Using Actively-learned Deep kernels to estimate epistemic uncertainty

Hena Ghonia [1 2] Shanaya Mehta [3]

[1]Mila – Institut Québécois d'Intelligence Artificielle [2]Université de Montréal [2]Assert AI

Context

- 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
- Epistemic uncertainty refers to the uncertainty arising from the lack of knowledge about the true underlying data distribution.

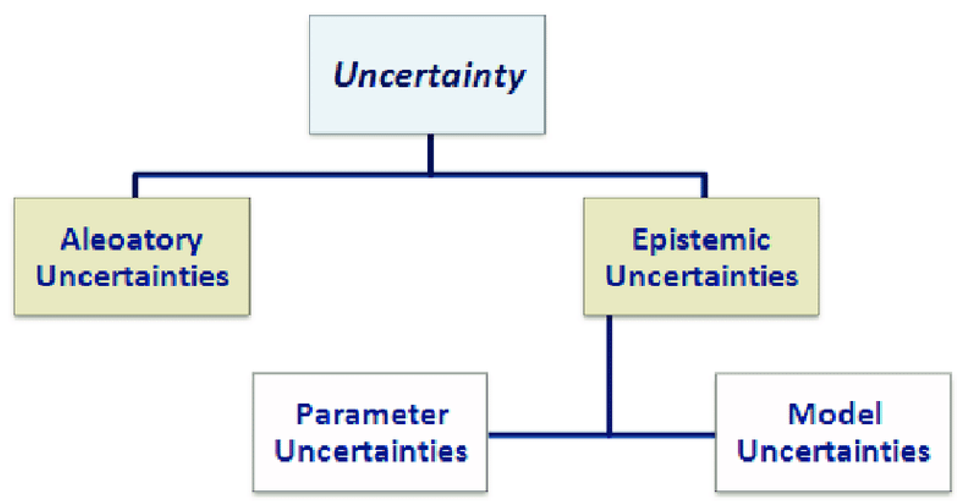

Figure 1. Type of data uncertainties[1]

- Our method consists of Actively learned deep kernels which has Long short-term memory (LSTM) and feed forward network as feature extractor with covariance kernel [3] to create expressive and scalable closed form kernels [4].

Prior work with Deep kernels

- 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.ober2021promises
- Gaussian Process:
  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 $\boldsymbol{\mu}_i = \mu\left(x_i\right)$ and covariance matrix $\left(K_{X,X}\right)_{ij} = k_\gamma\left(\mathbf{x}_i, \mathbf{x}_j\right)$ 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 by[5]:

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

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 X[5].

Approach

- 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):

$$U = \sqrt{\operatorname{cov}\left(\mathbf{f}_*\right) E\left[\mathbf{f}_*\right]^2}$$

