dataset, Shift-21K dataset and AirQuality dataset. Learning rate for feature extractor, mean

module and likelihood was 0.001 and for covariance module learning rate was 0.0001.
Technical specification: Model was trained on Tesla T4, with 16 GB memory on torch 2.0.1+cu118. Python version used was 3.10.12.

A.2 Empirical results

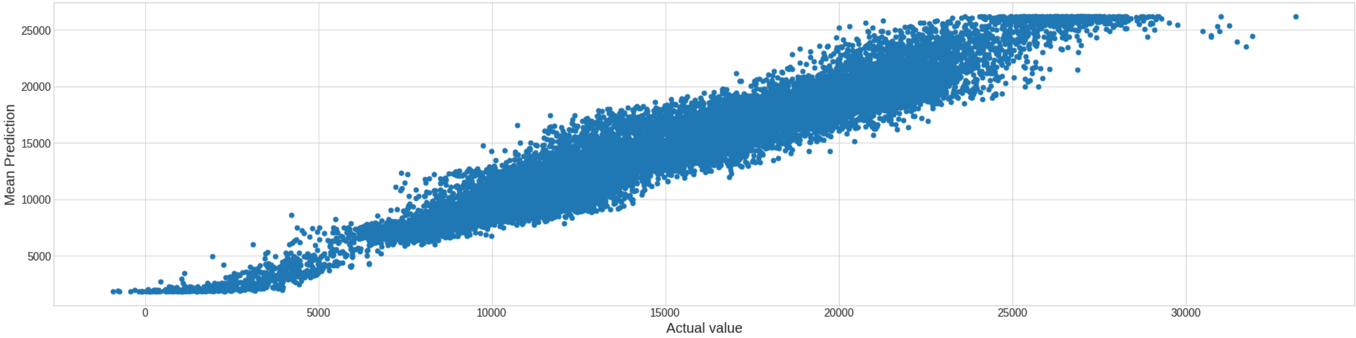


Figure 3: Actuals vs Predicted via Active learning approach on test set of Shift Dataset

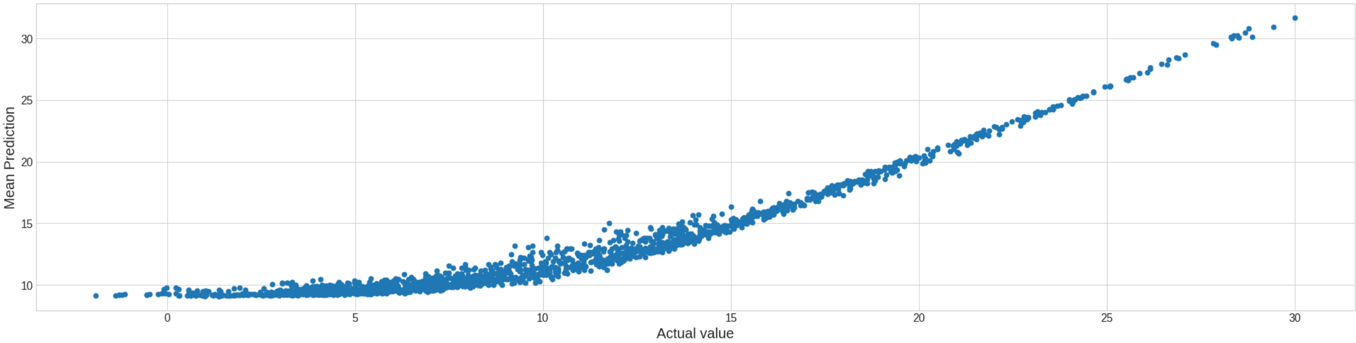


Figure 4: Actuals vs Predicted via Active learning approach on test set of Air Quality Dataset