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

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

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

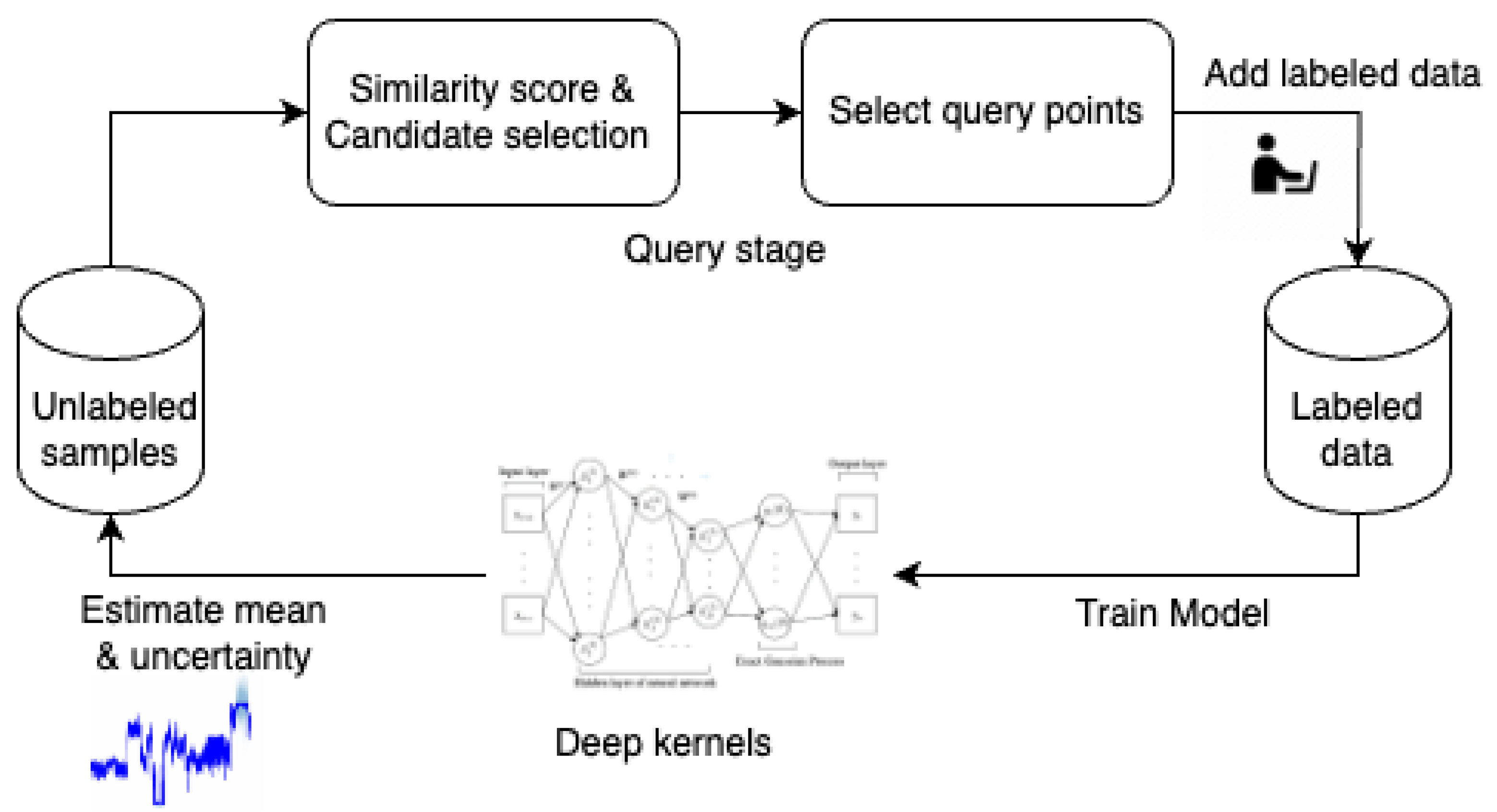

Figure 2. Batch active learning cycle

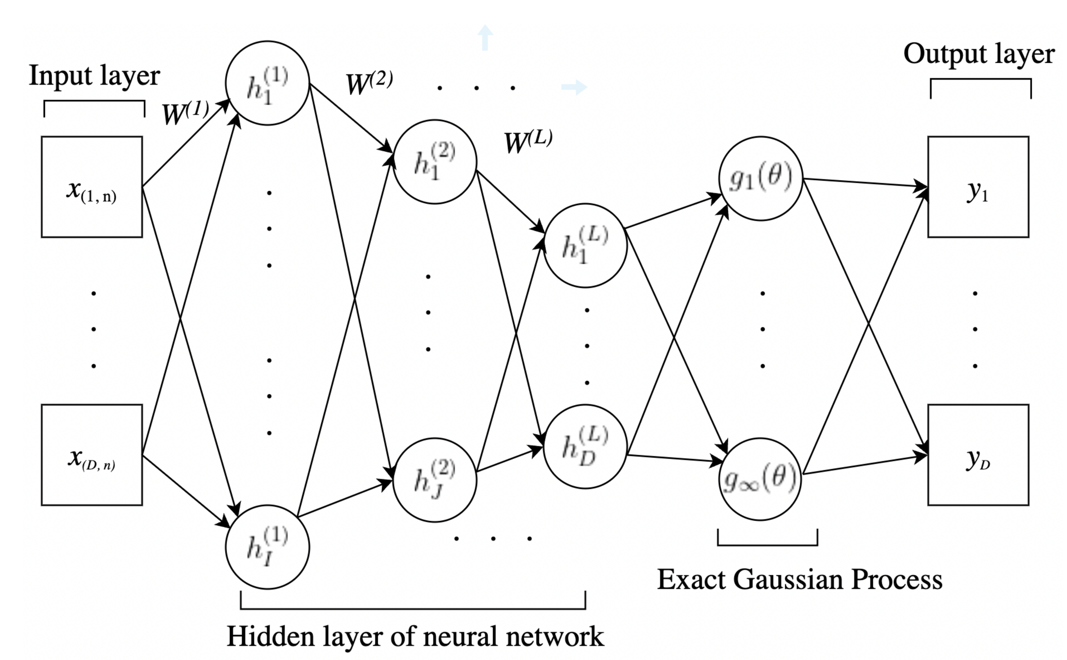

Figure 3. Architecture of Deep kernel

Results

For all six empirical experiments, spectral kernel performed better than RBF.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 below experiments:

Method	Dataset	Dataset attribute	MAE	RMSE	sMAPE	R2 score
DKL	Shift-559K	10	2784.4573	3412.79	0.10	0.59
Active-DKL	Shift-559K	10	**1260.52**	**1578.82**	**0.04**	**0.91**
DKL	Shift-21K	10	**1108.18**	**1439.54**	**0.03**	**0 .92**
Active-DKL	Shift-21K	10	1716.35	2116.19	0.03	0.83
DKL	Air Quality-9K	14	3.94	4.55	0.35	0.42
Active-DKL	Air Quality-9K	14	**2.68**	**3.59**	**0.18**	**0.64**

Table 1. Performance error metric on test set

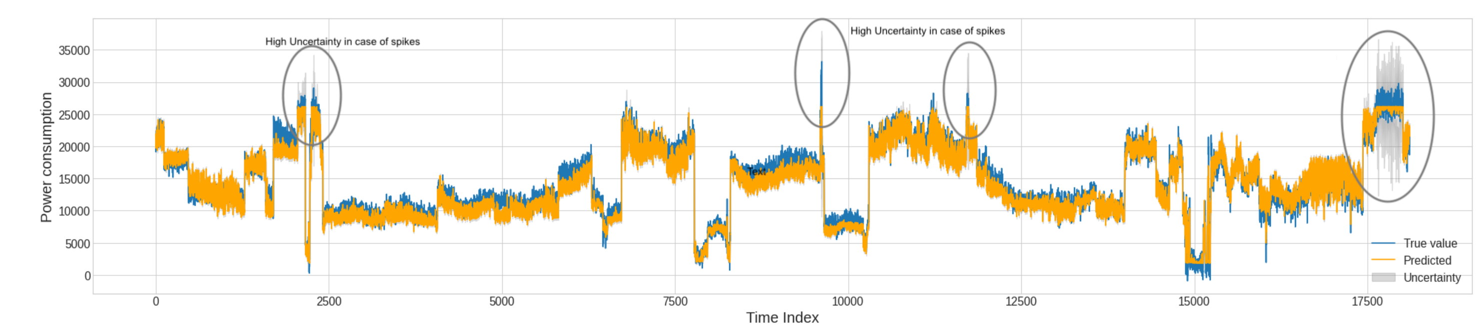

Figure 4. Uncertainty estimation using Active learning approach for Shift Dataset

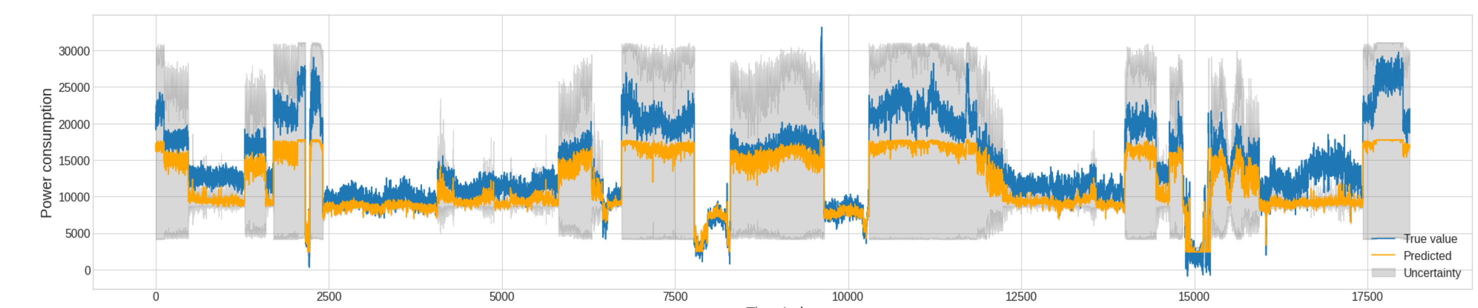

Figure 5. Uncertainty plot for Deep Kernel learning on test set of Shift Dataset

Discussion

- Active learning with deep kernels can be computationally expensive!
- Careful consideration should be given to the selection of informative examples.
- We select query points using pairwise distance, other approach such as using K-means clustering can also be used.
- Future direction to explore similar approach for classification and unstructured data

References

[1] Nabil Ben Said Amrani, Laurent Saintis, Driss Sarsri, and Mihaela Barreau. Evidential networks for evaluating predictive reliability of mechatronics systems under epistemic uncertainties. *Journal of Konbin*, 49:459–475, 06 2019.

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

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

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

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

https://github.com/Hstellar/Active-DKL

hena.ghonia@umontreal.ca