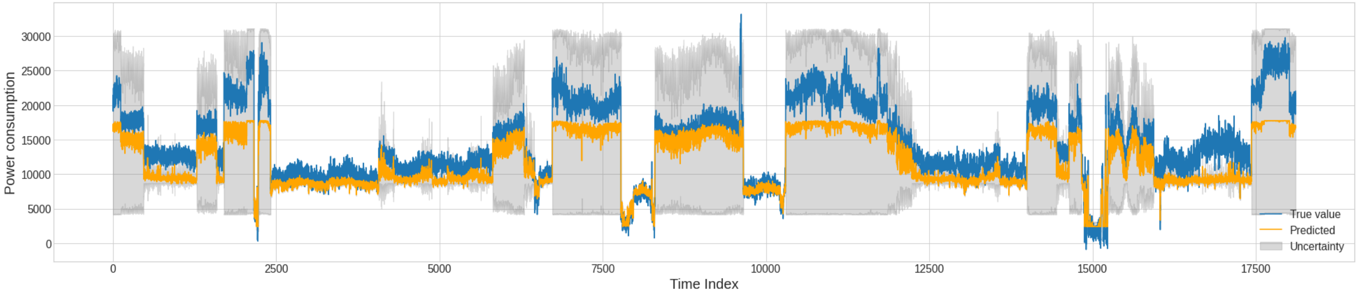


Figure 5: Uncertainty plot for Deep Kernel learning on test set of Shift Dataset

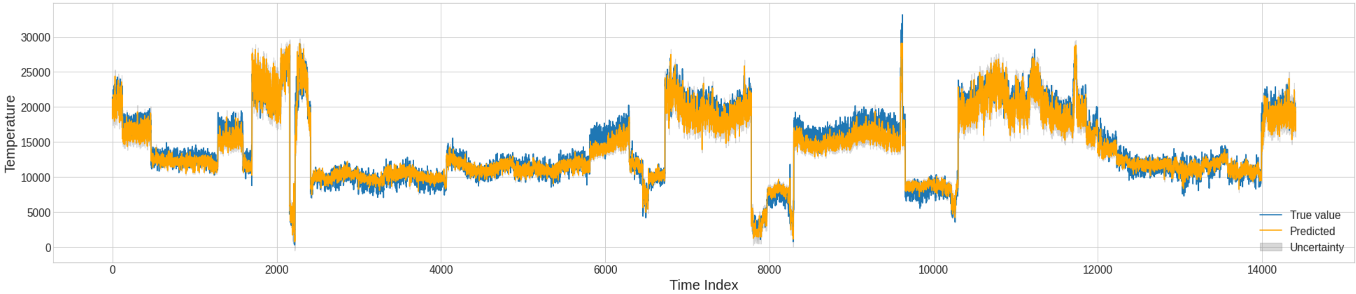


Figure 6: Uncertainty plot for Deep Kernel learning on test set of Shift-21K Dataset

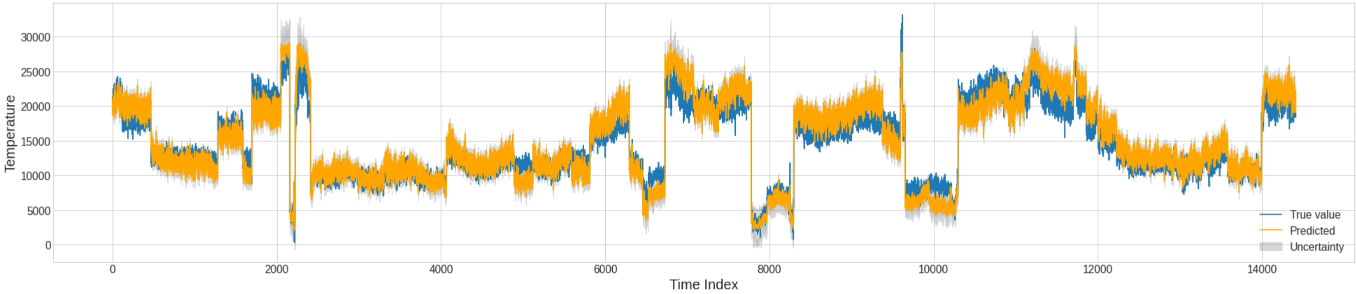


Figure 7: Uncertainty estimation using active learning approach for Shift-21K Dataset

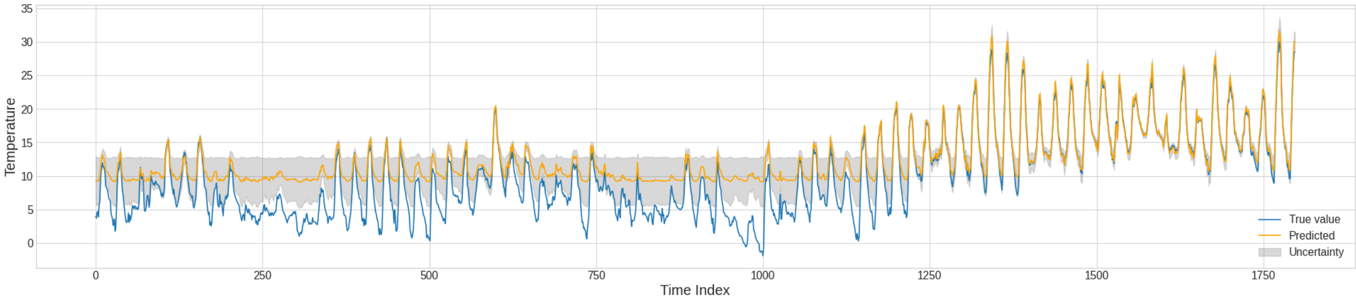


Figure 8: Uncertainty estimation using Active learning approach for AirQuality dataset

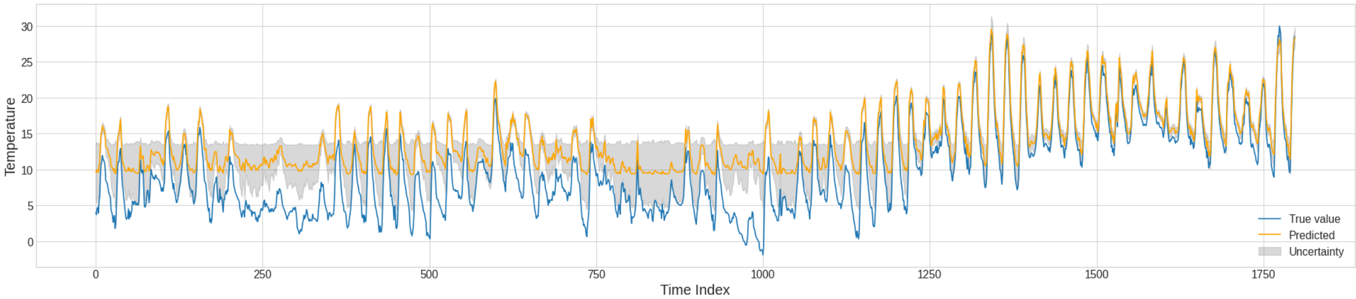


Figure 9: Uncertainty plot for Deep Kernel learning on test set of Air Quality dataset